Using 3D Point Cloud Data and Machine Learning to Assess Skeletal Remains

Ph.D. Thesis

Jessica Frances Lam
jfl6@le.ac.uk
School of Archaeology & Ancient History
INTREPID Forensics Programme
University of Leicester

Supervised by:

Dr. Jo Appleby[1] and Prof. Jeremy Levesley[2]

[1]School of Archaeology & Ancient History, University of Leicester
[2]Department of Mathematics, University of Leicester

July 7, 2020

Abstract

The purpose of this PhD project was to create a new method of sex assessment for crania using 3D point cloud data and machine learning. Through the process of investigating sexual dimorphism, this project has been the first to define the "discrimination factor", which provides both researchers and practitioners of forensic anthropology a new tool for quantifying sexual dimorphism and comparing morphological traits. This project also created a ground-truth database of 3D point cloud data by using structured light scanning to document 534 crania (263 belonging to females and 271 belonging to males) from four diverse skeletal collections (located in the United Kingdom, Japan, Italy, and South Africa). A program called CraniAlign was created in conjunction with Clotho AI to process the 3D point cloud data in a manner that was transparent, reliable, and allowed for automation. In the first study of its kind, CraniAlign was compared to DAVID 4, which is the industry standard, in order to demonstrate that proprietary algorithms are not ideal for research. Finally, the 3D point cloud data of 316 individuals (134 female, 182 male) were used to train and test artificial neural networks. Three methods were successfully created – one that sought to classify individuals according to sex regardless of the population to which they belonged; one for classifying individuals according to sex and population; and one for classifying individuals into population groups regardless of sex. All three methods yielded training accuracies of 97.1% - 100.0% and evaluation accuracies of 87.5% - 92.5%. This project was therefore the first to apply deep learning to the problems of sex, population, and population-specific sex classification using the entire geometry of the cranium, and has successfully established three methods with unprecedented performances when tested on samples which were not involved in the training of the models.

Acknowledgements

I would like to first and foremost acknowledge and wholeheartedly express my gratitude to my two supervisors, Dr. Jo Appleby and Prof. Jeremy Levesley. I am grateful for the fact that both of you supported me and believed in me from the start, and for agreeing to take me on as your student under such unexpected circumstances. The encouragement and positive feedback I received from the both of you have enabled me to see my PhD through to its submission, despite the hardships that I encountered. Thank you for being exactly the kind of supervisors I needed.

I am also very grateful to my thesis examiners, Prof. Richard Thomas and Dr. Elena Kranioti. Thanks to the two of you, the examination process was both enjoyable and challenging. Due to your insightful comments and constructive feedback, I was able to improve my thesis further into a satisfying piece of work. Thank you tremendously for being amazing and thorough examiners!

Thank you also to Dr. Lisa Smith for not only coordinating the INTREPID Forensics Programme but also for your unwavering support in helping me get to where I am now. Thank you for staying positive and helping me get my PhD on track - I would not have had the privilege to work with such incredible people if it weren't for you!

Thank you to Dr. Tracy Rogers (University of Toronto Mississauga, Director of the Forensic Science Program), for whom I would not even be a part of this programme, and also for graciously accepting to host me for my secondment. Gratitude is also extended to Mr. Tom Horton whose meticulous organizational skills and candid sense of humour has made everything function as smoothly and enjoyably as possible, and to Mr. Alex Murphy who did a great job in keeping things organized in Tom's absence.

I would like to also thank Dr. Etienne Pillin, whose expertise, support, and encouragement helped me stay somewhat sane throughout this entire process. Whenever I became particularly pessimistic about my ability to complete my PhD, you were always overwhelmingly positive and showed me that I had the necessary abilities to take my project in the direction I envisioned. This project, and my ability to finish, would not have been possible without you.

Contents

Chapter 1

Introduction & Background

1.1 Purpose & Significance

The purpose of this research project is to use machine learning to create a method of skeletal sex assessment using 3D models of the cranium, with the following aims: 1) the method has an accuracy (i.e. the frequency with which the parameters are correctly determined) that is comparable or greater than existing assessment methods using the cranium; 2) the error rate of the method is quantifiable and known; 3) the method is reproducible (i.e. the method is not subject to large inter and intraobserver errors); and 4) the method abides by the Daubert criteria (1993), which is necessary for its admissibility in court and forensic purposes.

In both bioarchaeology and forensic anthropology, a major component of assessing human skeletal remains is the biological profile, which primarily involves establishing the age, sex, and ancestry of the individual (Ubelaker 2008; Braz 2009; Gapert et al. 2009a; Rogers 2009; Calce 2012; Moore 2013; Lam et al. 2016). By supplementing the biological profile with trauma, pathology, and burial practice assessments, bioarchaeologists can analyze demographic information about past populations such as disease, mortality, and nutrition (Roberts and Manchester 2005; Walker 2008a). These analyses are pivotal in understanding how our species has adapted, both genetically and physiologically, to environmental and societal stresses in the past (Larsen and Walker 2004; Walker 2008a). Assessing human skeletons is also crucial in a modern context for forensic anthropologists. When establishing the identity of a skeletonized

individual, missing person profiles can be included or excluded as a possible match on the basis of the biological profile obtained from skeletal analyses (Rogers 1999; Calce and Rogers 2011; Moore 2013; Lam et al. 2016). Once the list of possible missing persons who fit the biological profile has been established, comparative methods can then be undertaken to establish the identity of the individual.

It is imperative that methods of assessing age, sex, and ancestry from the skeleton are accurate. If not, this can drastically affect the interpretations of a skeletal population in an archaeological context. For example, Mays (1993) used traditional regression methods for estimating the ages of juveniles in a Roman British sample, and determined that the juvenile mortality was caused by infanticide; however, using a Bayesian statistical model for estimating age, Gowland and Chamberlain (2002) interpreted the ages-at-death to be more consistent with a natural mortality distribution. Such contrasting interpretations can sometimes be addressed by considering evidence from historical documents or records to support one interpretation over another, but this disparity remains a severe problem in prehistoric archaeology since no records exist apart from the skeletons themselves. In forensic anthropology, an inaccurate biological profile is also problematic because it may cause the true identity of the skeletonized individual to be excluded from the list of possible missing persons. Not only would such a result hinder the identification process, but it can also damage the credibility of the forensic anthropologist if he/she is required to testify in court (Christensen 2004; Rogers and Allard 2004; Williams and Rogers 2006).

Traditionally, skeletal assessments are performed by an analyst using either metric or morphological methods. The accuracy of the results depend on two major factors: 1) the skill and previous experience of the examiner (White et al. 2012); and 2) the clarity of the method itself, a topic which is discussed by this PhD researcher elsewhere (Lam et al. 2016). While the latter is merely an issue with conveying how techniques should be performed, the former addresses the inherent subjectivity in using human analysts (Moore 2013). Morphological methods rely on the analyst's experience with osteological landmarks and trait expressions; therefore, an inexperienced analyst may assess landmarks differently than an experienced one. Even metric methods, which rely on measurements between defined osteological landmarks, are conditional to the analyst's ability to recognize those landmarks. To illustrate the degree of subjectivity in metric methods, a study by Adams and Byrd (2002) discovered that

there was high interobserver error in some osteological measurements performed by analysts with less than five years of osteometric experience, due to issues with landmark recognition. Even more concerning is that the ability to correctly recognize these problematic landmarks did not improve in analysts with greater levels of experience (Adams and Byrd 2002). It is unclear whether the issue is due to the clarity of the landmark definitions or to the subjective experience of an analyst being able to correctly find the osteological landmarks, but the study does highlight the need for reliability and repeatability in osteological analyses.

With the increasing capabilities of modern technology, skeletal assessment methods have been adapted for computerized analyses in order to make the methods more objective. Despite analyzing metric data with computers, however, the data are still largely obtained manually and therefore suffer from the same issues of subjectivity outlined above. This project seeks to address the bias associated with using human analysts to choose metrically-defined landmarks, as part of the larger aim of creating a new computerized method of sex assessment for determining the biological profile. Additionally, due to the fact that sexual characteristics are affected by ancestry, this method also seeks to classify individuals into geographical populations as part of the sex assessment process.

Finally, it is important to consider that techniques used by forensic scientists must abide by the Daubert Criteria (1993). Following the United States Supreme Court ruling in Daubert v. Merrell Dow Pharmaceuticals Inc. (1993), four criteria were established in order for scientific expert witness testimony to be admissible: 1) the judge is the gatekeeper of evidence and is ultimately the ruling power that can decide what evidence is and is not admissible; 2) the evidence must be relevant to the trial, and also reliable in its source and in how it was analyzed; 3) the conclusions presented from analyzing the evidence must be based on scientific knowledge and should be demonstrably proven to have been derived according to the scientific method; 4) the scientific methodology governing the theory or technique upon which evidence is analyzed is further defined by five requirements: a) the theory or technique must be falsifiable, refutable, and/or testable; b) the theory/technique should have been subject to prior peer review and publication; c) a known or potential error rate associated with the analysis should have been established; d) the testing of the theory or technique should have been subject to the proper standards and controls; and e) the analysis should be generally accepted by the relevant scientific community. In forensic research, the fourth Daubert criterion - and there-

fore the associated five requirements governing the theory or technique - is the most relevant. This research project therefore aims to create an analytical computer program that satisfies the fourth Daubert criterion, especially regarding the establishment of known error rates associated with skeletal analyses, which is variable and ultimately unquantifiable when human analysts are involved. To achieve this, the use of a computer program in this project to remove human error will increase the reliability of the method that is developed. A 3D ground-truth reference database of crania is used for both developing the assessment method and establishing a known, quantifiable error rate. The new method of sex and ancestry assessment therefore meets court admissibility standards to ensure that it will be useful to both bioarchaeologists and forensic anthropologists.

1.2 Skeletal Sex Assessment & Sexual Dimorphism in the Cranium

Biological sex refers to whether an individual is genetically male or female, which is then expressed physiologically in the skeleton due to hormones and other genetically-controlled variables (Sutter 2003; Moore 2013; Bulut et al. 2016). Sex should be distinguished from gender, which is a complex social construct that encompasses age, ethnicity, race, and social status (Joyce 2005; Hollimon 2011). This research project focuses exclusively on sex rather than gender, although gender can influence skeletal sex assessment if individuals undergo hormone replacement therapy or craniofacial surgery. Sexual dimorphism refers to the differences in body shape and size, as well as differences in rate and timing of development, between males and females in a single species (Stinson et al. 2012). Sexual dimorphism in the human skeleton therefore allows analysts to assess whether an individual is male or female. As noted by several authors (e.g. Frayner and Wolpoff 1985; Ross et al. 2003; Galdames et al. 2008; Veroni et al. 2010; Stinson et al. 2012; Garvin et al. 2014), sexual dimorphism is significantly affected by the environment, a prime example being that body size in males are affected more than females by nutritional deficits which results in less sexual dimorphism (Bogin 1999; Ross et al. 2003; Galdames et al. 2008). It is therefore imperative to understand how environmental variables can interact with genetic variables, in order to both identify reliable skeletal traits that truly reflect biological sex, and to interpret them properly if they are confounded by the environment. The

focus of this research project is therefore to identify a reliable method of using skeletal traits that exhibit sexual dimorphism, which is then investigated to determine population differences.

Population differences encompass environmental variables since, from an evolutionary perspective, a population is defined as a group of individuals sharing the same geographic area and culture (DiGangi and Hefner 2013). Culture and geography form the environment for a given population, and have been shown in numerous studies to influence human variation significantly (e.g. Kennedy 1995; Lahr 1996; Cartmill 1999; Edgar and Hunley 2009; Relethford 2009, etc.) Geography influences dietary resources available to individuals, whereas culture can influence how dietary resources are allocated and used, ultimately affecting nutrition (DiGangi and Hefner 2013). Geography also dictates pathogen and climate exposures whereas culture affects how society responds to such factors (DiGangi and Hefner 2013). Furthermore, the environment affects genetic patterns of evolution by influencing which traits are advantageous, neutral, or disadvantageous (DiGangi and Hefner 2013) in a manner that allows for biological distance - how closely related populations are to one another - to be calculated (Stojanowski and Schillaci 2006). Populations exhibit variable degrees of human variation, and therefore variable degrees of sexual dimorphism (González et al. 2007). The approach of using populations to understand these degrees of variation is referred to as ancestry assessment, which is one of the three main variables in a skeletal biological profile. This research project therefore studies ancestry insofar as to address the variation in sexual dimorphism between different populations, as ancestry has an important bearing on the ability to distinguish between the sexes.

Addressing sexual dimorphism in human populations is particularly challenging as males and females possess physical traits that share approximately 95% of the total range of variation (St. Hoyme and Işcan 1989; Schwartz 1995; Rogers 1999). Robustness and rugosity associated with musculature, as well as morphological differences related to child-bearing and childbirth in females, primarily account for the differences between males and females (Rogers 1999). Unsurprisingly, the pelvis has been therefore cited as the most reliable skeletal element for assessing sex (e.g. Đuric 2005; Decker et al. 2011; Spradley and Jantz 2011; Christensen et al. 2014; Krishan et al. 2016) for its role in reproduction. The skull, which refers to both the cranium and the mandible, has been traditionally viewed and taught as the second most reliable skeletal indicator of sex (Pickering and Bachman 1997; Byers 2002; Bass 2005; Moore

2013); however, empirical research throughout the past few decades have shown that post-cranial elements are actually more reliable (e.g. Berrizbeitia 1989; Robling and Ubelaker 1997; France 1998; Klepinger 2006; Spradley and Jantz 2011). For ancestry assessment, the skull is recognized as the most useful indicator due to the numerous population-specific traits that are available (most notably exemplified in research by Rhine 1990, Gill and Gilbert 1990, and Hefner 2009, with a combined trait list compiled and tested by Wood 2015). This research therefore seeks to combine sex and ancestry analyses using the features of the cranium in order to improve its usability for generating an accurate biological profile. The mandible is not included in this project, and all analyses focus solely on the cranium.

The skull as a whole is an important element in forensic anthropology for the purpose of identification. To illustrate this, Komar and Potter (2007) examined 773 cases from the New Mexico medical examiner's office between 1974-2006 and found that in cases where a skull was recovered, 87% of the individuals were successfully identified; if no skull was recovered, only 61% were identified. In the latter scenario, the researchers noted that identification rates were negatively impacted regardless of the percentage of postcranial elements that were recovered (Komar and Potter 2007), meaning that the presence of the skull significantly affects the ability to identify an individual. Although it has been widely accepted through anecdotal work in forensic anthropology that identification of human remains can be hindered if the remains are not fully recovered (Haglund and Reay 1993; Komar 2004), Komar and Potter's (2007) research is one of the earliest published attempts at actually quantifying and demonstrating this impact, as well as highlighting the importance of the skull in identification. Skulls, whether complete or fragmentary, are the most common skeletal element to be recovered (Bass and Driscoll 1983); it is therefore reasonable to improve techniques using the skull in order to maximize the ability to identify the individual. By creating a computerized method of population-specific sex assessment using the cranium, this research project provides analysts with another tool that can be used to create the biological profile for identification purposes.

Sexually dimorphic features in the skull have been noted in diverse populations, although to different extents (e.g. Ascadi and Nemeskeri 1970; Loth and Henneberg 1996, 1998; Vidarsdottir and O'Higgins 2001; Rosas and Bastir 2002; Thayer and Dobson 2010; Coquerelle et al. 2011). The theory behind sexual dimorphism in the skull lies in the differential growth and development that males and females undergo throughout life, due to differences in bone re-

sponse to hormone levels. Specifically, differences in the quantity of testosterone are primarily responsible for the sexual dimorphism that is observable in the skeleton (Sutter 2003). Testosterone is detectable by the tenth week of fetal development, peaking during the 15th week (Grumbach and Kaplan 1974; Challis et al. 1976). By the time an individual is born, sex differences exist in the upper midface due to this peak of testosterone in males during gestation (Bulygina et al. 2006; Coquerelle et al. 2011). From birth, both males and females follow similar craniofacial growth trajectories until puberty, at which point males exhibit a pronounced rate of craniofacial development and are farther along the craniofacial growth trajectory than females (Vidarsdottir 1999; Vidarsdottir and O'Higgins 2001; Bulygina et al. 2006; Coquerelle et al. 2011). Shape and developmental differences between the sexes are most prominent after puberty. The result is larger, more robust, and more rugged features related to muscularity in male crania, whereas females have smaller, rounder, and more gracile features (Işcan and Steyn 2013; Bulut et al. 2016). As females age, however, their craniofacial development continues such that post-menopausal females may resemble males (Walker 1995; Rogers 2005; Moore 2013). Consequently, Krogman and Iscan (1986) have suggested that cranial sex assessment should be limited to individuals aged approximately 20 - 55 years old, as they claim there is too much overlap between the sexes in young and old individuals.

In terms of overall cranial shape, differences between the sexes exist in the contour of several cranial bones. In females, the frontal and parietal bones exhibit bossing, which refers to a smooth, rounded eminence indicating the original centers of ossification (White and Folkens 2005). This creates a more rounded shape in the calvarium (i.e. skullcap) in females, in contrast to a less bulbous shape in males due to the downward elongation of the craniofacial features during male development (Rogers 2005). The result is that females tend to have broader, rounder foreheads whereas males tend to have slightly sloped foreheads (Wilkinson 2004). Despite the fact that these are well-known and accepted traits, there is discordance in the literature about the ability to quantify and/or define forehead sloping in a manner that allows males and females to be distinguished reliably (Rogers 2005; Williams and Rogers 2006; Ramsthaler et al. 2010; Bulut et al. 2016). Bulut and colleagues (2016) suggest that this may be due to a failure in humans to perceive the underlying spherical shape of the male forehead because of prominent features that protrude from the frontal bone, such as the supraorbital ridge and glabella. Figure 1.1 below compares and contrasts the difference in overall skull shape between males and females.
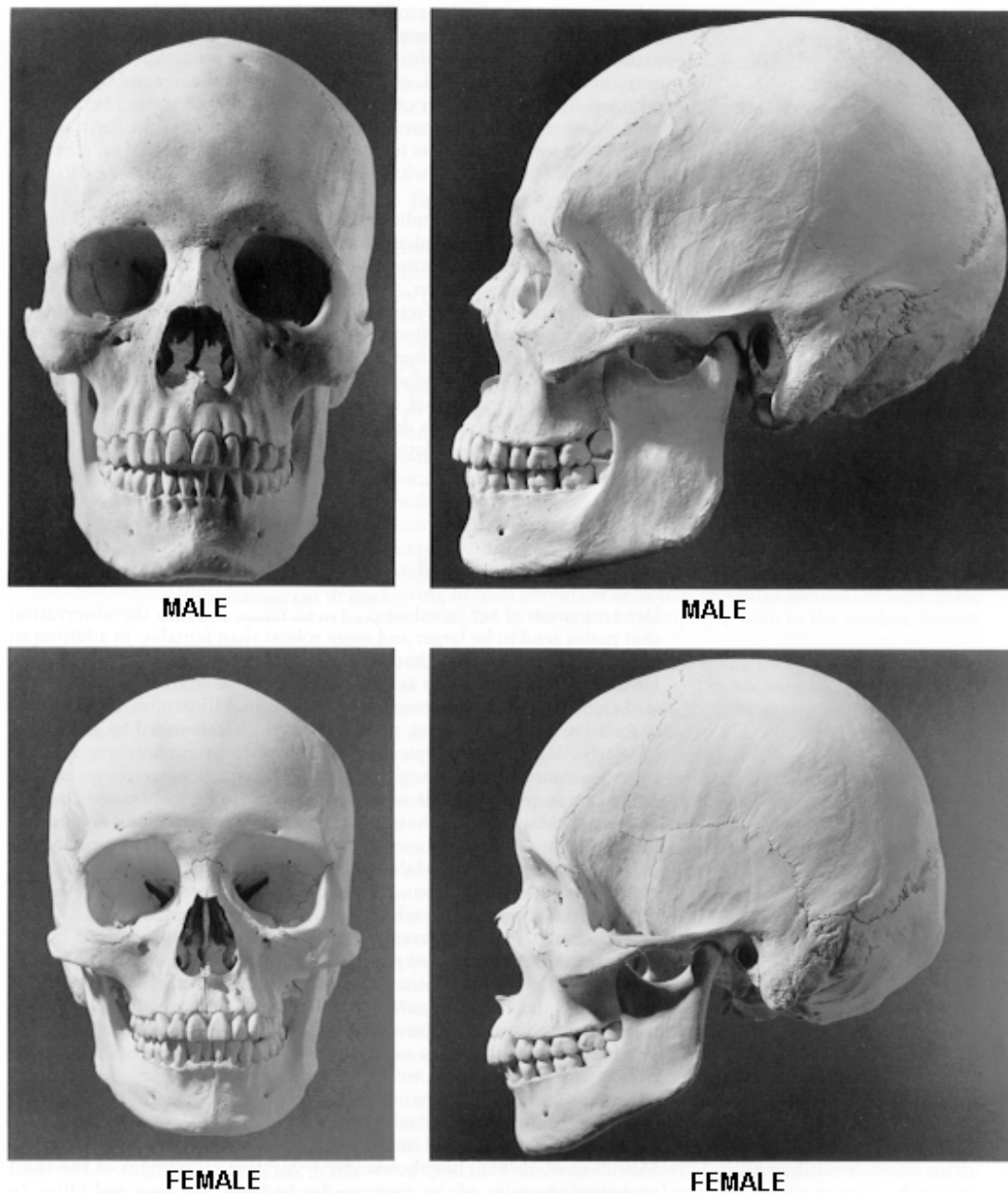
Figure 1.1: Overall shape and feature differences between male (upper images) and female (lower images) skulls. Note the difference in the roundness of the forehead; the shape of the nasal cavity; the orbit size and shape; the shape of the jaw; and the overall muscularity of the skull. Source: White 1991, pg. 321.

Glabella, which is defined as "the most forward projecting point in the midline of the forehead at the level of the supraorbital ridges and above the nasofrontal suture" (Bass 2005 in Chapter 1.3) has been recognized as the most sexually dimorphic and reliable part of the skull for assessing sex (Rogers 2005; Williams and Rogers 2006; Walker 2008b; Garvin et al. 2014). When viewed in the lateral profile, the glabella usually does not project anteriorly much or at all in females, whereas it is pronounced and well-developed in males (Buikstra and Ubelaker

1994). Similarly, the larger supraorbital ridge in males (labelled as the superciliary arch in Figure 1.2) results in a thicker orbital border than in females, the range of thickness which has been described as a "dull knife" to "approximating the width of a pencil" (Buikstra and Ubelaker 1994). The differences in the size of glabella and in the thickness of the supraorbital ridge are due to the growth of the inner table of the frontal bone relative to the outer table (Enlow 1982). The inner table is related to the growth of the frontal lobe, which is completed by the age of six; conversely, the outer table is part of the nasomaxillary complex which finishes remodelling a few years later depending on sex (Enlow 1982). Females do not have the outwards and downwards facial growth spurt that males undergo, meaning that their nasomaxillary complexes finish developing much earlier than males, while males will continue developing in this region for longer. The result is that the outer table continues growing for longer in males than in females, resulting in a thicker supraorbital ridge and glabella, as well as what appears to be a less rounded forehead as the frontal bossing (labelled as the frontal eminence in Figure 1.2) is lost.
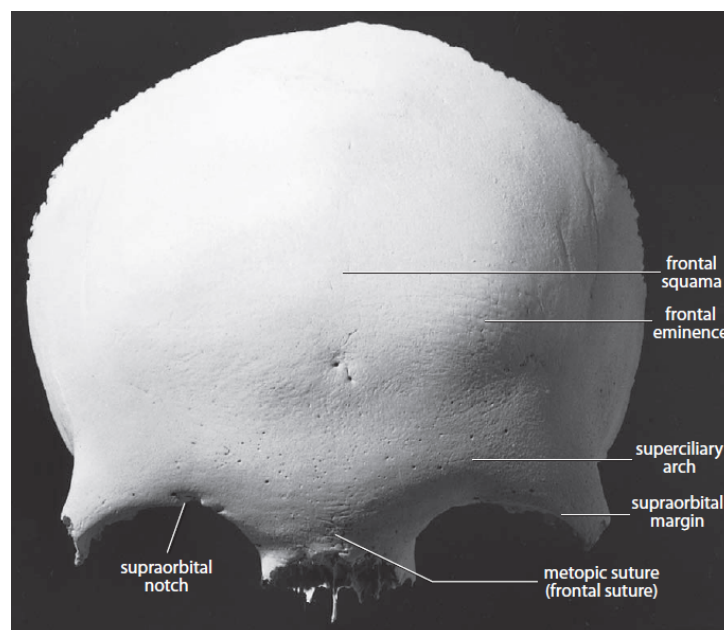
Figure 1.2: The anterior view of the frontal bone. Differences between the sexes exist in the frontal squama, which is flatter and exhibits a less prominent eminence in males; and the thickness of the supraorbital margin/superciliary arch. Source: White and Folkens 2005, pg. 88.

Sexually dimorphic features in the cranial base have been identified, namely the foramen magnum, the occipital condyles, the mastoid processes, and the rugosity of the nuchal crest (See Figure 1.3). The cranial base exhibits less growth-related changes relative to other

areas of the skull, so developmental differences between the sexes is likely to be less promi-nent than in other regions (Buschang et al. 1983; Gapert et al. 2013). Unsurprisingly, studies into the size and shape of both the foramen magnum and the occipital condyles show relatively low distinguishing abilities, ranging from 66.5% - 76% for the foramen magnum depending on the population studied (Galdames et al. 2009; Gapert et al. 2009a), and 67.7% - 76.7% for the occipital condyle (Gapert et al. 2009b; Macaluso Jr. 2011). In a study by Williams and Rogers (2006), evaluating the size of the occipital condyles scored extremely high in terms of intraob-server error (i.e. 12.5% of samples were scored inconsistently), demonstrating the subjectivity of evaluating this trait.
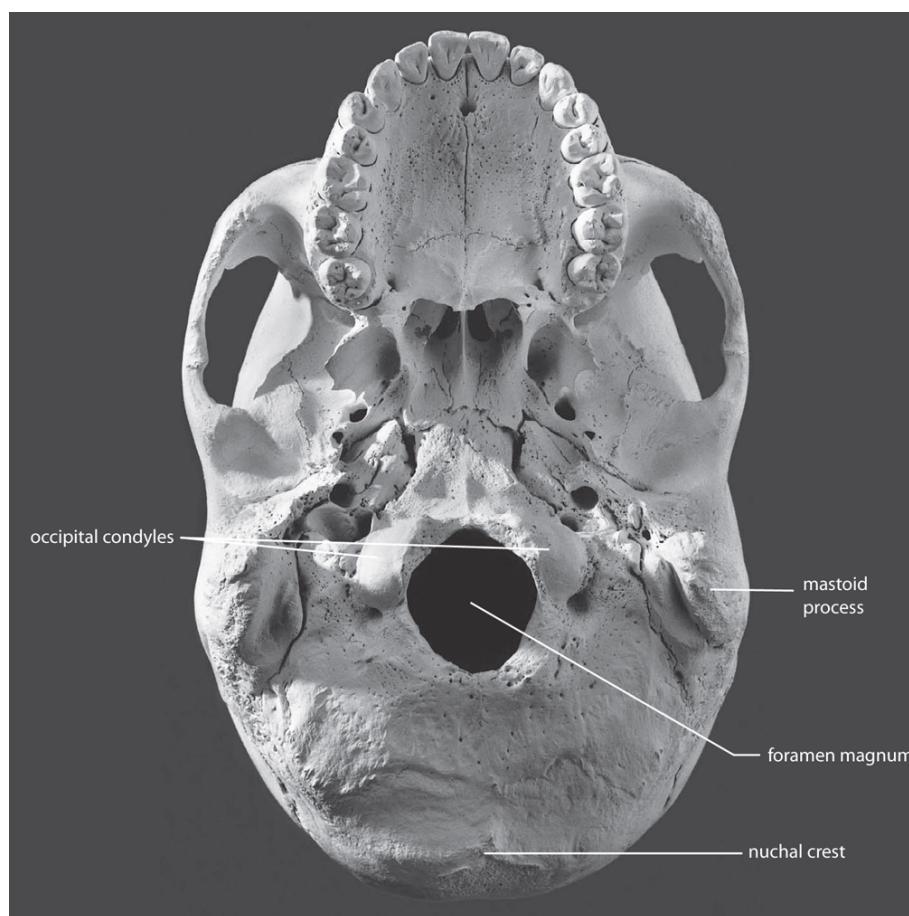
Figure 1.3: The posterior view of the cranium showing the cranial base. The foraman mag-num, occipital condyles, mastoid process, and nuchal crest show various degrees of sexual dimorphism depending on the population to be studied. Despite this, the mastoid process has been shown to fare better as a sex indicator than other features in the cranial base. Adapted from: White and Folkens 2005, pg. 81.

Unlike the foramen magnum and occipital condyle, the mastoid processes and rugosity of the nuchal crest are affected by the size of the muscles in the craniofacial region and in

the neck, which prompts bony growth at the insertion regions (Gapert et al. 2013). As males have more testosterone which promote muscle growth (Sutter 2003), they tend to have larger mastoid processes and a more rugged nuchal crest than females. The mastoid processes and the nuchal crest are therefore affected by factors other than differential developmental processes between males and females. Physical activity and lifestyle can affect muscularity such that athletic females may appear more masculine in these traits, and gracile males may be mistaken for females. Despite this potentially confounding variable, Williams and Rogers (2006) determined in their study that mastoid size was one of the best indicators of sex, since it was able to be used to predict sex with over 80% accuracy while the associated intraobserver error was less than 10%, although population differences do exist. Conversely, although a scoring system does exist for assessing the nuchal crest (Buikstra and Ubelaker 1994), Rogers (2005) and Williams and Rogers (2006) instead combine nuchal crest rugosity with a general assessment of the overall robustness or gracility of the skull. This approach is sensible because the cranial base is less sexually dimorphic than other regions, and assessing this region alone has little value if it is not compared to the rest of the cranium (Nikita 2014).

The differences in the splanchnocranium, i.e. the facial region of the cranium (White and Folkens 2005), exist due to the downward and outward growth that males undergo during adolescence (Rogers 1999). The nasomaxillary region in particular undergoes significant growth (Buschang et al. 1983; Humphrey 1998), meaning that as the maxilla elongates along with the orbital floor, the orbits become too large relative to the eye and its associated structures. To compensate for this enlargement, the orbital floor deposits bone in order to maintain an appropriate size (Enlow 1982), such that the orbits appear high in males relative to the rest of the face. Another consequence of this bone deposition is that the distance between the orbital floor and the nasal floor increases almost by a factor of two by the time the nasomaxillary growth is complete (Enlow 1982), creating a nasal cavity that appears longer and narrower in males compared to females. The nasomaxillary region, however, is also one of the primary regions used in ancestry assessment as variation has been noted according to ancestry in many studies (e.g. Enlow 1982; Mo 2005; Weinberg et al. 2005; Frelich and Hunt 2007, etc.). Therefore, any features in the nasomaxillary region should exhibit population-specific sexual dimorphism, so an ancestry assessment is required to be performed before analyzing these features.

Changes in the nasomaxillary region also affect the zygomatics, which, in addition to

being displaced downwards with the rest of the maxilla, also continue to grow laterally (Enlow 1982). The longer and more intense craniofacial growth spurt that males undergo result in the zygomatics being larger than in females, and in the zygomatic arches being positioned more laterally (Rogers 2005). As such, the posterior root of the zygomatic bone may be continuous with the supramastoid crest as part of the temporal line (Keen 1950). Whether the temporal line extends past the external auditory meatus (labelled as the suprameatal crest in Figure 1.4) is therefore an indicator of whether an individual is male or female, although it is affected by the development of the temporalis muscle (Keen 1950). Consequently, St. Hoyme and Işcan (1989) advise that this feature is only useful in populations that exhibit sexually dimorphic musculature. It is therefore preferable to assess traits in the nasomaxillary region if the analyst has had extensive experience with the population to which the unknown individual belongs, in order to understand the variation that may occur within the population as well as any variables such as muscularity that may influence traits.
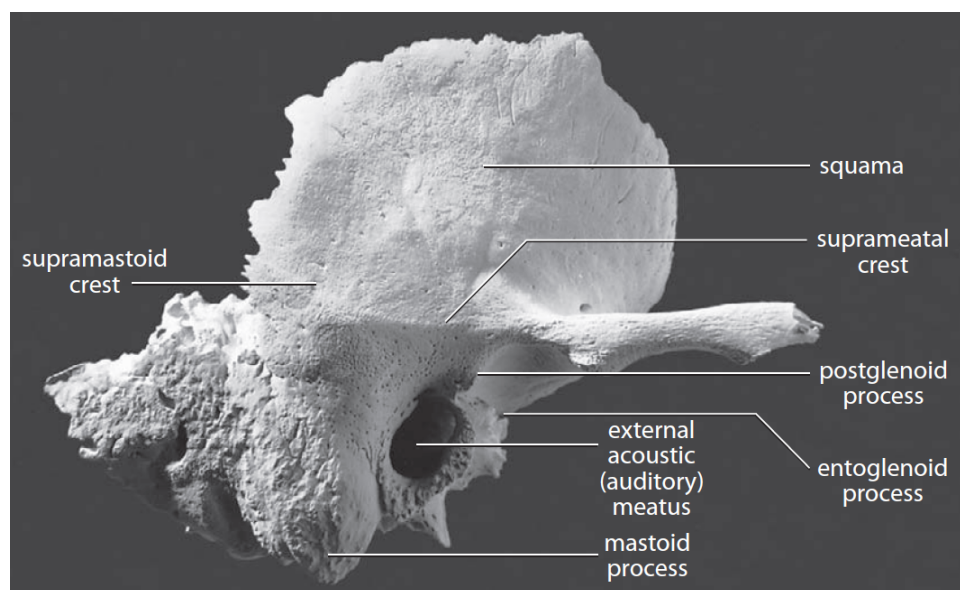
Figure 1.4: The temporal bone, showing an extension of the temporal line over the external auditory meatus (i.e. the suprameatal crest), which is a masculine feature due to increased muscularity. Source: White and Folkens 2005, pg. 96.

Using any one single trait to assess sex is never recommended, so although some traits may appear to be less sexually dimorphic than others, a combination of traits that are theoretically sexually dimorphic will give a more accurate combined approach (Williams and Rogers 2006). It is important to note that the varying success rates of each cranial trait for sex assessment are not only due to the inherent subjectivity of the analyst, but also due to the way in

which traits are assessed. For example, the glabella has been recognized as one of the most useful visual traits for distinguishing between males and females, but metric studies using the glabella have obtained low accuracy rates not much higher than chance (e.g. 57% accuracy rates were reported by Nikita (2014)). Similarly, mastoid size is well-accepted in the literature to be sexually dimorphic, but attempts at quantifying the mastoid triangle (i.e. by calculating either the 2D area or 3D volume between three craniometric points - asterion, porion, and mastoid) have had varied and mostly unsuccessful results (Kanchan et al. 2013). The accuracy rates therefore only partially reflect the trait's usefulness in indicating sex, and is heavily dependent upon whether it is assessed morphologically or metrically. If metric, the mathematical models used to determine what values are male and female hugely impact the results, as many studies seem to use mathematical models to find something that "works", instead of fitting the data to a model that actually represents the phenomenon that is being studied. For example, Walker (2008b) tested multiple types of discriminant function analyses to assess ranked ordinal scores of five morphological cranial traits, and concluded that logistic regression discriminant function analysis was the best option in order to minimize misclassification and sex biases. Such a brute force approach (in computer science terms) is not based on theoretical models, and consequently, interpreting results from such approaches are extremely limited since they do not actually reveal anything about the underlying phenomenon being studied. This research project foregoes the decision as to what mathematical models should be used for classification, and instead uses an approach based in machine learning that is appropriate for solving classification problems. In this way, the classification method in this project is not subject to the assumptions and limitations of function-based analyses. Furthermore, by using an approach that has been explicitly developed for the purpose of classification, the results are not at risk of being interpreted beyond the limits of the analytical method.

1.3   3D Methods of Analyzing Bone

With the boundaries of technology continually expanding, numerous methods of 3D data analysis have been explored for sex assessment. A review of the literature has revealed three major fields of 3D data analysis in osteology: 1) landmark-based methods; 2) comparative shape-based analyses; and 3) the creation of averaged 3D models to determine the representative male and female shapes to be used for comparison, which is an extension of shape-based

analysis. Each of these approaches carry significant advantages and disadvantages due to the various methods in which 3D data are acquired, the sampling of 3D data, and the analyses themselves. It should be noted that these three categories of methods are primarily based on the data input, which in turn govern the type of analysis to be used. Conversely, machine learning analytical methods can be applied to a wide variety of data (Mitchell 1997; Goodfellow et al. 2016). Although there have been recent studies in osteology that have utilized some of the machine learning algorithms (which are discussed below in more depth), this project focuses explicitly on deep learning applied to classification, which to the author's knowledge has not yet been applied to 3D models of bone. In the absence of a well-established body of literature on deep learning applied to 3D methods of analyzing bone, this section will instead explore the three major fields of 3D data analysis in osteology to date and outline the advantages and limitations of each. This section will conclude with a literature review of machine learning algorithms (not necessarily deep learning) applied to osteological analyses in order to highlight how machine learning can address the limitations posed by traditional methods of analysis.

1.3.1 Landmark-Based Methods

Landmark-based methods use a combination of several craniometric points in order to capture statistically significant shapes. All landmark-based methods which require a manual input suffer from unavoidable inter- and intra-observer error (Sholts et al. 2011), which differ according to the method used to measure the data. For example, the error associated with taking measurements on a skull between craniometric points will depend on the analyst's training and ability to locate landmarks, as well as the analyst's ability to correctly utilize sliding calipers and/or spreading calipers.

The most well-known landmark-based program for cranial sex and ancestry assessment is FORDISC (currently v.3.1) created by Jantz and Ousley (2005). For the purpose of creating FORDISC (2005), researchers from all across America pooled craniometric data from modern individuals from various ethnicities, although most of these ethnicities are from American populations. This database forms the basis for FORDISC (2005), which is aptly named after "Forensic Discriminant Functions", since it uses the database content to create made-to-order discriminant functions to assess sex and ancestry. A researcher wishing to use FORDISC (2005) must first input as many craniometric measurements that are available on the skull.

Based on which craniometric measurements are inputted, FORDISC (2005) will create a discriminant function from the database content, and categorize the skull in an ethnic-specific sex category. The output of the program also includes the accuracy of this function to be able to categorize skulls in the database, as well as the posterior probability and typicality probability. The posterior probability is the strength or significance of the resulting categorization of the sample, since it gives the probability of correct categorization (Jantz and Ousley 2005). The typicality probability is how similar the sample is to those in the database, based on the inputted craniometric data (Jantz and Ousley 2005).

The advantage to FORDISC is that it can be used to assess fragmented or damaged skulls, since the program can run properly even if only one measurement is inputted. Conversely, there are major limitations to FORDISC. Categorization of a sample into an ethnic-specific sex category occurs even if the group to which the sample truly belongs does not exist in the FORDISC database. There is no way of knowing whether a sample is actually correctly categorized other than by interpreting the typicality probability, which indicates whether a skull seems to be different than those in the database. The typicality probability, however, does not indicate whether a sample is atypical merely because they are at an extreme end of a male or female category (i.e. if they fall into the tail end of a probability distribution), or if it is because the sample truly does not belong to any of the given categories. Additionally, FORDISC uses 2D measurements taken from craniometric points without attempting to interpret shape, which in turn limits the ability to interpret the output.

More recently, Ross and Slice (2014) developed 3D-ID which addresses some of the limitations of FORDISC. As the name suggests, 3D-ID uses a 3D coordinate database of cranial measurements. Unlike FORDISC, 3D-ID creates a statistically-significant shape from inputted 3D coordinates of craniometric points. The program then analyzes the 3D shape created by the inputted coordinates, and categorizes the sample into a sex and/or ancestral group. Unlike the FORDISC database, the 3D-ID database contains 3D coordinate craniometric data from multiple different skeletal collections from around the world, so the representation is more global than that of FORDISC. The output of 3D-ID allows a deeper interpretation of the results since it involves shape analysis. This output also includes, for each available sex/ancestral group: the Mahalanobis squared distance, which is a calculation of how closely related the unknown sample is to the centroid or mean of the group; the posterior probability; and the typicality

probability. The two probabilities are defined in a similar to way to those in FORDISC. The addition of the Mahalanobis squared distance is useful because this value does not assume that the sample actually belongs to a group, and demonstrates how far away the sample is from the mean of the group. Therefore, where the sample falls in terms of standard deviations away from the mean can be interpreted.

Although 3D-ID expands upon the premise of FORDISC, it is similarly limited in that only coordinate data can be inputted, and only for the pre-defined craniometric points that are listed in the program. This is restrictive, as demonstrated by Coquerelle and colleagues (2011), since there are some statistically-significant shapes that cannot be captured by standard craniometric points. For example, Coquerelle and colleagues (2011) showed that by using semi-landmarks, i.e. taking coordinate points at regular intervals along a surface, a better representation of the geometric shape of the sample can be achieved. In turn, this approach can reveal shape-related growth differences that are not seen when using only craniometric points. Therefore, there is value in exploring point cloud data in which the entire geometry of the cranium is represented, such as in this project.

The 3D-ID program is based on geometric morphometrics (GMM), which is defined as a method of "analyzing and visualizing shape variation in the absence of size differences among specimens" (McKeown and Schmidt 2013). GMM is a powerful method of analysis since it can interpret shape-related variation separately from size if required, so size differences which may otherwise confound shape similarities can be isolated. The literature pertaining to GMM for adult sex assessment is almost exclusively based on the skull (e.g. Kimmerle et al. 2008; Shearer et al. 2012; Nikita 2014; Holton et al. 2016), and it has been noted that although absolute size differences exist between male and female skulls (Rogers 2005; Moore 2013), size should not be used for distinguishing between the sexes. Instead, cranial traits and features should be compared to the overall size and shape of the cranium for assessing sex (Nikita 2014). GMM is therefore best suited to ignore size differences and instead focus on shape-related differences in the skull.

To create a geometric shape for GMM analysis, craniometric landmarks or semi-landmarks are digitized and their position in 3D space is recorded. Within a group or category, these digitized coordinates are superimposed onto each other. As seen below in Figure 1.5, image (A), the coordinate data are not oriented the same way for each sample, and size also precludes the

ability to meaningfully interpret these coordinates. In order to orient, scale, and locate the coordinates with respect to one another, the Procrustes superimposition is used (Adams et al. 2004; Slice 2005, 2007; Mitteroecker and Gunz 2009). The result is seen in Figure 1.5, image (B), where clusters for each landmark or semilandmark exist. Most importantly, size differences between each of the samples are excluded, so shape differences account solely for the variation seen for each landmark or semi-landmark. From these clustered data, a shape must be created that represents the variation seen in this group. There are different mathematical approaches to create this shape, and the most common approach is calculating the Mahalanobis squared distance (McKeown and Schmidt 2013; Slice and Ross 2014). Regardless of which method is used, they all involve representing a landmark or semi-landmark solely based on the mean and variance of each cluster. The result is seen in Figure 1.5, image (C), where one group is represented by the shape created by the black circles and another group is represented by the shape created by the white circles. Two groups - in this example, an African-American group and a Euro-American group - can therefore be superimposed and fitted to each other using a regression method of one's choosing such that the difference between the two groups can be calculated statistically.
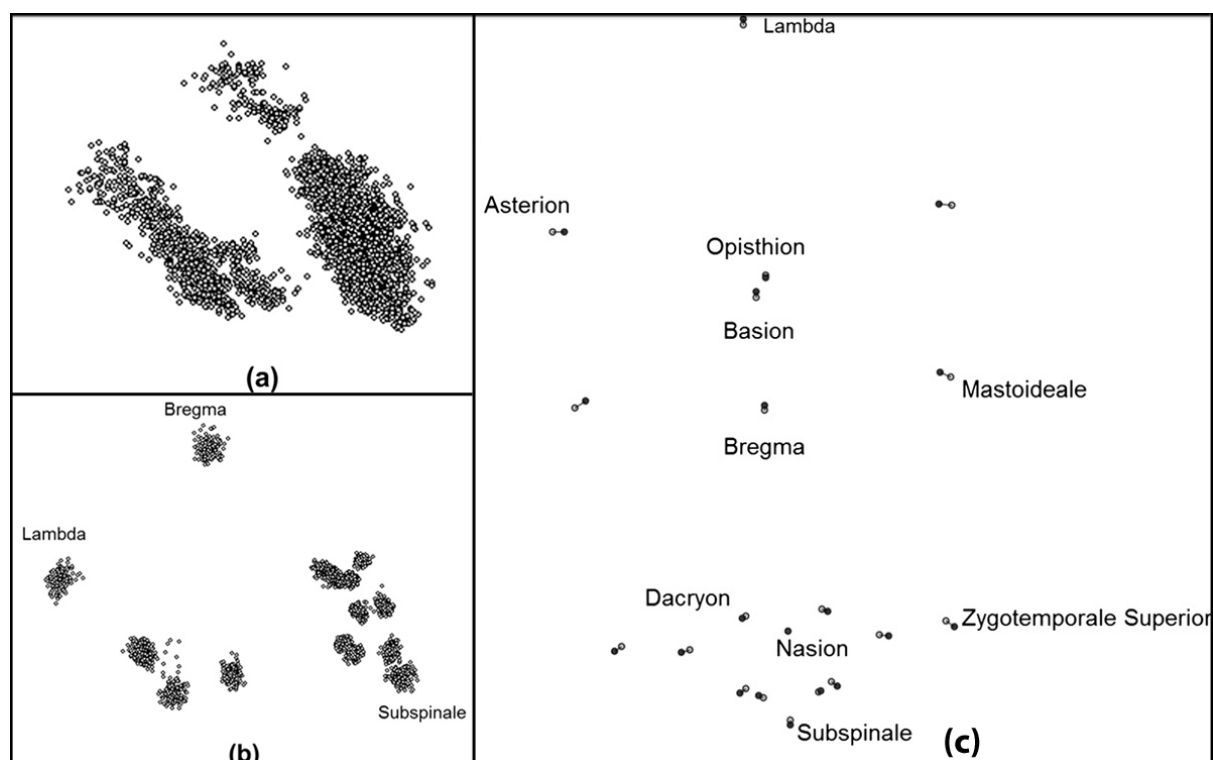
Figure 1.5: The processing steps in generating a statistically significant shape using GMM. (A) 3D coordinates from many different samples superimposed on top of each other, although the orientation and scaling of each sample has not been standardized yet. (B) The 3D coordinates from all samples have now been standardized using the Procrustes superimposition. The result is that coherent clusters are now visible, with each cluster corresponding to different craniometric landmarks. (C) From each cluster, a representative point is chosen to create an overall shape composed of different points. The black points represent one group of samples while the grey points represent another group. The differences between these two shapes therefore represent differences between the two groups. Source: McKeown and Schmidt 2013.

There is no arguable criticism for applying the Procrustes superimposition to the coordinate data, as this is a simply a way to scale and orient the data in the same manner. The main issue, however, is situated in the creation of a "statistical shape" from the clustered data. Explanations of how this is done mathematically in osteological GMM are not very clear, but all studies seem to use only the mean and the variance of each cluster (McKeown and Schmidt 2013). It is therefore uncertain if this approach assumes a normal distribution, which may be an inappropriate assumption if there is no evidence of a normal distribution; conversely, if this assumption is not made, there is still the limitation of only using mean and variance to define a representative point for each landmark. By the very nature of finding a representative point for each landmark, which is further limited by only considering mean and variance, GMM analyses do not and cannot take into account the full range of variation that may exist in the samples themselves. Additionally, Bulut and colleagues (2016) point out that since landmark-

based studies do not use the entire geometry, the ability for researchers to understand cranial traits is inhibited. The fact that not all shape information is used prompts researchers to choose different combinations of landmarks to create statistically significant shapes, resulting in a lack of standardization and agreement on which landmarks to use, and how many. Finally, the major epistemological issue with GMM research is that in studies that are unable to successfully classify individuals, it is unclear whether the failure is due to: 1) the fact that not enough landmarks or semi-landmarks are included to create shapes that are statistically significant; 2) if the landmarks or semi-landmarks used are actually not useful at all; 3) the improper imposition of mathematical models used for categorization; or 4) any combination of these. Therefore, while GMM is extremely prevalent and useful in the literature for understanding size and shape changes, such as in longitudinal or cross-sectional growth studies (e.g. Coquerelle et al. 2010), using GMM analyses to classify individuals is problematic.

1.3.2 Comparative Shape Analysis

The second group of 3D data analyses is comparative shape analysis not based on landmarks, meaning that the actual geometry of the sample (i.e. an entire 3D model of a bone) is used rather than an abstract shape such as in GMM. As GMM analyses dominate the literature, studies using comparative shape analyses are relatively few, and have only recently been published. An example of a comparative shape analysis is provided by Bulut and colleagues (2016), where 3D models of frontal bones were generated from computed tomography (CT) scans. The 3D models were then aligned to a 3D sphere, and the distance between the two were established in order to quantify the roundness of the frontal bone. This was done to establish if there is a statistically significant difference in the frontal bone curvature between males and females. Such an approach is advantageous since the entire frontal bone was used, meaning that all of the available information about the geometry of the bone was included in the analysis. Additionally, the "sphere overlap method" is completely landmark-free, and therefore avoids inter- and intraobserver errors associated with landmark identification (Bulut et al. 2016).

Before addressing limitations associated with landmark-based methods, however, the issues with the study in question should first be noted. The chosen methodology in Bulut and colleagues' study (2016) in order to quantify roundness was incorrect. By comparing a 3D

model to a spherical model, roundness cannot actually be quantified - what is actually quantified is the distance between the 3D model and the spherical model. This distance may vary between males and females, and thus is an acceptable premise to explore whether the frontal bone is sexually dimorphic in this way. Bulut and colleagues (2016) chose to extend their study to challenge the common conception that the frontal bone is rounder in females than in males, and this is an interpretation that simply cannot be made based on how the results were obtained. Furthermore, when aligning the 3D frontal bones to the sphere, Bulut and colleagues (2016) used a "best-fit algorithm", which is not described. It is therefore unknown what parameters were taken into account when aligning the two models, and the reliability of such an approach is questionable. Finally, Bulut and colleagues (2016) note that another limitation in the overlap method is that it is affected by the accuracy of the 3D models themselves; however, this is not further discussed. The accuracy of 3D models of bone is therefore addressed and discussed in Chapter 1.4 (Generating 3D Models of Bone), and is also demonstrated in Chapter 4 (Examining the Properties of 3D Models).

Although the execution of Bulut and colleagues' study (2016) was flawed due to method-ological and interpretation issues, the premise of attempting a landmark-free approach using the overlap method is not. The study rightfully demonstrates the advantages to such an ap-proach - it is quantifiable, does not rely on human measurement, and uses the full shape infor-mation from the bone. Furthermore, if alignment algorithms are used and properly explained, this method can be repeatable. As demonstrated in Bulut and colleagues' study (2016), limita-tions to interpreting results from this approach do exist. Comparative shape analysis can only demonstrate that there is a quantitative difference between the bone and the geometric model to which it is compared; it cannot speak to the nature of this difference (e.g. whether or not the difference is due to a difference in roundness). To investigate the reason for shape differences, GMM is much better suited, especially the use of semi-landmarks on a curved surface if round-ness is to be quantified. It is therefore the aim of this PhD project to use the full geometry of crania in order to avoid biases from human assessment to create a method for sex assessment. The results of the method are then compared to the results of human assessment in order to exemplify the advantages and disadvantages of both.

1.3.3 Averaged Representative 3D Models

Expanding on the premise of comparative shape analysis in which whole 3D models are used, and similar to the way in which GMM models use a mathematical method to determine representative points to create a shape, it has been proposed in several recent research endeavours to create representative 3D meshes to which to compare unknown samples. A mesh is created from a point cloud, which is simply a collection of 3D coordinates or points, and a mesh is formed by connecting these points to form polygonal surfaces. A representative 3D model for a group can be created by averaging the point clouds of each sample, and unknown samples can be compared to this representative model in order to determine if and how the sample differs (Furmanová et al. 2017).

Luo and colleagues (2013) have applied this premise to sex assessment using 3D meshes of skulls, and a Master's research project at the University of Toronto Mississauga also averaged 3D meshes to create representative male and female models for assessing the distal humerus for sex (K. Fleming, *personal communication, January 15, 2014*). The problem with averaging meshes is that the natural biological variation that exists between individuals is assumed to be distributed normally, and all traits/features are averaged in the same way, with the same weighting. Even if other mathematical models were to be used to generate a "representative" model, some kind of distribution must be assumed and there is currently no way of knowing what mathematical model can best approximate natural biological variation. The assumption as to what distribution best approximates this variation is the major, but unavoidable, limitation to using averaged representative 3D models. This is similar to the limitation with GMM shapes, but the error associated with averaged meshes is compounded many times more, since a mesh can be composed of thousands of points, each of which have to be averaged, whereas a GMM shape is limited to the number of craniometric landmarks and semilandmarks. Unfortunately, this limitation is not discussed in the literature, and results of studies using averaged representative 3D models continue to be overinterpreted. Finally, another limitation to consider is that the process of averaging entire 3D meshes or point clouds may result in a 3D model that exhibits features or a combination of features that do not exist naturally. It is therefore flawed to use such an idealized 3D model to represent an entire category.

The value in creating averaged 3D models perhaps lies in demonstrative or teaching

purposes, wherein for the sake of understanding a complex concept, oversimplified examples can be used. This is especially helpful in court where a jury is present, in order to explain basic differences between male and female skull morphology. Currently, however, there is little value in using averaged representative 3D models in research to investigate cranial sexual dimorphism, even if studies such as the one undertaken by Luo and colleagues (2013) claim the opposite.

1.3.4 Machine Learning in Osteological Analyses

Machine learning is a field in computer science that is dedicated to the study and implementation of algorithms that learn from data in order to perform a certain task (Mitchell 1997; Goodfellow et al. 2016), and encompasses many different algorithms such as decision trees, model-based clustering, and deep learning, the latter of which is the primary focus of this project. Deep learning builds upon the premise of Artificial Neural Networks (abbreviated as "ANNs"), which are modelled after the way humans learn by creating associations between acquired data (McCulloch and Pitts 1998). A single ANN consists of a sequence of nodes which are connected in a specified way, analogous to how neurons in the human brain are connected to each other. Two sets of data are typically used by deep learning algorithms - the training dataset and the evaluation dataset. When the training dataset is run through the ANN, the connection between the nodes are tuned (or "weighted") (Raschka 2015), similar to how synapses fire between neurons in humans in order to reinforce a concept. As such, the process of training the ANN involves running the data through the network for several iterations, which are also known as "epochs". At each epoch, the weights of the ANN are optimized in order to improve the accuracy of the model. Additionally, the dataset is usually divided into batches for two reasons: 1) since deep learning is usually applied to large datasets, it is not always feasible to run the entire dataset due to technological constraints; and 2) by splitting the dataset into batches, a degree of stochastic noise is added that can mitigate issues of overfitting the model to the dataset such that the model can be generalized. The accuracy obtained on the training dataset quantifies how well the model was able to perform a given task using the amount and type of data provided, and can therefore indicate the usefulness of the data to the given task as well as the model's theoretical suitability to perform the task. Once a model has been trained, the evaluation dataset is then used to test the performance of the model. The accuracy obtained

on the evaluation dataset indicates the model's ability to be generalized, since the evaluation dataset was not used in its training.

In the context of osteological analysis, classification is an issue that researchers have recently tried to address by delving deeper into the principles of machine learning. Classification in machine learning can be either supervised or unsupervised. Supervised machine learning means that the data to be classified has associated information associated to each sample (termed "labels") for the purpose of training the algorithm (Hastie et al. 2009). The categories in which the data is to be classified are therefore determined by the user, on which the algorithm is based. Conversely, unsupervised machine learning merely utilizes the input data to identify categories in which the data is classified. Both supervised and unsupervised machine learning have their merits in osteological analyses.

An example of supervised machine learning in osteology is the AncesTrees program, which was recently developed by Navega and colleagues (2015) for assessing ancestry based on metric analyses of the skull using decision trees algorithms. Although manual measurements similar to those in FORDISC were used for training this program, which suffer from human measurement errors, AncesTrees is unique in that it demonstrates how machine learning algorithms can be applied to classifying osteological data. When testing AncesTrees on a dataset that consisted of individuals from six ancestral groups, 75.0% of African individuals and 79.2% of European individuals were correctly categorized; when a model was created that only included these two groups, the performance increased to 93.8%. This study is an excellent example as to how the use of a machine learning tool can be used in osteology for classification.

It should be noted that the issue with using supervised learning is the fact that human biases exist as to how to annotate or categorize data. The annotation of data is not a problem if the associated information is ground-truth (i.e. the sex or population to which an individual belongs is known, and the bone that is being assessed is known to come from that particular individual). The associated information only becomes a problem if the information used by machine learning algorithms comes from assessed or estimated characteristics (e.g. if a machine learning algorithm were to be trained on a prehistoric dataset for which the sex of the individuals is given by an analyst's skeletal assessment), in which case, the resulting model cannot be as robust. A study by Algee-Hewitt (2016) best exemplifies the potential bias of using supervised learning in a practical application. Part of her study involved comparing human-determined

ancestral groups (i.e. ethnic groups since these groups encompass social implications) to the groups that the computer program identified using unsupervised, or unannotated, data. She found that there was no Hispanic group identified by the computer, which is a term that forensic anthropologists use to describe ancestry, especially in the United States. This finding supports the assertion that Hispanic is an ethnic group with heavy social connotations and a diverse biological basis (Algee-Hewitt 2016), and explains why forensic anthropologists have had such difficulty in defining a biologically meaningful Hispanic group. The use of unsupervised machine learning therefore has the potential to address classification tasks in a manner that is less influenced, if at all, by human annotation. For a more in-depth discussion into how unsupervised machine learning approaches are beneficial and appropriate for osteological analyses, refer to the paper by Trentin and colleagues (2018).

Recent studies have utilized machine learning in osteological analyses and have achieved very promising results which further demonstrates the need for machine learning tools to be made available to analysts. Examples include: Afrianty and colleagues (2014) who utilized the width, height, and thickness of patellae in order to assess sex using an ANN, and who achieved 96.1% accuracy; and Cavalli and colleagues (2017) who used the calvarium contour from CT scans in lateral view for ANN-based sex assessment and achieved accuracies up to 87%. Some studies have even compared the performance of machine learning algorithms to the traditional regression and multivariate methods (e.g. Curate et al. 2017) which is an important step in establishing machine learning as a justifiable and more robust tool for analysts. It is therefore the goal of this PhD project to further demonstrate the utility of machine learning in osteological analyses, and to advocate for its establishment in forensic anthropology.

1.4 Generating 3D Models of Bone

Traditionally, CT imaging has been used to create 3D images of bone, and a large body of research surrounding the use of CT images for age and sex assessments exists (e.g. Telmon et al. 2005; Sidler et al. 2007; Barrier et al. 2009; Ferrant et al. 2009; Grabherr et al. 2009; Decker et al. 2011; Villa et al. 2013). CT images are created by attenuating X-rays through an object, which create 2D cross-sectional images based on density. A 3D image is then formed by combining multiple 2D cross-sections (ASTM 1997). The advantage to using CT images is

that internal structures of an object can be visualized without destructive or invasive procedures (ASTM 1997; Schladitz 2011). Conversely, limitations exist in the resolution of the image, which is heavily dependent upon the physical scanning parameters and image processing algorithms. For example, a 2D Gaussian-like function termed the point-spread function is involved with the image-formation process (ASTM 1997). The point-spread function causes blurring in the CT image, which reduces resolution since small features will appear larger and sharp edges will be softened. Consequently, features or boundaries that are very close together might not be distinguishable. This can be a source of error when trying to take measurements from bone surfaces, because the boundary of the bone must be interpreted. Additionally, Villa and colleagues found that the texture quality of the resulting 3D model from CT scans is poor due to the thickness of the 2D slices used to create the 3D image, which can complicate morphological assessments that rely on bone texture (Villa et al. 2013).

Laser scanning has been explored to create 3D images of bone, and has been compared to CT scans to determine that laser scans do indeed produce higher quality 3D models than CT scans (Villa et al. 2013). Laser scanning can use light of any wavelength in order to measure the distance from the light source to the object. From this measurement, point clouds are generated that correspond to the external geometry of the object, and if the laser scanner has a camera incorporated into it, colour photographs can be taken in order to provide colour information to the 3D model (Sholts et al. 2010). Several scans of an object must be taken from different angles so that the point clouds can be registered or aligned to each other to form a complete representation of the object. The resulting 3D models are scaled and may have colour information. Sholts and colleagues (2010) tested the reliability of measurements taken from laser scan models of bone, and found that interobserver error was low at 2%. Laser scanning was therefore shown to generate higher quality and more reliable 3D models than CT scans. Limitations exist with laser scanning, however. Objects that have reflective or shiny surfaces, such as greasy bone, may cause the laser beam to be reflected off of the surface without returning to the scanner, or it can be reflected off of several surfaces before returning to the scanner with a weakened wavelength. Either of these scenarios results in unquantifiable error in the 3D model.

Photogrammetry has recently been investigated by researchers for documenting bone, due to its cost-efficiency and portability (e.g. Johansen 2014; Lam 2014; Saly 2014). To doc-

ument bones using photogrammetry, a camera is all that is strictly required, although some researchers choose to use light boxes to ensure an even lighting on the object, and rotating stands to facilitate the documentation process. A series of photos are taken of an object from multiple angles and heights, which allows the photogrammetry software to calculate the location of the camera relative to the object using multiple point triangulation (AgiSoft 2012). The purpose of multiple point triangulation is to understand the position of the object and the camera such that identical features in different photos can be aligned to each other using the Scale Invariant Feature Transform (SIFT) algorithm developed by Lowe (1999). SIFT uses the colour information in the photo for alignment, however, which means that if lighting is inconsistent across photos, either error is introduced into the resulting 3D model, or the alignment fails and a 3D model is unable to be created. Therefore, researchers do need to be trained in basic photography. It is important to note that not all photogrammetric software automatically scales objects, meaning that 3D models will need to be manually scaled, and thus may suffer from human error. Nevertheless, Johansen (2014) found that interobserver error was statistically insignificant when comparing measurements taken from actual crania to the corresponding measurements taken from the 3D models of the same cranium.

3D models of bone are advantageous because they provide an accurate method of documenting bone, in a manner which also allows intuitive interactions with the image to view the bone in whatever manner suits the analyst (Telmon et al. 2005; Villa et al. 2013; Lam 2014). 3D images of bone are also advantageous in cases where specimens in a collection need to be repatriated, since they offer a more holistic approach for documenting the collection compared to the limited notes, photographs, and test results that were undertaken prior to repatriation (Kakaliouras 2014). Similarly, sharing skeletal data is much easier since researchers can access virtual 3D models of bones rather than having to physically go to the skeletal collection. Due to the wealth of research potential that is associated with 3D models of bones, it is important that metric and morphological assessments of 3D bones are reliable.

While numerous studies have established that measurements are repeatable and reliable on 3D models of bones (e.g. Sholts et al. 2010; Decker et al. 2011; Johansen 2014), a persistent issue that has existed with 3D models is texture (Villa et al. 2013). This PhD researcher's Master's thesis (2014) investigated how the issue of texture could affect morphological assessments by assessing the auricular surface of os coxae for age using the Lovejoy

method (1985), which is heavily dependent upon bone quality and texture. The actual bone specimens were first assessed, before the 3D photogrammetric models were randomized and re-assessed blindly. This study found that there was no statistical difference in the accuracy obtained for either sample type, and that the same trends were demonstrated in the inaccuracy and bias charts between the two sample types. Therefore, the study concluded that there was no statistical difference in accuracy between assessing 3D models or bone specimens for morphological features, although it was noted that from a practical point of view, texture was extremely difficult to assess. It was also noted that correct assignment of 3D models to an age category was mostly reliant upon other morphological features. It was therefore asserted that methods developed on dry bone should be adapted for use with 3D representations. Villa and colleagues (2013) similarly state that skeletal assessments perform best on dry bone rather than with 3D models, although their study, along with others (e.g. Barrier et al. 2009; Grabherr et al. 2009) demonstrated that accurate morphological assessments are still possible using 3D models.

In conclusion, each method of generating 3D models has advantages and limitations in terms of time, cost, training, and protocol. The most important factor that all 3D modelling methods have in common, however, is that the results are repeatable and reliable. With the advancement of technology, the field of generating 3D models is continually improving, and new technology can be explored in research. This research project investigates one such new technology - Structured Light Scanning (SLS) - which is explored in Chapter 2 (Data Acquisition & Methodology).

1.5   Research Aims

The goal of this project is to create a new method of assessing skeletal individuals according to sex and ancestry/population using 3D cranial data, which will be facilitated by the use of machine learning. In order to achieve this, the following research objectives have been identified: 1) creating a 3D "ground-truth" cranial database consisting of modern individuals from various populations which will be used to generate and test the new method; 2) establishing the performance of current morphological assessments on correctly categorizing individuals from the database as a baseline; and 3) creating a proof of concept using machine learning to

assess the geometric shape of crania to determine the degree of sexual dimorphism globally, as well as within different populations.

1) Creating a 3D Ground-Truth Skeletal Database

A ground-truth database consists of samples of known origin, as opposed to samples with assessed or estimated characteristics. Such databases are important for extracting robust statistical conclusions based on the known characteristics of the sample, and therefore exist mostly for research purposes. A ground-truth skeletal database therefore consists of individuals with known biographic information, since the skeletons are often from donors or from cemetery burials with associated records. Issues with unknown or estimated age, sex, and other relevant information such as disease and/or cause of death are mitigated or eliminated completely depending on the completeness of the associated documents.

In this project, a 3D ground-truth cranial database was generated using structured light scanning. In order to populate this database, it was necessary to document several ground-truth skeletal collections from several geographically diverse institutions so that the variation of sexual dimorphism in different populations was captured. Due to the requirement of using individuals of known origin, most collections that satisfy this requirement are modern collections (i.e. within the last 100 years), with exceptions such as documented cemetery burials, which can date back to approximately the 1800s. It was ideal that samples included in this study did not exhibit pathological features that could obscure or alter sexually dimorphic traits on the cranium, which could skew the interpretation of such features when assessing sex and ancestry. By the same principle, good preservation of the specimens was also necessary.

To choose which skeletal collections were contacted for inclusion into this project, the following criteria were used. First, the skeletal collection should include male and female adults, and the crania should be available and in good condition (i.e. good preservation; few pathological expressions which would interfere with sexual trait analyses; little to no fragmentation). For the purposes of this study, adult individuals are defined as those who are 18 years of age or older at death, since differences between males and females are usually well-established at this point in development, and assessments of sex can be made with confidence (Braz 2009).

Collections that contained a large number of suitable samples were prioritized over smaller collections. Additionally, the most suitable collection for a given geographical region

was chosen. This project aimed to include individuals from European, Asian, and African populations in order to create a dataset with diverse populations. Originally this project also aimed to include North American individuals but it was not possible to incorporate North American collections within the scope of this PhD. A list and description of each collection that was documented are given in Chapter 2.1 (Skeletal Collections).

The scanning and acquisition of the raw point cloud data needed to be processed into coherent point clouds representative of the geometry of the crania. The final results then needed to be examined for their reliability and reproducibility. Establishing the reliability and reproducibility of 3D point cloud data on real datasets is not common practice, even though it should be in order to uphold scientific rigour. This PhD project therefore defines the terms "reliability" and "reproducibility" in the context of point cloud data, establishes a method by which to determine these two parameters, and determines the reliability and reproducibility of the point cloud data used in this project (see Chapter 4, Examining the Properties of 3D Models).

2) Establishing the Performance of Current Morphological Assessments

In order to have a baseline comparison to which to compare the performance of the machine learning method of sex assessment, the same individuals included in the 3D database were also assessed using morphological traits identified to be of high quality (Williams and Rogers 2006). It was necessary to perform these morphological assessments on the actual bone rather than the 3D models in order to better reflect the true performance of these morphological traits which were developed on dry bone; applying these methods to 3D models may not yield the same accuracy, as Villa and colleagues (2013) suggest and as ascertained in this researcher's Master's thesis (Lam 2014). By comparing the performance between the morphological assessments and the method generated using machine learning, insight can be gained into the morphological traits which are subject to the limitations of human perception, and improvements to the current existing morphological traits used can be identified. Chapter 3 (Cranial Sexual Dimorphism in Various Populations) presents the results of the morphological assessments on four skeletal collections from different populations, which are then compared to the performance of the machine learning method in Chapter 5 (Exploring Cranial Sexual Dimorphism with Deep Learning).

3) Establishing a Proof of Concept with Machine Learning to Generate a New Method of Sex Assessment

Although many features of the crania have been identified and repeatedly examined for sexual dimorphism (e.g. Buikstra and Ubelaker 1994; Williams and Rogers 2006; Gapert et al. 2009a; Humphries and Ross 2011; Bulut et al. 2016; Jung and Woo 2016; Krishan et al. 2016, etc.), readily available approaches that combine sexual dimorphism and population variation with 3D data are lacking in abundance in the literature. What little studies that do exist to address this issue focus on using craniometric points in combination with GMM (e.g. Ross et al. 2010; Navega et al. 2015). Although such studies are very valuable and also provide statistical methods for analyzing samples, the use of craniometric points with GMM analyses do not account for the full range of variation that exists simply because the full geometry of crania is not utilized. This project therefore uses point cloud data that do represent the full geometry of the crania in order to avoid potentially excluding features that are abstract to humans but useful for programs in classifying according to sex. Additionally, the use of GMM for classification is limited to creating a "statistical shape" from clustered data, which does not actually account for any external variation that may exist because only the center of the clusters are used to create a model. Conversely, machine learning offers more powerful methods of classifying geometric data because it attempts to create models that can account for stochastic noise (and are therefore applicable to external samples not involved in the making of the model) while having the ability to continually learn from the data. In Chapter 5 (Exploring Cranial Sexual Dimorphism with Deep Learning), machine learning is applied to 3D point cloud data in order to investigate sexually dimorphic traits and how these traits vary across populations.

The analyses using machine learning in this project are considered to be successful if the method created has an accuracy of at least 80% when tested on both the samples used to create the method and, more importantly, when tested on a 20% holdout sample dataset which is not part of the creation of the method. PointNet (Qi et al. 2016) is the machine learning (specifically, deep learning) algorithm used in this project. Due to the fact that it is quite new and therefore untested for its applicability to cranial point cloud data, the use of PointNet (Qi et al. 2016) is used as a proof of concept to establish how deep learning can be applied to classifying cranial point cloud data according to sex and population. Improvements as to how computer science and bioarchaeology can collaborate are identified and discussed in the context of how deep learning can be further applied to real datasets to help solve current research questions in bioarchaeology.

Chapter 2

Data Acquisition & Methodology

The data collected in this project can be categorized into two groups: 1) ordinal data from visual assessments of sexually dimorphic characteristics (discussed below in 2.2); and 2) 3D point cloud data representing the geometry of the cranial samples (the acquisition and processing are discussed below in 2.5 and 2.6, respectively). 2D photographic recording was also undertaken to supplement the data by providing a reference of the visually assessed characteristics and also to serve as another means of visually recording data by which to check the 3D point cloud data. The photographs, however, are not formally used in any analyses, since the analyses focus on the ordinal data and the point cloud data.

To acquire both types of data, cranial samples were used from four skeletal collections - St. Bride's Fleet Street Collection in London, U.K. (SB); Nagasaki University Modern Cadaver Collection in Nagasaki, Japan (NU); Milano Skeletal Collection curated by LABANOF in Milan, Italy (ML); and Pretoria Bone Collection in Pretoria, South Africa (PR) . The number of available crania differed at each skeletal collection. In cases where there were more than 150 crania in a collection, 75 male and 75 female individuals were randomly chosen for inclusion. This was done to optimize the amount of time spent at a collection and the time spent processing and analyzing the data. Due to differences in lab set-up, collection access, and how many crania were autopsied, assessing and documenting 150 crania took approximately between 100 - 150 hours. This translates into approximately 2.5 - 4 weeks of data collection per skeletal collection. It should be noted that the number of samples used for collecting the ordinal data (i.e. the number that were visually assessed and scored) differs from the number of samples that were

scanned and used to create point cloud data. The major reason for this discrepancy is that some samples could be visually assessed but were deemed unsuitable to be included in the ground-truth database due to issues of preservation or disease. Even though some of these degraded samples were not scanned, they were still visually assessed in order to increase the amount of ordinal data in this project. There was also one case in the PR collection where the point cloud data file was found to be corrupt, and the data could not be accessed or recovered. All samples included in this project have therefore been visually assessed and have associated ordinal data; however, not all samples have associated 3D point cloud data. The result is that this project encompasses 637 individuals (312 female, 325 male) with ordinal data, and 534 with 3D point clouds (263 female, 271 male; before processing of the point cloud data). The distribution of this sample size is given below in Table 2.1.

Table 2.1: The distribution of the samples according to population and sex in the ordinal dataset and the point cloud dataset.

Dataset	Population	# of Females	# of Males	Total #
Ordinal	SB	92	95	187
Total # = 641	NU	75	75	150
	ML	70	80	150
	PR	75	75	150
Point Cloud	SB	44	41	85
Total # = 534	NU	75	75	150
	ML	70	80	150
	PR	74	75	149

The age profiles of the samples used in the ordinal and point cloud datasets are given below in Tables 2.3 and 2.4, respectively, and have been categorized by age range into the categories denoted in Table 2.2. There is a heavy bias of older individuals across all collections, but this is an unfortunate and unavoidable limitation of most skeletal collections. The implication of this bias is that the resulting classification methods developed in this project will need to be applied to young adults (i.e. age categories 4 and below; 59 years of age and younger) with caution, especially considering that the machine learning algorithms will inherit this bias during their training.

Table 2.2: The age categories and their associated age range used for trait-specific age and sex analyses.

Age Category	Age Range (years)
1	18 - 29
2	30 - 39
3	40 - 49
4	50 - 59
5	60 - 69
6	≥ 70

Table 2.3: The distribution of the samples according to population and age category in the ordinal dataset.

Population	Age Category					
	1	2	3	4	5	6
SB[1]	23	24	20	29	52	38
NU	5	6	23	17	39	60
ML	3	7	7	8	21	104
PR	9	9	15	25	20	72
Total #	**40**	**46**	**65**	**79**	**132**	**274**

[1] One individual does not have a known age, and is therefore excluded from this table

Table 2.4: The distribution of the samples according to population and age category in the point cloud dataset.

Population	Age Category					
	1	2	3	4	5	6
SB[1]	7	10	11	14	23	19
NU	5	6	23	17	39	60
ML	3	7	7	8	21	104
PR	9	9	15	25	20	71
Total #	**24**	**32**	**56**	**64**	**103**	**254**

[1] One individual does not have a known age, and is therefore excluded from this table

Choosing which crania to include in this project was usually done before visiting the collection and before the analyst was able to see the samples (with the Milano collection as the only exception; this is discussed below in 2.1.3 (Milano Skeletal Collection - Milan, Italy (ML)).

Consequently, out of the 150 samples chosen for inclusion, some were excluded once issues of preservation become clear upon seeing the sample. Therefore, instead of simply creating a pool of 150 samples, all samples in a collection were first divided into groups according to sex, and then each sample in each group was numbered according to a randomly generated, non-repeating sequence. The first 75 in the male and female groups were pooled, and assessed in order of their given sample number (i.e. as determined by how the institution numbered them). The samples remaining in each group starting from number 76 onwards acted as a reservoir. If any of the 150 chosen samples were deemed unsuitable for assessment, the analyst requested a colleague to look up the sex of the unsuitable sample and choose a replacement sample from the appropriate reservoir group. This ensured that the number of males and females included in the data collection were as equal as possible, and also ensured that the analyst remained unbiased when assessing the samples, since the sexes of the samples were not revealed to her.

2.1 Skeletal Collections

The following skeletal collections were chosen to represent an overall diverse sample for the ground-truth database. While it is impossible to be able to properly represent each and every population that exists within the scope of this PhD project, broad geographical regions are included for diversity. The curator of each collection was contacted, and provided with a brief explanation of the project. Before the commencement of data acquisition at the institution curating the collection, ethical approval was first obtained from the University of Leicester (refer to Appendix A) and, if required, from the institution to be visited.

2.1.1 St. Bride's Fleet Street Collection - London, United Kingdom (SB)

The St. Bride's Fleet Street Skeletal Collection (to be referred to as 'SB' in this document) comprises of cemetery burials of 227 individuals (213 adults, 14 juveniles) who died between 1740 - 1852 (Gapert et al. 2013). There is some indication that the Fleet Street individuals were of middle socioeconomic status, as the lower churchyard is deemed to have contained individuals of low status, and the high status individuals were kept in crypts. Due to the associated coffin plates, it is possible to identify the individual in each burial, and thus, this

collection is appropriate for inclusion into this project, since sex and age are known. A few exceptions exist, where certain burials overlapped or shifted such that the excavated individuals do not seem to fit the identities given by the coffin plates. These individuals are noted by the curator of the collection, and were excluded from this study.

This collection was included to represent British European individuals, although it must be noted that it is a 19th century collection and may therefore exhibit secular changes. It has been suggested by researchers that the cranium has become taller and narrower in the past two centuries, based on craniometric measurements (Jantz and Jantz 2016) and morphological assessments (Kilroy and Tallman 2019). It should be noted, however, that Kilroy and Tallman (2019) found that the nasal bone contour and the loss of a postbregmatic depression in females were the only cranial traits to show a sexual bias when comparing 19th century collections to collections that contain 20th century and 21st century individuals. It is therefore expected that secular changes will not significantly affect the sex or ancestry classifications performed in this project for two major reasons: 1) in the visual scoring portion of this project, the nasal bone contour is not a trait that is assessed; and 2) the application of machine learning for classification is expected to take into account variations that are as minor as those reported by Kilroy and Tallman (2019), especially since the algorithm will be trained using a mixed pooled sample from SB and other skeletal collections.

2.1.2   Nagasaki University Modern Cadaver Collection - Nagasaki, Japan (NU)

The Nagasaki University Modern Cadaver Collection (NU) is composed mostly of donors who died between 1950 and 1970 (Tsurumoto et al. 2013), although this collection is still being expanded. As of 2016, there are 138 females and 232 males in the collection. Seventy-five males and 75 females were chosen by the lab technician so that the analyst was blind to the selection, although the numbering of the skeletons are according to sex (i.e. the skeletons numbered 1-232 are male; 233-370 are female). The sample number was censored during the assessment process using sticky notes to hide the number written on the bone, but it was sometimes impossible not to see the number if it was written on or adjacent to a surface that needed to be assessed, such as the nuchal crest on the occipital.

The assessment protocol was altered to account for the lack of space in the lab and accessibility to the collection. The analyst and her assistant were not allowed to access the

actual collection themselves; instead, a lab technician fetched the samples and brought them into the lab room. Furthermore, the space only allowed for 15 crania to be out at once. Due to these two constraints, the 15 crania (consisting of both males and females) were first assessed, scanned, and then randomized by the analyst's assistant so that the analyst could perform the re-assessment in a different order. Once completed, the 15 crania were packed away in their boxes for the lab technician to put away, and the next 15 crania were brought in.

2.1.3   Milano Skeletal Collection - Milan, Italy (ML)

LABANOF (Laboratorio di Antropologia e Odontologia Forense) is an organization associated with the University of Milan, and has a partnership with the City of Milan to create a modern documented skeletal collection for research purposes, especially relating to forensic anthropology. The Milano collection (ML) consists of more than 1700 skeletons which were exhumed 15 years after coffin burials. Both males and females are represented in this collection, and range from 28 years old to 103 years old (Cappella et al. 2016). Although this collection contains more than 1700 individuals, only about 400 skeletons have been cleaned and are available for study as of 2017. Furthermore, since backhoe excavators were exclusively used for exhumations, the skeletons have suffered much post-mortem damage. While at LABANOF, 460 skeletons were visually assessed for their suitability for inclusion into this project based on the priority criteria (see Table 2.8), and whether or not they were cleaned of dirt and mould. Due to the extensive post-mortem damage primarily caused by excavation, and the fact that some were not cleaned, only 150 skulls were available for study. All 150 samples were therefore included in this project, and consist of 80 males and 70 females. Due to the limited number of cleaned skulls in good condition, it was unavoidable to have uneven sample sizes between males and females. Furthermore, the difficulty in locating samples in various rooms across campus, as well as the fact that it was only possible to access certain areas at certain times, prompted an approach similar to the one used in Nagasaki to be adopted wherein 24 samples were assessed, scanned, randomized, and re-assessed before packing them away and retrieving the next set of 24 samples.

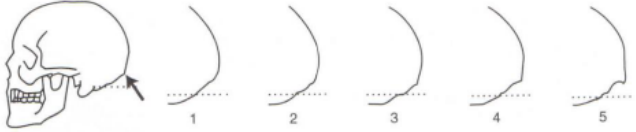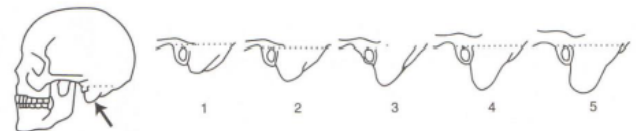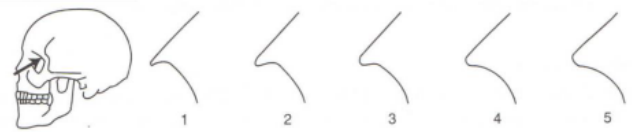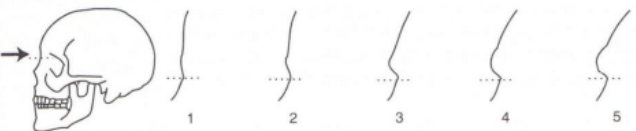### 2.1.4   Pretoria Bone Collection - Pretoria, South Africa (PR)

For African individuals, the Pretoria Skeletal Collection in South Africa was chosen to be included as it includes 704 male and female skulls that are well-preserved. This collection began in 1942 and has since been reorganized to promote its usefulness to researchers (L'Abbé et al. 2005). The 150 individuals that were assessed in this collection are primarily divided into two "ancestral" groups - "Black" and "White". These terms are widely used in South Africa to refer to one's ethnic group, both socioculturally and in the academic environment; however, it is very common to use these terms to refer to "ancestry" in both South African society and in the South African academic community, which is why "ancestry" is put in quotation marks. The definition of who is considered "Black" and "White" is based upon both physical and social criteria, such as: whether one's hair is thick enough to hold a pencil without it falling (a "Black" characteristic); the slope of one's forehead; and whether one is a foreigner or not (all foreigners are considered "White") (L'Abbé 2017, *personal communication*). There is therefore a high potential for biological variation within the "Black" and the "White" groups to the point where it is difficult to use these terms alone to narrow a missing persons list for forensic purposes. Additionally, there is also the fact that there is a clearly defined social dichotomy between the "Blacks" and the "Whites" in South Africa, meaning that relationships between members of different groups are virtually non-existent. The Pretoria Bone Collection is thus comprised of two (at the very least, if not more) genetically, socially, and culturally diverse populations that co-existed in one geographic area. For this reason, the analyses for this collection (provided in Chapters 3 (Cranial Sexual Dimorphism in Various Populations) and 5 (Exploring Cranial Sexual Dimorphism with Deep Learning)) include an examination according to "ancestral" groups to discern patterns that may aid forensic anthropological endeavours.

2.2   Performing Sex Assessment

To obtain the ordinal data for this project, it was necessary to identify methods of sex assessment that were both applicable to this project's research questions, and established in the literature to be applicable to various populations. One of the most well-established methods of assessing sex using the cranium is given in Buikstra and Ubelaker's *Standards* (1994), and consists of four morphological traits. The scoring system, as well as the trait descriptions, are

given below in Table 2.5. The scoring system consists of five stages, from 1 - 5, where 1 is definite female; 2 is probable female; 3 is intermediate; 4 is probable male; and 5 is definite male. Due to the standardized scoring, Buikstra and Ubelaker's method (1994) has become widely used in both forensic anthropology and in bioarchaeology. These four cranial traits were therefore scrutinized in this project to assess their accuracy and applicability to different skeletal populations.

Table 2.5: The traits and scoring methods given by Buikstra and Ubelaker (1994) for sex determination. Images are adapted directly from Buikstra and Ubelaker 1994.
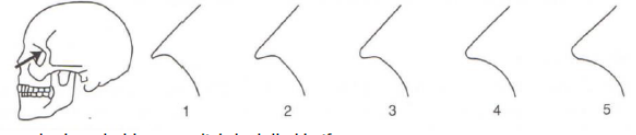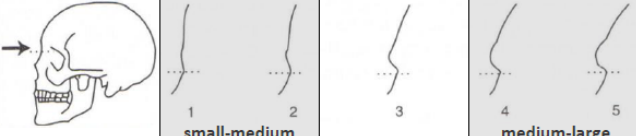

A more recent study by Williams and Rogers (2006) evaluated the accuracy and reliability of 21 traits on the skull. They established six "high-quality" traits, which were defined as having an intraobserver error of 10% or less, and an accuracy of assigning sex for 80% or more of the population on which they tested their method. These six traits are: mastoid process; glabella; zygomatic extension; nasal aperture; general size and architecture; and gonial angle. The mandible was not considered in this project, so the gonial angle was not assessed or included. Consequently, the five remaining traits given by Williams and Rogers (2006) were used, and are given in Table 2.6.

Table 2.6: The five high-quality traits and their descriptions as given by Williams and Rogers (2006).

FEATURE	MALE ♂		FEMALE ♀	
Size & architecture:	big/rugged	_____	small/smooth	_____
Supraorbital ridges:	medium to lge	_____	sml to medium	_____
Nasal aperture:	high, thin sharp margins	_____	lower, wider rounded margins	_____
Zygomatic:	extends	_____	does not	_____
Mastoid:	medium-lge	_____	sml-medium	_____

It should be noted that two of the five traits (mastoid process and supra-orbital ridge/ glabella) are also used in Buikstra and Ubelaker's method (1994), suggesting that these two traits have a very reliable and accurate distinguishing ability between the sexes. Unlike Buikstra and Ubelaker (1994), however, Williams and Rogers (2006) use a binary scoring system for assessing cranial and mandibular traits. While this approach potentially decreases inter- and intra-observer error (simply due to the fact that there are less options from which to choose), it is less precise when attempting to understand how a trait is distributed between the sexes (e.g. is the trait expression normally distributed in both males or females, or is it skewed, and if so, how?). Therefore, for the purposes of this project, the scoring system used by Buikstra and Ubelaker (1994) was preserved for all traits described in *Standards*. The two traits (i.e. mastoid process and supra-orbital ridge/glabella) that overlap with those given by Williams and Rogers (2006) also followed this scoring method, although Williams and Rogers' binary trait descriptions (2006) were used to describe the two male and two female scores. The remaining three traits (i.e. zygomatic extension, nasal aperture, and size & architecture) given by William and Rogers (2006) were assessed as either male or female, as originally intended, although an intermediate option was added to denote uncertainty if a trait was unable to be categorized as male or female. The resulting list, descriptions, and scoring of the seven cranial traits used in this project are given below in Table 2.7.

Table 2.7: The seven traits used in this project to assess sex, based on Buikstra and Ubelaker (1994) (given in white) and Williams and Rogers (2006) (given in grey).

Trait	Visualization & Score (circle)
Nuchal Crest *(Lateral profile)* Rugosity associated to attachment of nuchal musculature; ignore contour of underlying bone	1 = smooth, no bony projections visible in lateral profile 5 = massive nuchal crest that projects a considerable distance; well-defined bony ledge or hook
Mastoid Process **(Assess R & L)** Compare size with surrounding structures (e.g. EAM & zygomatic process); most important variable is volume of mastoid process, not length	small-medium medium-large 1 = very small process; projects a small distance below inferior margin of EAM & digastric groove 5 = length and width several times that of EAM
Supra-Orbital Margin **(Assess R & L)** Hold finger against margin of orbit at lateral aspect of supraorbital foramen; hold edge of orbit between fingers to determine thickness	1 = extremely sharp bolder, e.g. slightly dulled knife 2 = thick, rounded margin with curvature approximating a pencil
Supra-Orbital Ridge/ Glabella *(Lateral profile)* Compare with diagrams	small-medium medium-large 1 = smooth contour of frontal, little or no projection at midline 5 = massive glabellar prominence, rounded loaf-shaped projection (well-developed)
Zygomatic Extension **(Assess R & L)**	1 = does not extend past EAM 5 = extends past EAM
Nasal Aperture	1 = lower, wider, rounded margins 3 = intermediate 5 = high, thin, sharp margins
Size & Architecture	1 = small/smooth 3 = intermediate 5 = big/rugged

All of the specimens were photographed with a Nikon AW-1 camera, along with close-ups of the traits with a scale reference. These photographs served to check the associated 3D scans - for example, if a hole was seen in the scan, the photographs could be examined to understand why (e.g. the hole in the scan was an error due to the fact that the area of the cranium was stained black and therefore not captured by the scanner; or the hole is actually exhibited in the bone due to damage or trauma).

2.3   The Premise of Structured Light Scanning

In order to understand the 3D data used in this project, it is necessary to define several terms. "3D model" is a vague term that refers to any digital three-dimensional representation

that can be manipulated in 3D space. Fundamentally, a 3D model consists of a point cloud, which is a collection of points in 3D space each consisting of a set of 3D coordinates and possibly colour information (Pitzer 2015). Due to the fact that the definition of a point cloud includes 3D coordinate information, a point cloud is considered a type of 3D model. It is possible to use a point cloud to create a mesh, which is another type of 3D model. A mesh consists of creating triangles by treating the 3D coordinate information from a point cloud as vertices, and then connecting neighbouring vertices (which can be simply called "neighbours") together (Pitzer 2015). These triangles are commonly termed "elements" of a mesh. The result is that the shape of the 3D model is represented by something that resembles a wire mesh, with the size of the elements limiting the amount of detail represented. If desired, a texture can be applied to the mesh, which defines the appearance of the surface of the mesh using an image. Figure 2.1 below presents a simple visual comparison between a point cloud, a mesh, and a mesh with texture.
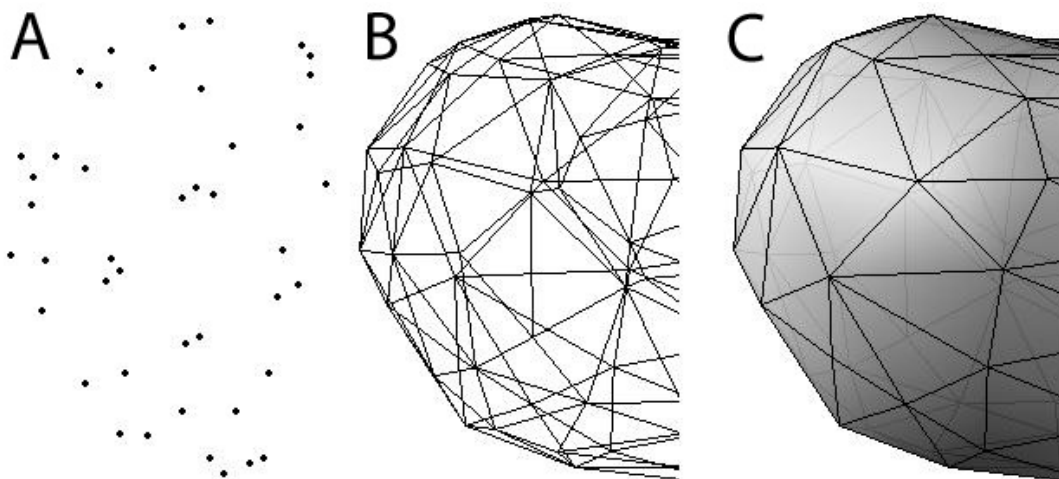

Figure 2.1: A) is a point cloud, consisting of points with 3D coordinates; B) is a mesh, in which the point clouds are used as vertices to create triangular elements; C) is a mesh with a texture (represented by the gray surface). Modified from Marjanovic 2007.

This project involved the use of meshes and point clouds, although the latter was used in the analyses for two main reasons. Firstly, most of the algorithms that were of interest in this project use point cloud data rather than meshes. The use of point cloud data actually seem to be more preferable within the computer software community as evidenced by the existence of an entire C++ library solely devoted to point cloud analysis (i.e. Point Cloud Library (PCL) (Rusu and Cousins 2011)). No such exclusive library exists for meshes as far as the author is aware. Secondly, point cloud data are subject to less interpretation and assumptions than meshes.

Point cloud data are simply 3D coordinates, whereas during the creation of meshes, different mathematical models exist in order to create the elements in a mesh (Pitzer 2015). These models are not always documented or readily available, so the use of meshes is accompanied by a degree of uncertainty in their creation. The generation of point clouds, as well as their subsequent subsampling for analytical purposes (which is discussed in Chapter 4), does carry some assumptions of the same nature, but these assumptions are compounded in meshes.

Structured Light Scanning (SLS) was used to create the 3D data for this project. SLS is a method of generating 3D models based on the distortion of numerous light patterns that have been projected onto an object (Liscio 2014). An SLS apparatus primarily consists of a projector and a camera mounted onto a bar such that the distance between the two is fixed. The object to be scanned is placed a set distance away from the SLS. The projector then projects a series of light patterns onto the object, such as stripes of different widths, which are distorted due to the surface of the object (DAVID-4 2017) as exemplified below in Figure 2.2. Simultaneously, the camera takes images of each distorted light pattern. The SLS program - which in this project is the DAVID 4 program (DAVID-4 2017) that accompanied the SLS - automatically creates a mesh that represents the external surface of the object, based on the series of light patterns and their distortions. An additional and optional step is for the projector to project red, green, and blue light onto the object to collect colour information, which is then applied to the mesh as a texture (DAVID-4 2017). From this mesh, it is possible to discard texture and element information such that only the point cloud data remains.
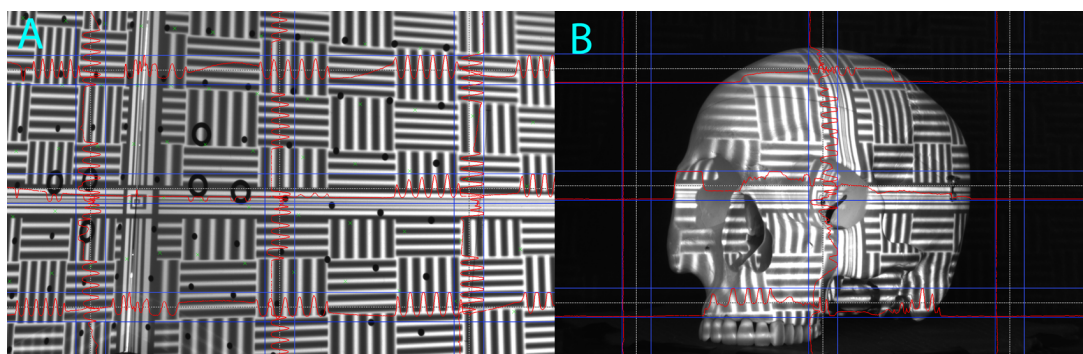

Figure 2.2: An example of how the distortion of a projected light pattern can be used in SLS scanning. A) shows the original light pattern, projected onto a flat surface. Note the pattern distortion in B) when an object with a curved external surface is placed in front of the SLS. Using several different light patterns with stripes of different dimensions, the external surface of the object can be represented.

It is important to note that an SLS can only create a mesh of the surface that is visible to

the camera at the time of scanning. In order to create a complete 3D model, several scans of the same object must be performed with the object in different positions. This is most practically achieved by placing the object on a rotatable surface and scanning at set degree intervals. The result is that there are several meshes of the same object in different positions that must be aligned to one another to create a coherent 3D model (see Figure 2.3 below). A certain amount of overlap must be present in the scans in order for this alignment to succeed, although the optimal amount of overlap required is dependent upon the algorithm used for alignment. An optimal overlap has never been established in the literature, although alignment algorithms such as Super 4PCS (Mellado et al. 2014) are programmed such that alignments should theoretically be possible given any amount of overlap. Regardless of which algorithms are used for alignment, two different approaches are usually used and combined - global registration and fine registration. Global registration focuses on aligning two point clouds or meshes roughly such that they are fairly correctly positioned with respect to one another (Mellado et al. 2014). Fine registration aligns two point clouds or meshes that are already roughly aligned, and seeks to minimize the point-to-point (technically vertex-to-vertex in meshes) distance between the two scans in the overlapping region (Mellado et al. 2014). As a general rule, and in most cases, fine registration does not succeed without first performing a global registration in order to position the two scans correctly.
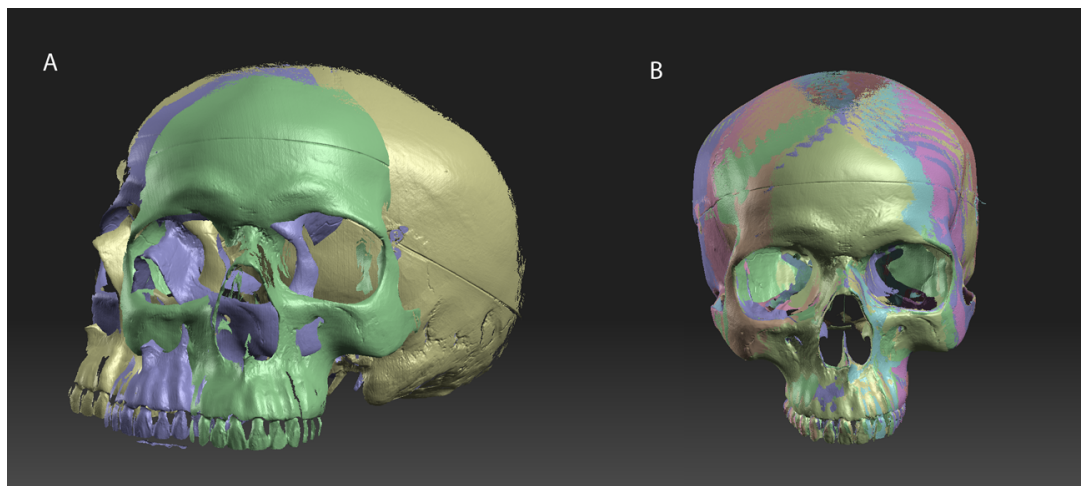

Figure 2.3: A) represents the raw data generated from the SLS, with each scan coloured differently. Only three scans are shown here in order to exemplify the fact that they are positioned differently and require alignment. B) shows all scans when they are correctly aligned to one another.

2.4 Setting Up & Calibrating the DAVID SLS-3 Scanner

Before any data collection could take place, the scanner had to be properly set up. The camera and projector were already mounted onto a bar, which was then secured onto a tripod. The tripod was then raised such that the camera and projector were angled down slightly. This ensured that the light from the projector sufficiently highlighted the features of the object with as little shadow on the object as possible. The distance between the camera and projector were adjustable, but was kept at 250 mm as recommended by the DAVID manual. The camera needed to be angled in order to create a triangle between the projector, the object, and the viewframe of the camera, and was therefore set to the default 22°, with an aperture of f/16. This value is the smallest aperture value on the camera and was chosen to maximize the depth of field of the images, resulting in the largest area possible to be in focus. The trade-off to using a small aperture was that images were darker due to the limited amount of light entering through the aperture. This was mitigated by ensuring that the lab space was as well-lit as possible, increasing the brightness of the projector, and/or slowing down the shutter speed of the camera to allow more light into the camera. Decreasing the aperture further than f/16 would have caused the resulting scans to lose sharpness (DAVID-4 2017), so a smaller aperture was not used.

To ensure that the SLS was set up properly, a plastic 360° protractor was placed on the table, along with a turntable covered with felt. The turntable was marked with a white line across the two movable layers, and the lines were aligned to 0° on the protractor. The cranium (for reference) was placed on the turntable. Using the "Setup" tab in the DAVID 4 program to visualize the sample using the camera, the camera focus, shutter speed, SLS distance, projector brightness, and tripod height were adjusted so that the cranium was clearly in view from any angle. The protractor was then taped onto the table to ensure that it did not move while scanning. The resulting set-up is seen below in Figure 2.4.
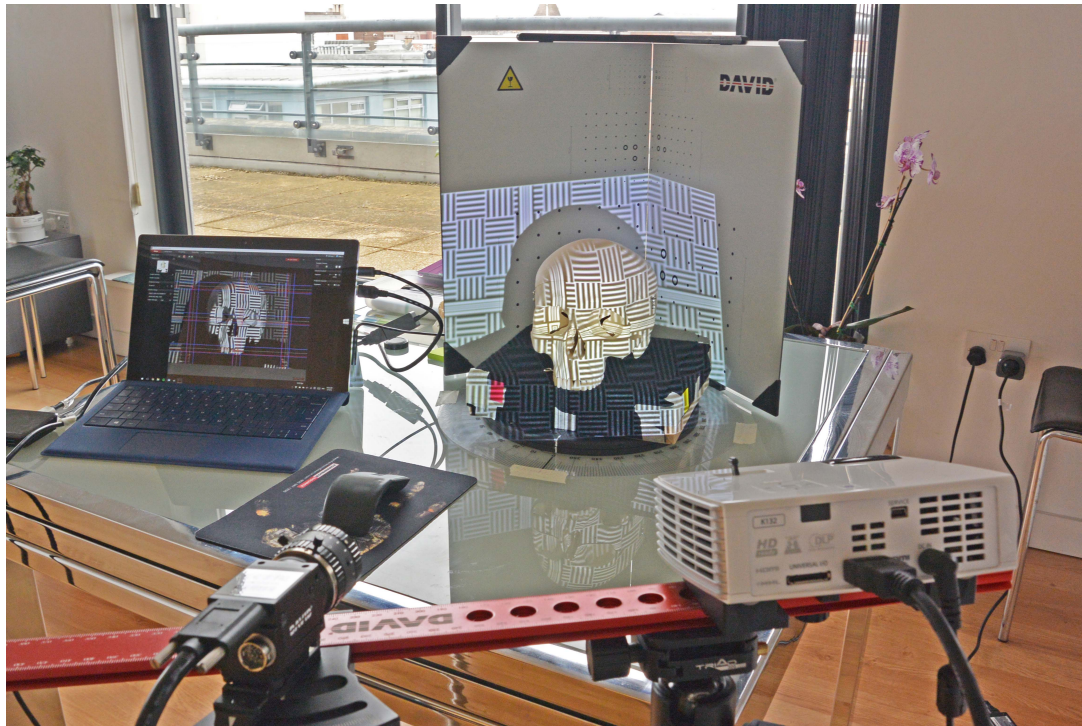
Figure 2.4: The set-up of the SLS (the camera and projector mounted onto the red bar on a tripod) and the sample. The protractor is taped onto the table with a black turntable placed on top, the white marker set to 0°. The cranium is in view and in focus on the tablet screen, indicating that the angle of the SLS and the cranium are correct. It should be noted that this photograph features a black box covered in felt to support the cranium, which also has coloured tape surrounding it. The varying colours were initially added to help with alignment of the consequent scans, but it was later determined that this is not necessary. The box has therefore been removed from the data collection procedure and is an extraneous feature in this photograph.

The SLS then needed to be calibrated such that the camera distance and angle relative to the projector could be calculated. With these parameters, it is possible for the software program to automatically scale the scans such that the object dimensions are preserved. Different calibration settings are needed depending on the size of the object to be scanned, and four available sizes are provided by the calibration boards. It is advisable to use the calibration board with the smallest area that can still encompass the object to be scanned in order to ensure maximum possible resolution during scanning. For this reason, the calibration board used for the crania was 120 mm. For autopsied samples, the 60 mm board was used for the calva. Due to the need to re-calibrate depending on whether a cranium or an autopsied calva was scanned, it was not practical to scan an entire cranium (if autopsied) before scanning the next. Therefore, all crania were first scanned during the first round of visual assessments, before re-calibrating and scanning the calva during the re-assessments.

To calibrate, the calibration board needed to be placed approximately the same distance away from the SLS as the object. The turntable and cranium were therefore set aside, leaving just the protractor. The middle of the protractor was marked with a crosshair, which was used as a marker to place the calibration board. The calibration board consists of two foldable panels with dots, and for the purposes of calibration, the panels needed to be kept perpendicular (at 90°) to each other. The intersection of these two panels was then placed on top of the crosshair of the protractor. It is important to ensure that the SLS was able to capture as many dots as possible on the appropriate calibration board, so the height or tripod angle may needed to have been adjusted accordingly. Lastly, the dots on the calibration board needed to be in focus on the computer screen, and upon viewing the calibration board itself, the pattern denoting the center of the projection needed to be sharp. After these adjustments were made, it was imperative to ensure that the camera angle, focus, and position relative to the projector did not change throughout the entire scanning procedure.

Using the "Calibration" tab, the appropriate calibration area was inputted (120 mm for crania, 60 mm for autopsied calva), and the calibration process was started. This prompted a series of light patterns to be projected onto the calibration board to establish the camera and projector's positions. A white balance was performed automatically by projecting red, green, and blue patterns onto the calibration board to ensure that colour information was captured properly. Once the calibration was complete, the projector displayed a checkered board onto the calibration board, corresponding to each of the dots (see Figure 2.5). This exemplified a successful calibration since the projector can only achieve this if the software program understands its position relative to the calibration board.
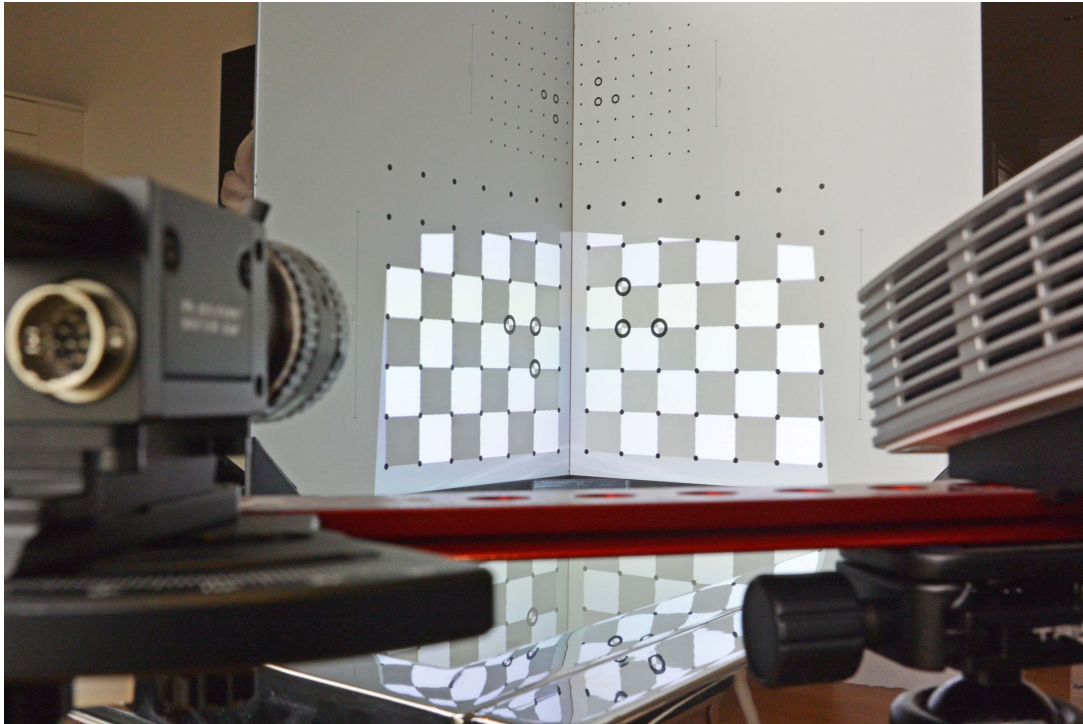
Figure 2.5: A checkered pattern is projected from the SLS onto the 120 mm calibration board, corresponding to the dots on the board. This indicates that the SLS has successfully performed calibration by calculating the distance and angles between the camera, projector, and board.

Once the calibration was performed, the calibration boards were shifted backwards and the turntable replaced onto the protractor. The calibration boards were then covered with black felt to create a black background during scanning. Anything that is black is automatically considered a "hole" by the SLS, and as such was not captured; therefore, by creating a black background, the background in the scans were automatically removed.

2.5 Scanning Crania with the DAVID SLS-3 Scanner

In the DAVID 4 Pro software program, it is possible to see which scans contribute to the overall 3D model, and where holes - or lack of information - exist. In addition to ensuring that all the scans cover the entire sample to avoid holes, it is important to establish an adequate amount of overlap between each subsequent scan. An overlap is necessary so that the software program can properly align and stitch the scans together in the correct orientation. The number of scans must therefore be optimized in terms of adequate overlap, coverage of the sample, and the time needed to produce each 3D model. A set refers to the scans taken of an object in a given orientation as the object is rotated at pre-determined intervals for a full 360 degrees.

The number of sets and the orientation of the cranium were thus determined by trial scans.

Scanning at 45° intervals (8 scans) when the cranium was in an upright position did not provide enough overlap for the software to recognize the orientation of each scan. The number of scans was therefore increased by reducing the scanning intervals to 30°. As seen below in Figure 2.6, the superior and inferior aspects of the cranium were not captured, so additional sets of scans with the cranium lying down on its lateral sides were required. These sets were able to be aligned properly with the first set when scanning at 45° intervals to create an overall model with no holes (Figure 2.7). The orientation of the cranium for the first scan in each set is displayed in Figure 2.8.
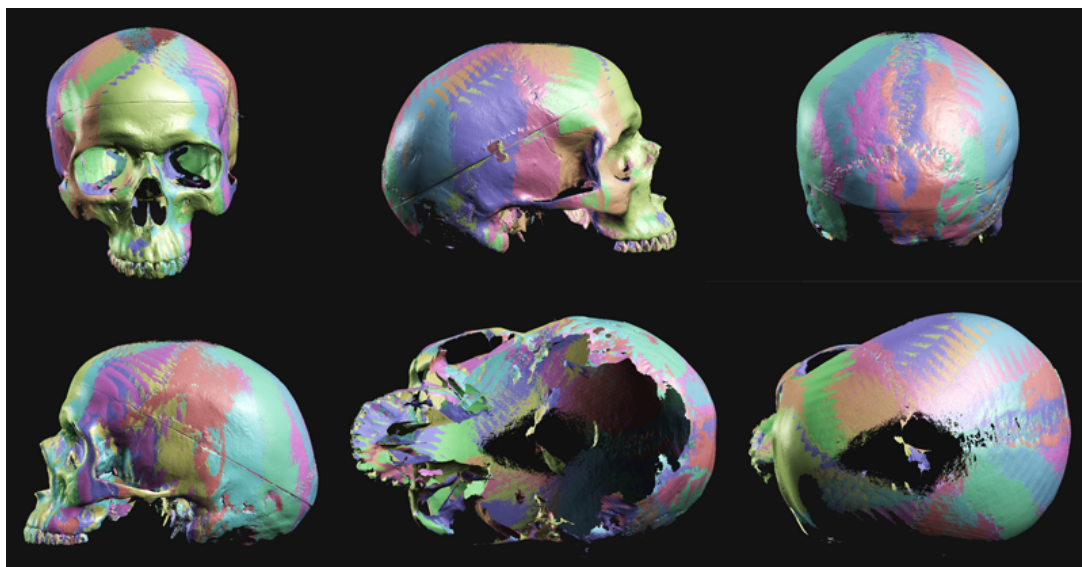
Figure 2.6: Six different views (anterior, posterior, lateral, superior, inferior) of the same cranium when scanned at 30° intervals. The texture, or colour information, is turned off to visualize which scans contributed to which parts of the model. As seen in the superior and inferior views, holes exist because these aspects were not captured.
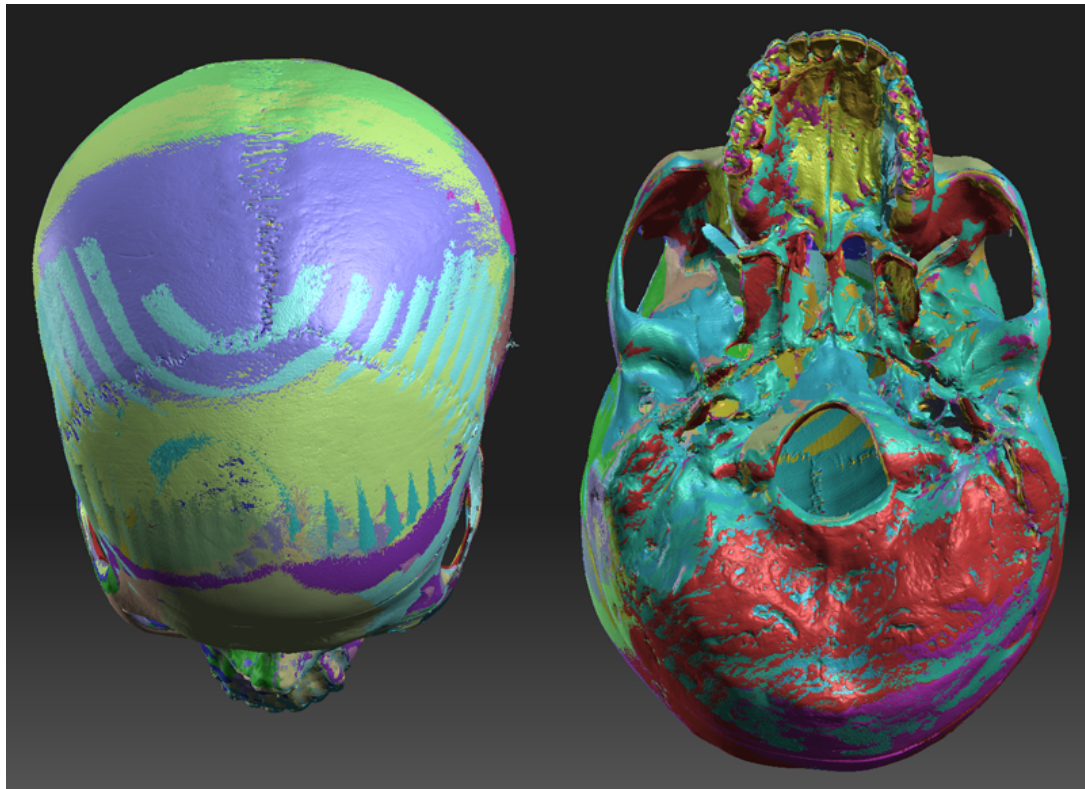
Figure 2.7: The inferior and superior view of the cranium sample after two additional sets of scans were performed. The holes previously seen in the superior and inferior aspects are now closed.

Figure 2.8: The orientation of the cranium for the first scan in each of the three sets. A) the cranium is set upright; B) the cranium is laid down on its left side; C) the cranium is laid down on its right side.

It should be noted that autopsied crania were scanned differently depending on how the calva was cut, and usually consisted of 15 - 22 scans. The first 12 scans were with the autopsied crania placed inferior-side up (i.e. with the cut surface down) and scanned at 30° intervals. The next set of scans was for capturing the details on the inferior surface of the crania, which was oftentimes not captured by the initial set depending on the cut (e.g. straight, egg-shell, or V-shaped) and how high up on the crania the cut was done. Due to the variation

in how the crania were autopsied, the remaining 3 - 10 scans were done either by propping up the crania such that the inferior surface was more or less parallel to the camera and scanning at 45° from 3 different positions, or placing the crania on its side and performing two sets of five scans at 45°, with the crania lying on the left and right side respectively for each set. The interior part of the cranium was not of interest in this study, so these two sets of scans only covered 180° around the anterior, inferior, and posterior sides.

Finally, the scanning parameters needed to be established. First, the quality of the scan was set to the maximum level, "Quality". This means that all of the available light patterns (a total of 29) were used to sample the object, thus recording the maximum amount of spatial information. The "Auto add texture" option was enabled, meaning that the colour information of the object was recorded. To ensure that the scans were saved within the project, the "Auto add to list" option was also enabled.

The shutter speed was adjusted on a case-by-case basis, and depended on the lightness or darkness of a sample, as well as the ambient light in the room. A very dark and soil-stained sample, for example, required a longer shutter speed so that the details on the sample were visible in the scan; conversely, a bleached sample was very bright and the shutter speed needed to be increased. The red sinusoidal lines within the blue vertical and horizontal lines were used to choose the correct shutter speed, since the red sinusoid should not get cut off (indicating that the exposure is too bright), and the peaks should be as close to the blue lines as possible (if they are too low, the exposure is too dark). A screenshot indicating correct exposure is given as an example below in Figure 2.9.
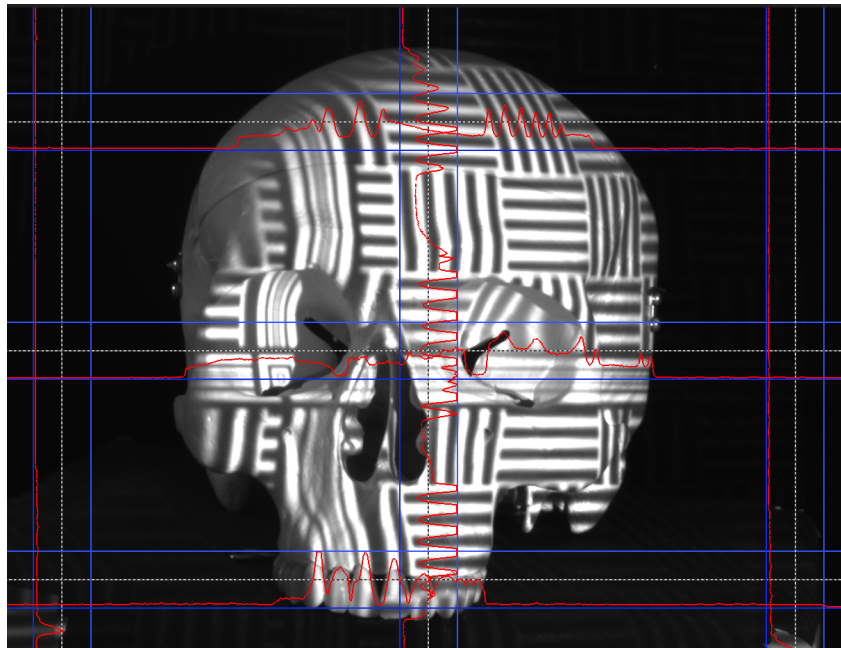
Figure 2.9: An example of correct exposure, which can be adjusted by increasing or decreasing the shutter speed of the camera. The red sinusoidal lines are close to the blue bars and overall the cranium is well-lit.

One of the goals of this project was to create a ground-truth 3D database, so it needed to be populated with reference models that can be readily used and are as complete as possible. Therefore, not all samples were scanned. Due to the way in which points are collected and a point cloud is created, scanning samples that are too dark, have extremely shiny surfaces, or have moveable features that introduce error into the scan (e.g. hair) would not result in a good or useable 3D model, so such samples were excluded. Additionally, samples that exhibited damage or fragmentation were likewise excluded from scanning and from being included into the database. The following table lists the exclusion criteria which was used to determine which samples were excluded from scanning. Only those that belonged to the third category denoted a sample which was still assessed visually but was excluded from scanning, whereas the first and second categories were both included for scanning. The reason for splitting the inclusion criteria into two categories was to ensure that if for any reason the data collection was shortened or was unexpectedly halted, the best samples in the collection that are most suitable for inclusion into the database were prioritized, resulting in the best possible outcome of a worst-case scenario. Samples that were assigned as second priority were those that may not yield the best quality or results after scanning but still contained useful information/features despite this limitation.

Table 2.8: Prioritization criteria for scanning.

Priority	Criteria
1 (inclusion)	100% complete (or very nearly), almost ideal bone texture and quality (not eroded or damaged), good preservation, little to no pathology or hair obscuring features, no deformation or warping
2 (inclusion)	$\geq$75% complete, adequate bone texture and quality (at least 75% of the bone should be well-preserved and/or have unobscured features, some hair/pathology/small deformations but should not affect scan quality; alternatively, dark-coloured specimens that are unlikely to yield excellent scans but may still be useful despite decreased quality)
3 (exclusion)	<75% complete, poor preservation/bone texture; alternatively, the sample is too dark (close to black) or shiny; or too much hair that would affect scans

2.6   Creating Coherent 3D Point Clouds From Scans

Before any alignment and fusion were possible, the scans first needed to be cleaned which involved cutting out irrelevant features that were captured during the scanning process, such as the turntable, tape, or background. An example of a scan which requires cleaning is given below in Figure 2.10.
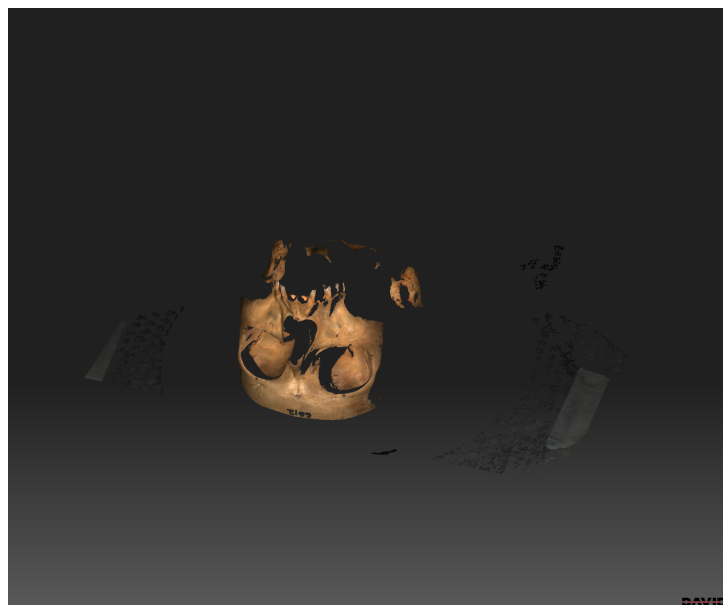

Figure 2.10: A scan that requires cleaning. Note the irrelevant features that were captured during scanning - the tape and the felt tablecloth.

Using the graphical user interface (GUI) of DAVID 4, the irrelevant features were cut

out of the scans so that only the features of the skeletal sample remained. Once this was done for all scans of a given sample, the files were exported from DAVID 4 into .obj files, which is a non-proprietary file format. Exporting was necessary so that the scans could be aligned and fused using CraniAlign. DAVID 4 was not used for several reasons: 1) Although it is fairly straightforward to align and fuse scans manually in the DAVID 4 program, the program that is provided with the SLS scanner does not allow for an automatic program to align and fuse the scans. This must therefore be done manually, and is extremely time-consuming. 2) An SDK (Software Development Kit) is available with the industrial version of the DAVID 4 program which allows for such automation, but is an extremely expensive add-on. 3) Since the algorithms for alignment and fusion are proprietary in the DAVID 4 software, any error in the resulting 3D models due to these processes are unknown. 4) The DAVID 4 alignment program includes some degree of randomness (DAVID-4 2017), which compounds the issue of unknown error. The use of DAVID 4 is inappropriate for research purposes for these four reasons, especially because it does not allow error in the resulting 3D models to be quantified. To address this limitation, CraniAlign was created to automate and control the alignment and fusion process. The parameters of CraniAlign are explained in 4 (Examining the Properties of 3D Models) and compared to DAVID 4. A summary of how the 3D data in this project were acquired and processed - which includes a cursory summary of how CraniAlign was utilized in the data processing - is provided below in Figure 2.11.
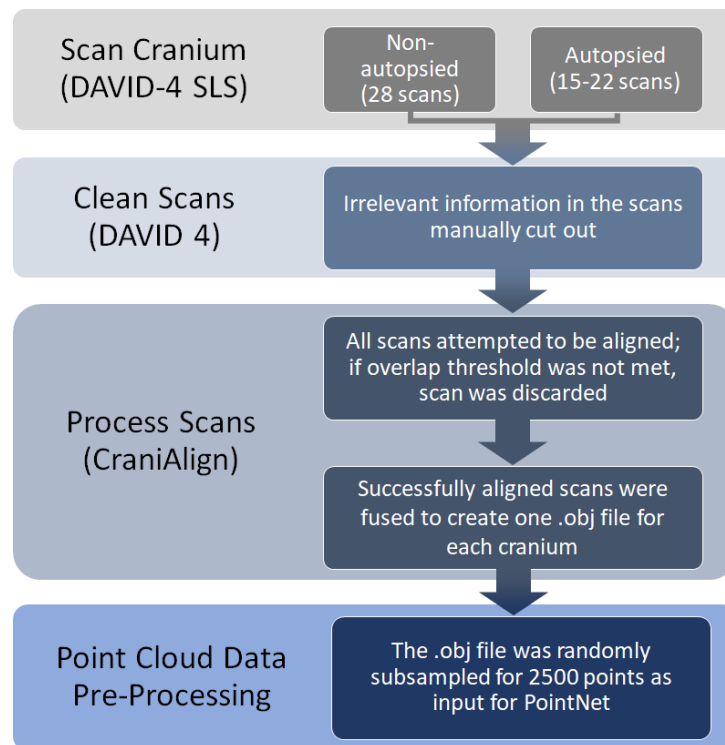
Figure 2.11: A workflow of the 3D data acquisition and processing required in order to achieve 3D point cloud data that were usable in the data analysis stage (see Chapter 5 for how the data was analyzed).

Chapter 3

Cranial Sexual Dimorphism in Various Populations

This chapter discusses the results and implications of the visual assessments on dry bone, both from a research perspective as well as the impact the results have on this PhD project. Following the visual assessment protocol outlined in Chapter 2 (Data Acquisition & Methodology), individuals from the four skeletal collections were assessed based on the degree of sexually dimorphic trait expression in the cranium. Though it is well-established that visual assessments are subjective to varying degrees (as discussed in Chapter 1), performing these visual assessments is a necessary precursor to understanding which traits vary according to sex and population, which in turn was useful for comparing against and understanding the analyses of the 3D data (see Chapter 5). An important output of this chapter is the formulation of a "discrimination factor", which is a novel approach for quantifying the usefulness of a trait in indicating sex.

3.1 Visual Assessment Results

For each skeletal collection, the results and discussion are broken up into three major sections: the overall results, which report on the overall accuracy of classifying males and females based on all cranial traits assessed during the two rounds of visual assessments; the results of examining the accuracy of each trait both according to sex and age category; and

finally, a discussion and summary of the results from the first two sections. After the results of the visual assessments have been reported and discussed for each collection, the results are combined in order to discuss their implications on a global, interpopulation scale.

The overall results include the accuracy of classification for both rounds of assessments and the intraobserver error. Intraobserver error was determined by calculating the number of instances in which an individual was categorized differently between the two rounds, regardless of whether the individual was categorized correctly. In order to interpret the intraobserver error, the Kappa statistic (Cohen 1960) was also calculated. The Kappa statistic, represented by κ, is calculated as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o = frequency of observations in agreement and P_e = hypothetical probability of agreement due to chance. P_e is calculated using the following equation:

$$P_e = \frac{1}{N^2} \sum_{i=1}^{n_c} n_i \, m_i$$

where N = total number of observations; n = total number of observations in a given category (i) for the first round of observations; m = total number of observations in a given category (i) for the second round of observations; and n_c = total number of categories. In the case of reporting the overall results achieved in this project for each skeletal collection, $n_c = 3$ (male, female, and indeterminate). A perfect agreement is indicated by $\kappa = 1$; a level of agreement due purely to chance is indicated by $\kappa = 0$; and a negative κ indicates that the level of agreement is worse than chance. In an attempt to interpret κ in a manner amenable to non-researchers, Viera and Garrett (2005) proposed the following interpretation (Table 3.1) which is used to judge the findings in this study.

Table 3.1: The qualitative interpretations of the Kappa statistic (κ) proposed by Viera and Garrett (2005).

κ	Level of Agreement
< 0	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement

Examining each trait individually was done by first reporting the total number assessed for males and females for that trait. Some traits are paired, meaning that the total number of observations for those traits are twice the number of individuals. Intraobserver error was also reported, along with kappa, although it should be kept in mind that there were instances in which a sample was assessed once, but then during the re-assessment was gauged to be too damaged/pathological to assess, or vice-versa. The total numbers for the intraobserver error therefore only account for the number of times a sample was assessed twice, which may be lower than the total number assessed for the trait. Following Williams and Rogers' (2006) definition of a high-quality trait, intraobserver error will be deemed acceptable if it is $\leq$ 10%.

Next, the frequency of males and females assigned to each score (1 - 5 for the Buikstra & Ubelaker traits; 1, 3, or 5 for the Williams & Rogers traits) was established in order to be analyzed. The frequency, rather than a count, was chosen for three main reasons. Firstly, not all of the cranial traits were available to be assessed on every individual, resulting in different numbers of observations for each trait. By using frequency rather than a count, results were more comparable between traits. Secondly, and by the same logic, some cranial traits are paired and there were instances in which only one side could be assessed, leading to different numbers of observations for a paired trait. Frequency for that trait was then calculated by summing the observations from both sides. Thirdly, the use of frequencies allowed probability distribution graphs to be created for each trait, which would not have been possible if average scores were used. The advantage to using a probability distribution graph rather than reporting average scores is that a probability implies the chance of an individual to be scored a certain way, rather than insinuating that the average score is reflective of an absolute truth. The issue of reporting results that are based on subjective observations - which is a problem with scoring

cranial traits - is therefore taken into account by reporting the results as a probability. Finally, another advantage to using probability distributions and frequencies is that the probability of an individual in the given population to be assigned a certain score for the particular trait in question was able to be established. By extension, it is possible to calculate the probability of a male and a female belonging to the population in question to be assigned to different categories (i.e. scored differently) based on the given trait - this shall be defined as the discrimination factor (d):

$$d = 1 - \sum_{s=1}^{n_s} P(F = s)\, P(M = s)$$

where $d \in [0, 1]$; $P(F) =$ frequency of females having been assigned a particular score (s); $P(M) =$ frequency of males having been assigned a particular score (s); and $n_s =$ the number of scoring categories used, i.e. 5 for the Buikstra & Ubelaker traits or 3 for the Williams & Rogers traits. The discrimination factor (d) therefore directly represents the usefulness of the trait in question because it is a measure of discrimination between the sexes; additionally, $1 - d$ quantifies the amount of overlap between male and female trait expressions. It must be made clear, however, that the latter is not related to the overlap integration between males and females that is typically used to portray sexual dimorphism (see Figure 3.1). Instead, $1 - d$ gives the probability that a male and a female in a given population are given the same score, which is more practically useful than simply calculating the area of overlap between two functions.
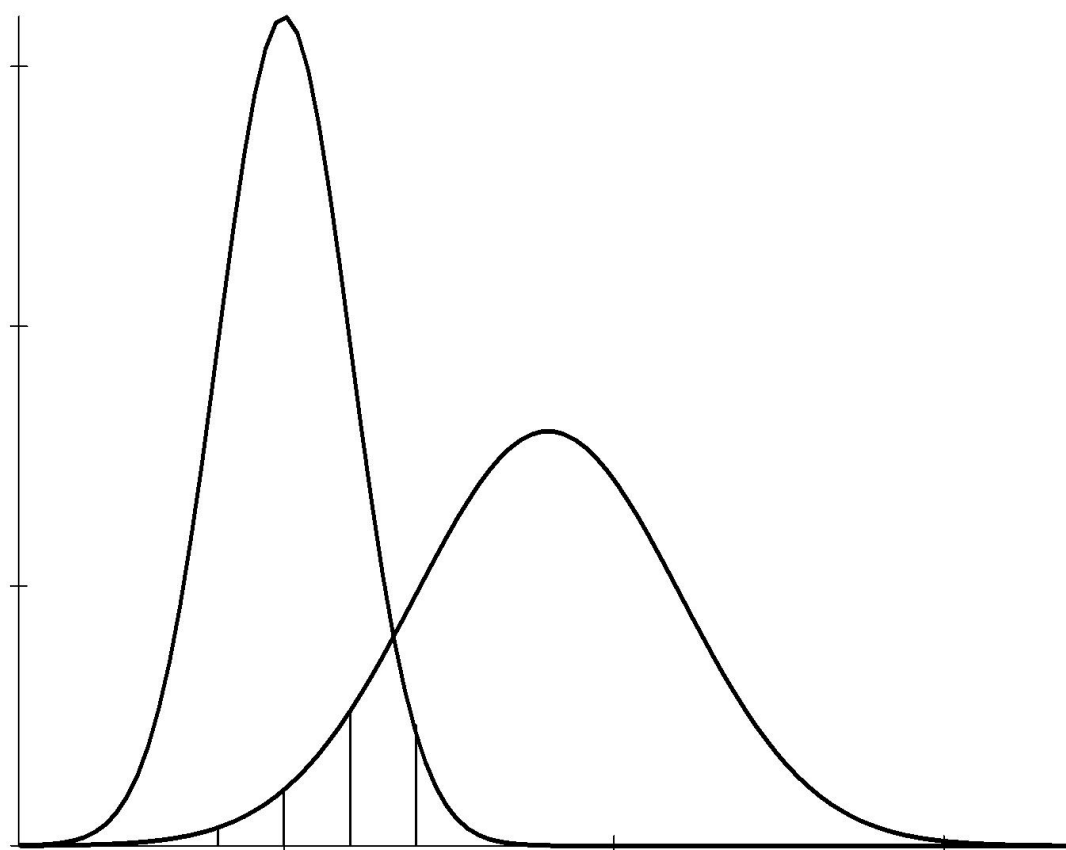
Figure 3.1: A graph of how sexual dimorphism is typically represented, with the overlap between males and females given by the lined area. This can be quantified by calculating the integral of the lined area, which is not as useful as calculating the discrimination factor d as defined in this study. This image was taken from Wikipedia to exemplify the fact that it is quite common to depict sexual dimorphism in this way, even outside academia. It is therefore important to clarify that the discrimination factor is *not* related to this depiction, which would be an erroneous but intuitive assumption given the way "overlap" is typically portrayed.

If d is low, the trait in question either does not display enough sexual dimorphism in the population to be useful as an indicator of sex (i.e. the overlap in male and female trait expression is high), or the expression of the trait was too difficult/variable for the researcher to score properly. When possible, the discussion for each skeletal collection includes a literature review of the results from other researchers who also assessed the same trait in order to ascertain whether the latter was a possible factor influencing the quality of the results. From a practical viewpoint, a trait with a discrimination factor greater than or equal to 0.800 will be considered an acceptable indicator of sex following the 80% accuracy cut-off value that Williams and Rogers (2006) identified for high-quality traits. Probability distribution graphs for each trait and each collection were generated, but for the sake of concision, only those of traits that have a discrimination factor greater than or equal to 0.800 are included in the body of this thesis. For reference and

completion, however, all graphs are included in Appendices C, D, E, and F.

Each trait was also examined by age to determine whether the ability to distinguish between males and females was affected by age. It was therefore of interest in this study to investigate whether the age at which a trait became more or less discriminatory could be established. A scatterplot was created of trait score vs. age for each sex, and four standard fitting functions were used - linear, quadratic, cubic, and logarithmic - in order to determine which, if any, modelled the data best. Once this was done, the average absolute error (AAE) and the coefficient of determination (R^2) were calculated for each function. The AAE was chosen to be calculated rather than the least mean squares error (LMSE), which is more typically used and reported in bioarchaeological studies, because it was possible to define a meaningful cut-off value that is more easily understood than if LMSE's were used. In this study, any functions with an AAE of 0.50 or less was considered a possible candidate to model the relationship between age and trait score. Choosing a cut-off value of 0.50 means that trait scores could vary up to 1 score apart, which is consistent with Buikstra and Ubelaker's (1994) scoring system in which females are either scored 1 or 2 and males are either 4 or 5. Functions with an AAE greater than 0.50 or with an R^2 value of less than 0.50 were not considered or discussed because these functions do not model the data well, although they are included in Appendices C, D, E, and F for the sake of transparency. The function with the lowest AAE therefore theoretically represents the function that best explains the relationship between age and trait score. This analysis provides two pieces of information - the AAE and R^2 quantify how good the fit of the function is to the data, and therefore how well the function explains the relationship between age and trait expression; and the function itself explains how the trait changes according to age for each sex (or if it does at all).

Age and sex were also investigated together. First, the individuals were sorted into age categories defined in Chapter 2 (Data Acquisition & Methodology), Table 2.2. The scoring was compared between the sexes to determine if it was significantly different between males and females of the same age category. This was done by performing Mann-Whitney tests between both sexes for each age category. Mann-Whitney was chosen because this test is appropriate to compare two groups of ordinal, non-parametric data. Mann-Whitney is also applicable if there is the possibility of having uneven sample sizes in the two groups (Field 2013), which is the case with the data presented here. The results of Mann-Whitney tests include the Mann-

Whitney statistic (U) which is based on the sum of ranks in a given group and the sample sizes of each group, and the associated p-value. Additionally, following good practice in statistics, the standardized test statistic (z) and the effect size (r) were calculated. The standardized test statistic (z) is a more understandable quantification of the difference in the medians of the two groups being compared, and is used when calculating the effect size (r) which is simply a normalized quantification of how sex affects trait scoring, when sample size is taken into account. Achieving a significantly different result for any age category (i.e. p < 0.05) indicated that the distribution of scoring for males and females were different enough such that the trait is expressed differently for that age range. By comparing the scoring between the sexes for each age category, the approximate ages at which these changes become significant in adults could be established. The results of the Mann-Whitney tests were also substantiated by calculating the discrimination factor (d) for each age category, which gave the probability of a male and a female of the same age category to be scored the same way. The output of the Mann-Whitney tests for each trait are provided in Appendices C, D, E, and F, and a simplified version that compares all traits in each collection is presented in the body of this chapter.

It should be noted that the non-simplified version of the Mann-Whitney tables in the Appendices serve as look-up tables for each trait. These look-up tables allow a practitioner to gauge the strength of their result if they are assessing the sex of an unknown cranium, provided that the population to which the unknown individual belongs is known and their age is estimated. By reporting the generated p-value and discrimination factor for the trait that is assessed for sex, the practitioner can quantify the strength of their conclusion. Furthermore, the look-up tables generated in this research project can be used to interpret scores of 3 which are otherwise considered "indeterminate", thereby indicating whether males or females are more likely to receive a score of 3. Even if age is not known, the overall discrimination factor for that trait (i.e. the age-agnostic discrimination factor) can be used to interpret indeterminate scoring. The output of this research therefore allows indeterminate scoring to be interpreted in a manner that can indicate sex. Consequently, a score of 3 is no longer "indeterminate", and the ability to quantify such an interpretation provides a vital component to any analysis both in research and for court purposes.

3.1.1 SB Collection Results

The data collection performed on this skeletal collection formed the basis for all future data collection protocols, as SB was the first collection to be documented. For this reason, the exclusion/inclusion criteria outlined in Chapter 2 (Data Acquisition & Methodology) were not consistently followed for the SB collection as these criteria were being developed contemporaneously.

Of the 213 adults in the collection, 187 individuals were assessed for sex (92 female and 95 male). Using the combination of traits and assuming each trait was equally weighted, 114 individuals were correctly categorized in the first round of assessment (60.96%); 19 were incorrectly categorized (10.16%); and 54 were indeterminate (28.88%). For the second round of assessment, 105 individuals were correctly categorized (56.15%); 14 were incorrectly categorized (7.49%); and 68 were indeterminate (36.36%). The breakdown of correct categorization for both rounds of assessment is given below in Tables 3.2 and 3.3. The intraobserver error for the SB collection was 55/187 (29.41%), and κ was calculated to be 0.533 which indicates a moderate degree of agreement (refer to Table 3.1). This indicates that while scoring differed 29.41% of the time, the degree to which the scoring was in agreement was moderate.

One male individual did not have a known and recorded age, so his results have been excluded from all age-related analyses. This accounts for the discrepancy between the total number of males for this collection and the total number of males in age-related results.

Table 3.2: An overview of the classification results from the first round of visual assessments on the SB Collection.

	Correct	Incorrect	Indeterminate
Females	78/92 (84.78%)	2/92 (2.17%)	12/92 (13.04%)
Males	36/95 (37.90%)	17/95 (17.89%)	42/95 (44.21%)
Total	**114/187 (60.96%)**	**19/187 (10.16%)**	**54/187 (28.88%)**

Table 3.3: An overview of the classification results from the second round of visual assessments on the SB Collection.

	Correct	Incorrect	Indeterminate
Females	71/92 (77.17%)	2/92 (2.17%)	19/92 (20.65%)
Males	34/95 (35.79%)	12/95 (12.63%)	49/95 (51.58%)
Total	**105/187 (56.15%)**	**14/187 (7.49%)**	**68/187 (36.36%)**

In order to investigate the usefulness of each trait as an indicator of sex, the discrimination factor (d) was calculated for each trait, and scoring consistency was investigated by establishing the interobserver error and the associated kappa statistic. The results for each trait are given below in Table 3.4, while Figure 3.2 displays the discrimination factor for each trait according to age category. The median trait scores according to age and sex were also established for each trait, and the distribution of scoring was examined using the Mann-Whitney statistical test to determine if there was a significant difference between males and females in each age category. The full results of the Mann-Whitney statistical tests for each trait are found in Appendix C, and the summary table is given below in Table 3.5.

Table 3.4: The usefulness of each trait as an indicator of sex in the SB collection, given by the discrimination factor (d) and the ability to score the trait consistently which is represented by interobserver error (i) and the kappa statistic (κ).

| Trait	d	Overall		Females		Males	
		i	κ	i	κ	i	κ
Nuchal Crest	0.819	61/167 (36.53%)	0.538	26/80 (32.50%)	0.560	35/87 (40.23%)	0.488
Mastoid Process	0.815	186/339 (54.87%)	0.292	98/170 (57.65%)	0.181	88/169 (52.69%)	0.330
Supraorbital Margin	0.738	137/259 (52.90%)	0.281	61/132 (46.21%)	0.351	76/127 (59.84%)	0.195
Glabella	0.883	65/153 (42.48%)	0.432	19/78 (24.36%)	0.494	46/75 (61.33%)	0.176
Zygomatic Extension	0.652	67/342 (19.59%)	0.608	32/150 (21.33%)	0.575	35/172 (20.35%)	0.479
Nasal Aperture	0.567	35/118 (29.66%)	0.451	18/62 (29.03%)	0.267	17/56 (30.36%)	0.523
Cranial Size	0.784	31/142 (21.83%)	0.632	7/74 (9.46%)	0.513	24/68 (35.29%)	0.450
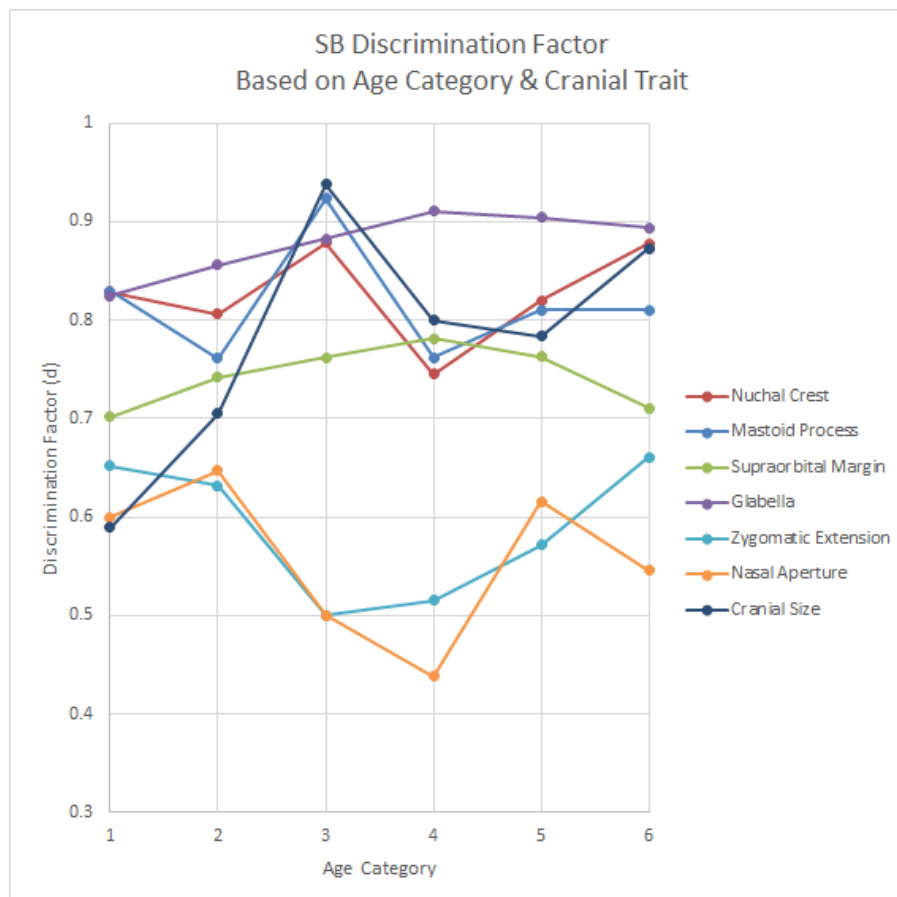
Figure 3.2: A line graph of the discrimination factor for each trait according to age category in the SB collection. This provides a visual comparison of each trait as an indicator of sex as well as a representation of how the discrimination factor of a trait changes according to age.

Table 3.5: The median scores of each trait for SB males and females in each age category. Age categories that do not display a statistically significant difference in scoring distribution according to the Mann-Whitney statistical tests (i.e. $p < 0.05$) are greyed out.

Sex	Age Category	Nuchal Crest	Mastoid Process	Supraorbital Margin	Glabella	Zygomatic Extension	Nasal Aperture	Cranial Size
F	1	2.0	1.0	2.0	1.0	1.0	1.0	1.0
M		4.0	3.0	2.0	3.0	5.0	2.0	3.0
F	2	2.0	2.0	2.0	1.0	1.0	1.0	1.0
M		3.5	3.0	2.0	3.0	5.0	3.0	3.0
F	3	2.0	1.5	2.0	1.0	1.0	1.0	1.0
M		4.0	4.0	3.0	3.0	3.0	3.0	3.0
F	4	3.0	2.0	1.0	1.0	1.0	1.0	1.0
M		2.5	3.0	3.0	4.0	5.0	1.0	5.0
F	5	2.0	2.0	2.0	1.0	1.0	1.0	1.0
M		3.0	3.0	3.0	3.0	5.0	3.0	3.0
F	6	1.0	2.0	2.0	1.0	1.0	1.0	1.0
M		4.0	3.0	2.0	4.0	5.0	3.0	3.0

Discussion & Conclusion of the SB Collection Results

Overall, combining the results of all traits to assess sex was useful for recognizing females (84.78% and 77.17% were categorized correctly for the two rounds), but very poor for recognizing males (37.90% and 35.79% were categorized correctly). Most males were actually categorized as indeterminate (44.21% and 51.58%), which is consistent with the researcher's overall impression that males were quite gracile in this population. It is therefore reasonable that when using a global scale of trait scoring such as Buikstra and Ubelaker's Standards (1994), males would fall around the middle and thus be categorized as indeterminate. In this study, however, the results show that males can in fact be distinguished from females in this population using three traits - the glabella, the nuchal crest, and the mastoid processes, which all achieved discrimination factor values greater than 0.800.

The glabella was the best indicator of sex in the SB population, with a discrimination factor of 0.883. The discrimination factor also remained acceptably high (i.e. over 0.800) for all six age categories, and the trait distributions between males and females were significantly different in all six age categories as well. The median score for females remained at 1 in all age categories, whereas the median score for males varied between 3 and 4, exemplifying the fact that a score of 3 could be used to distinguish males from females in this population even though a score of 3 is normally considered indeterminate.

The next best indicator of sex in the SB population was the nuchal crest, with an overall discrimination factor of 0.819. The discrimination factor remained acceptably high in all but one age category, with significantly different trait distributions between males and females in all but one of the categories (category 4, 50 - 59 years old). There was more variation in both males and females, with females having median scores between 1 - 3 depending on the age category, and males having median scores between 2.5 - 4.

The third best indicator of sex in the SB population was the mastoid process, with an overall discrimination factor of 0.815. The discrimination factor remained acceptably high in four out of the six age categories, but males and females had significantly different trait distributions in all categories. Median scores for females varied between 1 - 2, which is consistent with Buikstra and Ubelaker's scoring system, whereas the median score for males was 3 for all age categories.

Stevenson and colleagues (2009) used individuals from the SB collection, as well as others from the Hamann-Todd Osteological Collection and the Robert J. Terry Anatomical Collection, to create decision trees based on a combination of morphological traits in the skull to categorize individuals by sex. The combination of traits with the highest accuracy was the glabella, mental eminence (which is a mandibular trait), and mastoid size. Although the researchers included individuals of English, European American, and African American ancestry, they found that this combination of traits produced decision trees that best predicted the sex of the English individuals in the SB collection. The results of Stevenson and colleagues' study (2009) are consistent with the findings in this chapter for the SB collection, for which the glabella and mastoid process were among the three best traits as well. Their results are directly applicable to the results in this chapter because they also used the same scoring system given by Buikstra and Ubelaker (1994) for their data collection. Despite analyzing their data using different methods, however, their results are consistent with the results obtained in this chapter, which lends credence to the conclusions drawn from both their study and the ones in this chapter.

3.1.2 NU Collection Results

Out of the 150 individuals (75 female and 75 male), and using an equally-weighted combination of all the traits, 86 individuals were correctly categorized in the first round of assessment (57.33%); 9 were incorrectly categorized (6.00%); and 55 were indeterminate (36.67%). For the second round of assessment, 88 were correctly categorized (58.67%); 10 were incorrectly categorized (6.67%); and 52 were indeterminate (34.67%). The breakdown of correct categorization for both rounds of assessment is given below in Tables 3.6 and 3.7. The intraobserver error for the NU collection was 33/150 (22.00%), and κ was calculated to be 0.716 which indicates substantial agreement (refer to Table 3.1).

Table 3.6: An overview of the classification results from the first round of visual assessments on the NU Collection.

	Correct	Incorrect	Indeterminate
Females	52/75 (69.33%)	1/75 (1.33%)	22/75 (29.33%)
Males	34/75 (45.33%)	8/75 (10.67%)	33/75 (44.00%)
Total	**86/150 (57.33%)**	**9/150 (6.00%)**	**55/150 (36.67%)**

Table 3.7: An overview of the classification results from the second round of visual assessments on the NU Collection.

	Correct	Incorrect	Indeterminate
Females	53/75 (70.67%)	1/75 (1.33%)	21/75 (28.00%)
Males	35/75 (46.67%)	9/75 (12.00%)	31/75 (41.33%)
Total	**88/150 (58.67%)**	**10/150 (6.67%)**	**52/150 (34.67%)**

In order to investigate the usefulness of each trait as an indicator of sex, the discrimination factor (d) was calculated for each trait, and scoring consistency was investigated by establishing the interobserver error and the associated kappa statistic. The results for each trait are given below in Table 3.8, while Figure 3.3 displays the discrimination factor for each trait according to age category. The median trait scores according to age and sex were also established for each trait, and the distribution of scoring was examined using the Mann-Whitney statistical test to determine if there was a significant difference between males and females in each age category. The full results of the Mann-Whitney statistical tests for each trait are found in Appendix D, and the summary table is given below in Table 3.9.

Table 3.8: The usefulness of each trait as an indicator of sex in the NU collection, given by the discrimination factor (d) and the ability to score the trait consistently which is represented by interobserver error (i) and the kappa statistic (κ).

Trait	d	Overall		Females		Males	
		i	κ	i	κ	i	κ
Nuchal Crest	0.811	27/120 (22.50%)	0.711	14/52 (26.92%)	0.617	13/68 (19.11%)	0.752
Mastoid Process	0.821	110/300 (36.67%)	0.523	49/150 (32.67%)	0.538	61/150 (40.67%)	0.451
Supraorbital Margin	0.723	95/296 (32.09%)	0.527	45/146 (30.82%)	0.494	50/150 (33.33%)	0.494
Glabella	0.814	32/136 (23.53%)	0.667	13/67 (19.40%)	0.567	19/69 (27.54%)	0.625
Zygomatic Extension	0.459	27/300 (9.00%)	0.795	16/150 (10.67%)	0.782	11/150 (7.33%)	0.789
Nasal Aperture	0.669	26/114 (22.81%)	0.711	14/72 (19.44%)	0.668	12/72 (16.67%)	0.708
Cranial Size	0.803	35/150 (23.33%)	0.639	13/75 (17.33%)	0.531	22/175 (29.33%)	0.519
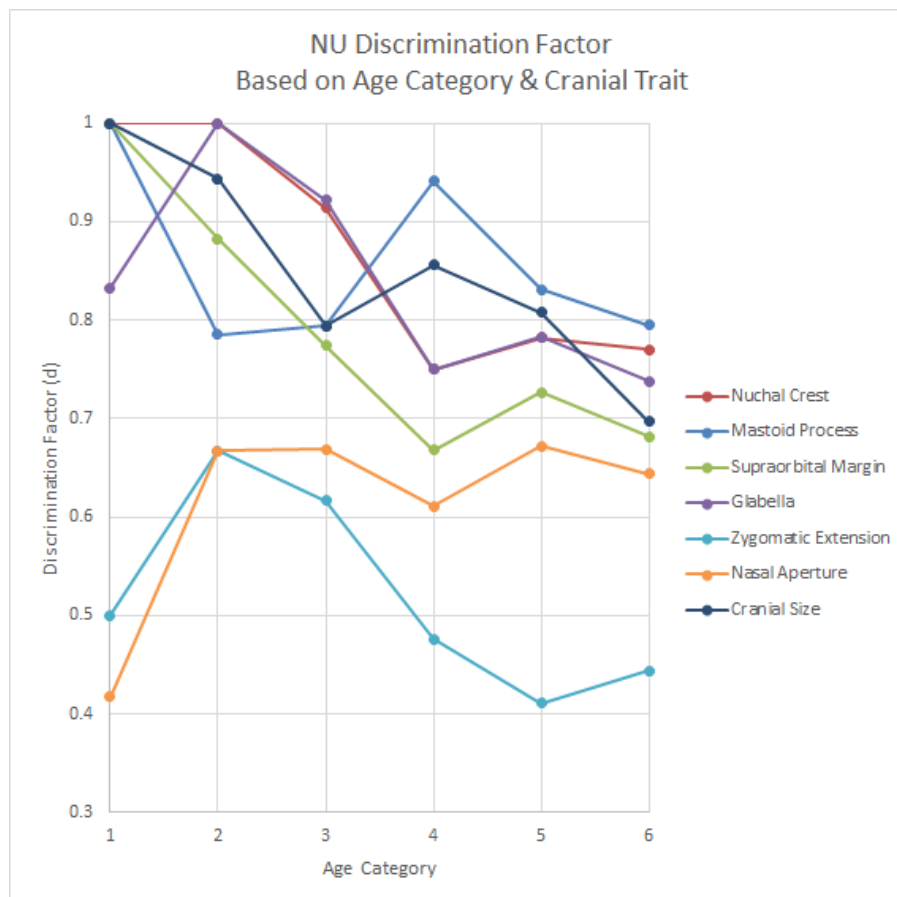
Figure 3.3: A line graph of the discrimination factor for each trait according to age category in the NU collection. This provides a visual comparison of each trait as an indicator of sex as well as a representation of how the discrimination factor of a trait changes according to age.

Table 3.9: The median scores of each trait for NU males and females in each age category. Age categories that do not display a statistically significant difference in scoring distribution according to the Mann-Whitney statistical tests (i.e. $p < 0.05$) are greyed out.

Sex	Age Category	Nuchal Crest	Mastoid Process	Supraorbital Margin	Glabella	Zygomatic Extension	Nasal Aperture	Cranial Size
F	1	1.0	1.0	1.0	1.0	3.0	5.0	1.0
M		5.0	3.0	3.0	2.0	5.0	5.0	5.0
F	2	1.0	2.0	2.0	1.0	1.0	3.0	1.0
M		4.0	3.0	3.0	3.5	5.0	5.0	5.0
F	3	1.5	2.0	1.0	1.0	1.0	1.0	1.0
M		4.0	3.0	2.0	3.0	5.0	5.0	5.0
F	4	2.0	1.0	2.0	1.0	5.0	2.0	1.0
M		3.0	3.0	2.0	2.0	5.0	5.0	3.0
F	5	2.0	2.0	2.0	2.0	5.0	3.0	1.0
M		3.0	3.0	2.0	3.0	5.0	3.0	3.0
F	6	2.0	2.0	2.0	1.0	5.0	1.0	1.0
M		3.0	3.0	2.0	2.0	5.0	5.0	3.0

Discussion & Conclusion of the NU Collection Results

Overall, combining the results of all traits to assess sex was more useful for recognizing females than males (69.33% and 70.67% accuracies as opposed to 45.33% and 46.67% between the two rounds of assessments), but total accuracy is not much better than chance (57.33% and 58.67%). Approximately a third of all individuals were categorized as indeterminate during both rounds of assessment (36.67% and 34.67%), meaning that the combination of traits used was not useful for assessing sex. There are, however, four traits which individually have discrimination factor values greater than 0.800 - the mastoid process, the glabella, the nuchal crest, and the cranial size.

The mastoid process had the highest discrimination factor, with a value of 0.821; however, it only seems to be useful in categories 1, 4, and 5 and does not seem to be related to age in any way. Furthermore, the intraobserver error rates were quite high - 40.67% for males and 32.67% for females, although scoring consistency was moderate. Despite the mastoid process having the highest discrimination factor, it is not a very reliable trait due to the high intraobserver error and low discrimination factor values for three age categories. The median score for males was consistently 3 for all age categories, which is indeterminate according to the scoring system by Buikstra and Ubelaker's *Standards* (1994) as well as by Williams and Rogers' paper (2006). Combined with the fact that the median for females fluctuated between 1 and 2 depending on the age category means that it is difficult to create a localized scale for the mastoid process scoring such that males and females can be distinguished more easily.

The trait with the second highest overall discrimination factor is the glabella, with a value of 0.814. The discrimination factor remains acceptable only for the first three age categories (≤ 49 years old). The fact that the results in this study demonstrate that the glabella has a high discrimination factor for those 49 years old or younger makes it a more reliable trait than the mastoid process, for which no such results were discernible. It must be noted, however, that the number of observations for the first three age categories is small, and therefore the results should be interpreted cautiously. Nevertheless, the results indicate that the glabella is potentially a valuable indicator for sex in the first three age categories, although it suffers from the same issue as with the mastoid process in that the median score for males fluctuated between 2 - 3.5. This makes it difficult to normalize the scores such that distinguishing between males and females can be done with a higher accuracy.

The nuchal crest also had an acceptable overall discrimination factor, with a value of 0.811; however, the discrimination factor is only acceptable for the first three age categories ($\leq$ 49 years old), which is the same issue as the glabella. Similarly, the sample sizes are small for the first three age categories, even moreso than for the glabella, and so these results must be interpreted cautiously.

Lastly, the cranial size had an acceptable overall discrimination factor of 0.803, and remained acceptable in all but two categories - 3 and 6 (40 - 49 years old, and $\geq$ 70 years old, respectively). Similar to the mastoid process, there was no evident relationship regarding the age at which the discrimination factor changes. The intraobserver error was lower than the mastoid process, however, although it was still higher than what is deemed acceptable for both males and females. Interestingly, while the median score for females remained at 1 throughout all age categories, the median score changed from 5 to 3 at age category 4 (50 - 59 years old) for males. It is important to remember that assessing cranial size also includes assessing the ruggedness of the cranium, so one possible interpretation of these results is that the ruggedness of the cranium becomes less prominent in older males. Contrarily, however, no relationship was found between trait expression and age in the scatterplots for males, so this remains a possible interpretation that must be explored in more depth before drawing concrete conclusions.

In conclusion, it is difficult to ascertain which trait best indicates sex in the NU collection, mostly due to the fact that there are so few individuals in the lower age categories. The small sample sizes for these categories limits the ability to draw robust conclusions regarding age. Therefore, the only strong conclusions that can be drawn are those that take into account the sample as a whole, without sub-dividing it by age or age category. The overall discrimination factors therefore indicate which traits were most indicative of sex in the NU collection, which are, in order of highest to lowest: the mastoid process, the glabella, the nuchal crest, and the cranial size.

To the researcher's knowledge, there have been no published studies that use the NU collection for investigating cranial sexual dimorphism. A study was conducted, however, on modern Japanese skeletons from the Jikei Medical University in Tokyo by Işcan and colleagues (1995) in which cranial dimensions were used in stepwise discriminant analyses to determine which measurements were good predictors of sex. The results from the study found that mas-

toid height alone was the best predictor of sex, which is consistent with the results in this chapter where the mastoid process as a whole was found to have the highest discrimination factor. Conversely, however, the study by Işcan and colleagues (1995) found that cranial size seems to have increased in females in recent generations, thus decreasing sexual dimorphism in Japanese crania. Although these results seem to suggest that cranial size is not a good indicator of sex, it must be noted that Işan and colleagues (1995) used cranial measurements to make this determination, whereas cranial size in this research project not only included an assessment of size but also of general rugosity or gracility. This added dimension to the assessment of cranial size in this project may have therefore contributed to the fact that cranial size had the fourth highest discrimination factor in the NU collection, with a value higher than 0.800.

3.1.3   ML Collection Results

Milano Skeletal Collection (curated by LABANOF) - Milan, Italy (ML)

Out of the 150 individuals (70 female and 80 male), and using an equally-weighted combination of all the traits, 76 individuals were correctly categorized in the first round of assessment (50.67%); 14 were incorrectly categorized (9.33%); and 60 were indeterminate (40.00%). For the second round of assessment, 88 were correctly categorized (58.67%); 17 were incorrectly categorized (11.33%); and 45 were indeterminate (30.00%). The breakdown of correct categorization for both rounds of assessment is given below in Tables 3.10 and 3.11. The intraobserver error for the ML collection was 33/150 (22.00%), and κ was calculated to be 0.709 which indicates substantial agreement (refer to Table 3.1).

Table 3.10: An overview of the classification results from the first round of visual assessments on the ML Collection.

	Correct	Incorrect	Indeterminate
Females	54/70 (77.14%)	0/70 (0.00%)	16/70 (22.86%)
Males	22/80 (27.50%)	14/80 (17.50%)	44/80 (55.00%)
Total	**76/150 (50.67%)**	**14/150 (9.33%)**	**60/150 (40.00%)**

Table 3.11: An overview of the classification results from the second round of visual assessments on the ML Collection.

	Correct	Incorrect	Indeterminate
Females	59/70 (84.29%)	0/70 (0.00%)	11/70 (15.71%)
Males	29/80 (36.25%)	17/80 (21.25%)	34/80 (42.50%)
Total	**88/150 (58.67%)**	**17/150 (11.33%)**	**45/150 (30.00%)**

In order to investigate the usefulness of each trait as an indicator of sex, the discrimination factor (d) was calculated for each trait, and scoring consistency was investigated by establishing the interobserver error and the associated kappa statistic. The results for each trait are given below in Table 3.12, while Figure 3.4 displays the discrimination factor for each trait according to age category. The median trait scores according to age and sex were also established for each trait, and the distribution of scoring was examined using the Mann-Whitney statistical test to determine if there was a significant difference between males and females in each age category. The full results of the Mann-Whitney statistical tests for each trait are found in Appendix E, and the summary table is given below in Table 3.13.

Table 3.12: The usefulness of each trait as an indicator of sex in the ML collection, given by the discrimination factor (d) and the ability to score the trait consistently which is represented by interobserver error (i) and the kappa statistic (κ).

Trait	d	Overall		Females		Males	
		i	κ	i	κ	i	κ
Nuchal Crest	0.772	29/150 (19.33%)	0.756	15/70 (21.43%)	0.689	14/80 (17.50%)	0.785
Mastoid Process	0.827	102/297 (34.34%)	0.554	45/139 (32.37%)	0.528	57/158 (36.08%)	0.508
Supraorbital Margin	0.723	122/300 (40.67%)	0.414	46/140 (32.86%)	0.475	76/160 (47.50%)	0.313
Glabella	0.845	38/149 (25.50%)	0.660	12/70 (17.14%)	0.694	26/79 (32.91%)	0.553
Zygomatic Extension	0.515	52/300 (17.33%)	0.653	17/140 (12.14%)	0.743	35/160 (21.88%)	0.555
Nasal Aperture	0.634	31/132 (23.48%)	0.609	7/63 (11.11%)	0.761	24/69 (34.78%)	0.453
Cranial Size	0.721	36/150 (24.00%)	0.599	10/70 (14.29%)	0.405	26/80 (32.50%)	0.507
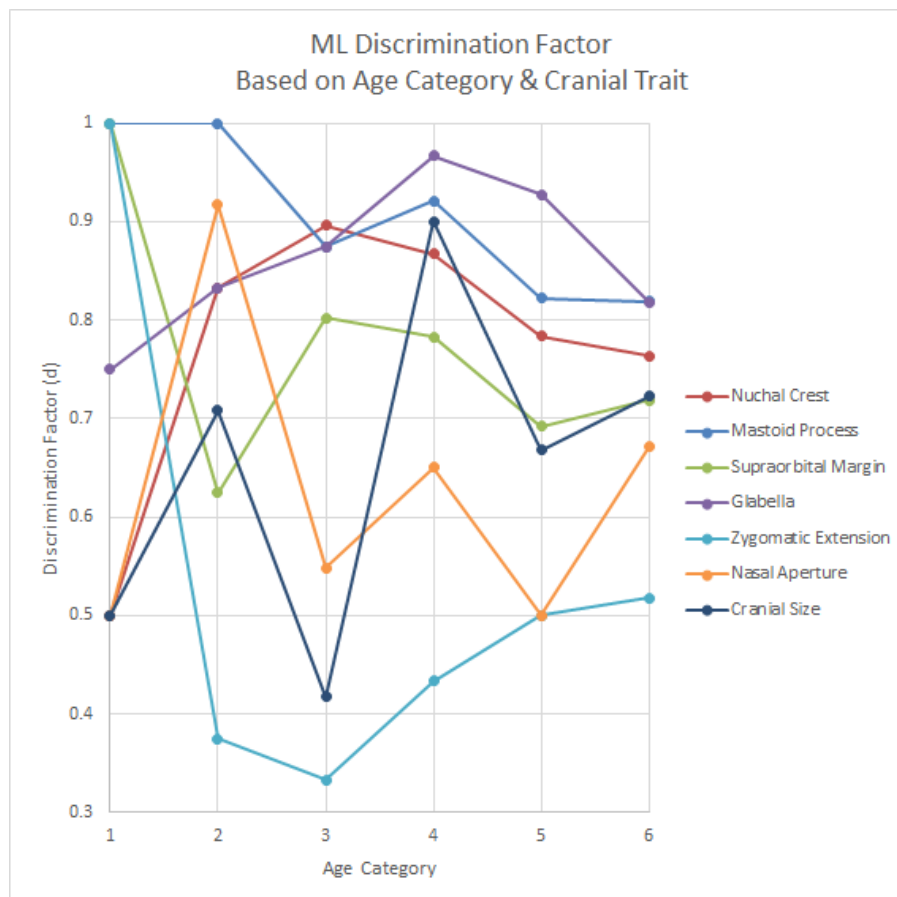
Figure 3.4: A line graph of the discrimination factor for each trait according to age category in the ML collection. This provides a visual comparison of each trait as an indicator of sex as well as a representation of how the discrimination factor of a trait changes according to age.

Table 3.13: The median scores of each trait for ML males and females in each age category. Age categories that do not display a statistically significant difference in scoring distribution according to the Mann-Whitney statistical tests (i.e. $p < 0.05$) are greyed out.

Sex	Age Category	Nuchal Crest	Mastoid Process	Supraorbital Margin	Glabella	Zygomatic Extension	Nasal Aperture	Cranial Size
F	1	2.0	2.0	1.0	1.0	1.0	5.0	1.0
M		2.5	4.0	3.0	2.5	5.0	4.0	2.0
F	2	3.0	2.0	2.5	1.0	1.0	3.0	2.0
M		2.5	4.0	2.5	3.0	1.0	1.0	3.0
F	3	2.0	2.5	1.0	1.0	1.0	1.0	1.0
M		3.0	4.0	2.0	3.5	1.0	1.0	1.0
F	4	1.0	2.0	1.0	1.0	5.0	3.0	1.0
M		3.0	4.0	2.5	3.0	5.0	3.0	4.0
F	5	2.0	2.0	1.0	1.0	1.0	1.0	1.0
M		3.0	3.0	2.0	4.0	3.0	1.0	3.0
F	6	2.0	2.0	2.0	2.0	1.0	1.0	1.0
M		3.0	4.0	2.0	3.0	5.0	5.0	3.0

Discussion & Conclusion of the ML Collection Results

Overall, using a combination of all the traits, categorizing males and females faired little better than by chance (50.67% accuracy for the first round of assessment, and 58.67% accuracy for the second). A large majority of individuals were indeterminate (40.00% for the first round, and 30.00% for the second round). It was particularly difficult to categorize males correctly (27.50% correct categorization for the first round and 36.25% correct categorization for the second round). For females, an acceptable accuracy rate was only obtained in the second round of assessment where 84.29% were correctly categorized (as opposed to 77.14% for the first round). In both rounds of assessment, no female was incorrectly categorized as a male. Only two traits had discrimination factor values that were acceptable - the glabella, and the mastoid.

The glabella had the highest discrimination factor at 0.845, and was acceptably high in all age categories except the first one (≤ 29 years old, $d = 0.750$). Due to the small sample size in the first age category (two males, four females), however, it is unclear whether the glabella is useful or not in the first age category. Nevertheless, the discrimination factor remains high in all other age categories - especially category 6 which encompasses the majority of the samples - which indicates that the glabella is a useful sexually dimorphic trait in the ML collection. This is despite the fact that the median score for males ranged from 2.5 - 4, which overlaps with what is considered indeterminate.

The mastoid had a slightly lower discrimination factor of 0.827, but the discrimination factor remained acceptably high in all age categories, unlike the glabella. The trait distributions also remained significantly different in all age categories. The median score for females fluctuated from 2 - 2.5 and for males, the median score fluctuated between 3 - 4.

To the researcher's knowledge, there is only one study that uses the crania from the ML collection to assess sex, which was undertaken by Manthey and colleagues (2018). In this study, craniometric measurements were taken and used in FORDISC to both improve the database used to create discriminant functions for categorization and to test FORDISC's current output on a European population, since FORDISC's database is mainly based on American individuals. The results of the study by Manthey and colleagues (2018) show a bias toward female categorization, with almost 25% of males incorrectly categorized as females. The authors attribute this to the fact that the Italian crania seem to differ markedly from the American crania

in terms of their cranial vault dimensions, as well as a less pronounced development of the glabella and mastoid height in Italians compared to Americans.

In this study, what is of interest are two results/conclusions that are drawn by the authors. Firstly, almost 25% of males were incorrectly categorized as females using craniometric points, in contrast to the 17.5% and 21.25% that were incorrectly categorized using morphological traits in this project. The implications of this statement are blatant: the assessment of morphological traits done by a human analyst, which is more subjective, achieved a lower rate of incorrect categorization than that achieved by a computer program, albeit one that was biased towards American samples rather than Italian (although it can be argued that the human analyst also has a similar bias, having been trained on and exposed to North American collections for her undergraduate and Master's degrees). Aside from this possible population-specific bias, another reason why the morphological traits achieved a lower incorrect categorization rate could be due to the fact that craniometric points do not always accurately represent a shape. In fact, craniometric points are most often used to create an abstract shape on which analyses are based (see Chapter 1.3.1 for a discussion on this topic). Considering the findings in this project and those of Manthey and colleagues (2018), it is further evident that morphological assessments offer insight that craniometric points do not.

The second conclusion drawn by Manthey and colleagues (2018) that is of interest to this project is their claim that incorrect categorization was partially due to the decrease in sexual dimorphism of the glabella and the mastoid - which, in contrast, were the only two traits in this project that had acceptable discrimination factor values. This discrepancy could be due to several factors - firstly, although the paper states that Italians have "less pronounced development of glabellar projection and mastoid height" (Manthey et al. 2018), the data that they provide actually seem to demonstrate the opposite. According to their data, Italians and Americans seem to have a very similar amount of sexual dimorphism in terms of the absolute difference between males and females in each population (see Figure 2 in their article). Secondly, and expanding on the first point, there does not seem to be an actual quantification of sexual dimorphism in their study used to support their claim, unlike in this project where sexual dimorphism is defined by the discrimination factor. The conclusion from this PhD project - which is that the glabella and the mastoid processes are the best traits to use to distinguish males and females in the ML collection - is therefore more robust than those made by Manthey and colleagues (2018) regard-

ing these two traits, since this project substantiates this claim with quantifiable data. Thirdly, there is again the issue of craniometric points not properly capturing shape information. In this case, it is likely that the craniometric points used for glabellar projection and mastoid height do not correspond well with the morphological equivalent of assessing the glabella and mastoid process.

3.1.4 PR Collection Results

Pretoria Bone Collection - Pretoria, South Africa (PR)

The breakdown of the 150 individuals (75 female, 75 male) used in this project is as follows: 32 "Black" females; 25 "Black" males; 43 "White" females; and 50 "White" males. Using an equally-weighted combination of all the traits, 87 individuals were correctly categorized in the first round of assessment (58.00%); 11 were incorrectly categorized (7.33%); and 52 were indeterminate (34.67%). For the second round of assessment, 87 were correctly categorized (58.00%); 11 were incorrectly categorized (7.33%); and 52 were indeterminate (34.67%). The breakdown of correct categorization for both rounds of assessment is given below in Tables 3.14 and 3.15. The intraobserver error for the PR collection was 30/150 (20.00%), and κ was calculated to be 0.733 which indicates substantial agreement (refer to Table 3.1).

Table 3.14: An overview of the classification results from the first round of visual assessments on the PR Collection. B = "Black", W = "White".

	Correct	Incorrect	Indeterminate
Females	64/75 (85.33%) *31/32 (96.88%) B* *33/43 (76.74%) W*	1/75 (1.33%) *0/32 (0.00%) B* *1/43 (2.33%) W*	10/75 (13.33%) *1/32 (3.13%) B* *9/43 (20.93%) W*
Males	23/75 (30.67%) *6/25 (24.00%) B* *17/50 (34.00%) W*	10/75 (13.33%) *1/25 (4.00%) B* *9/50 (18.00%) W*	42/75 (56.00%) *18/25 (72.00%) B* *24/50 (48.00%) W*
Total	**87/150 (58.00%)** *37/57 (64.91%) B* *50/93 (53.76%) W*	**11/150 (7.33%)** *1/57 (1.75%) B* *10/93 (10.75%) W*	**52/150 (34.67%)** *19/57 (33.33%) B* *33/93 (35.48%) W*

Table 3.15: An overview of the classification results from the second round of visual assessments on the PR Collection. B = "Black", W = "White".

	Correct	Incorrect	Indeterminate
Females	65/75 (86.67%) *31/32 (96.88%) B* *34/43 (79.07%) W*	1/75 (1.33%) *0/32 (0.00%) B* *1/43 (2.33%) W*	9/75 (12.00%) *1/32 (3.13%) B* *8/43 (18.60%) W*
Males	22/75 (25.00%) *5/25 (20.00%) B* *17/50 (34.00%) W*	10/75 (13.33%) *2/25 (8.00%) B* *8/50 (16.00%) W*	43/75 (62.67%) *18/25 (72.00%) B* *25/50 (50.00%) W*
Total	**87/150 (58.00%)** *36/57 (63.16%) B* *51/93 (54.84%) W*	**11/150 (7.33%)** *2/57 (5.26%) B* *9/93 (9.68%) W*	**52/150 (34.67%)** *19/57 (33.33%) B* *33/93 (35.48%) W*

In order to investigate the usefulness of each trait as an indicator of sex, the discrim-

ination factor (d) was calculated for each trait, and scoring consistency was investigated by establishing the interobserver error and the associated kappa statistic. The results for each trait are given below in Table 3.16, while Figures 3.5 (overall results), 3.6 (results for "Black" individuals), and 3.7 (results for "White" individuals) display the discrimination factor for each trait according to age category. The median trait scores according to age and sex were also established for each trait, and the distribution of scoring was examined using the Mann-Whitney statistical test to determine if there was a significant difference between males and females in each age category. The full results of the Mann-Whitney statistical tests for each trait are found in Appendix F, and the summary tables are given below in Tables 3.17 (overall results), 3.18 (results for "Black" individuals), and 3.17 (results for "White" individuals).

Table 3.16: The usefulness of each trait as an indicator of sex in the PR collection, given by the discrimination factor (d) and the ability to score the trait consistently which is represented by interobserver error (i) and the kappa statistic (κ). B = "Black", W = "White".

Trait	d	Overall		Females		Males	
		i	κ	i	κ	i	κ
Nuchal Crest	0.775 *0.752 B* *0.786 W*	26/144 (18.06%)	0.746	11/73 (14.07%)	0.711	14/71 (21.13%)	0.723
Mastoid Process	0.793 *0.796 B* *0.795 W*	93/299 (31.10%)	0.627	39/149 (26.17%)	0.715	54/158 (36.00%)	0.490
Supraorbital Margin	0.747 *0.737 B* *0.767 W*	125/298 (41.95%)	0.409	44/150 (29.33%)	0.506	81/148 (54.73%)	0.271
Glabella	0.788 *0.809 B* *0.781 W*	41/143 (28.67%)	0.607	22/73 (30.14%)	0.481	19/70 (27.14%)	0.642
Zygomatic Extension	0.580 *0.609 B* *0.564 W*	33/299 (11.04%)	0.779	18/149 (12.08%)	0.717	15/150 (10.00%)	0.758
Nasal Aperture	0.653 *0.671 B* *0.656 W*	45/138 (32.61%)	0.489	25/70 (35.71%)	0.387	20/78 (29.41%)	0.545
Cranial Size	0.763 *0.719 B* *0.774 W*	44/150 (29.33%)	0.529	6/75 (8.00%)	0.611	36/75 (48.00%)	0.260
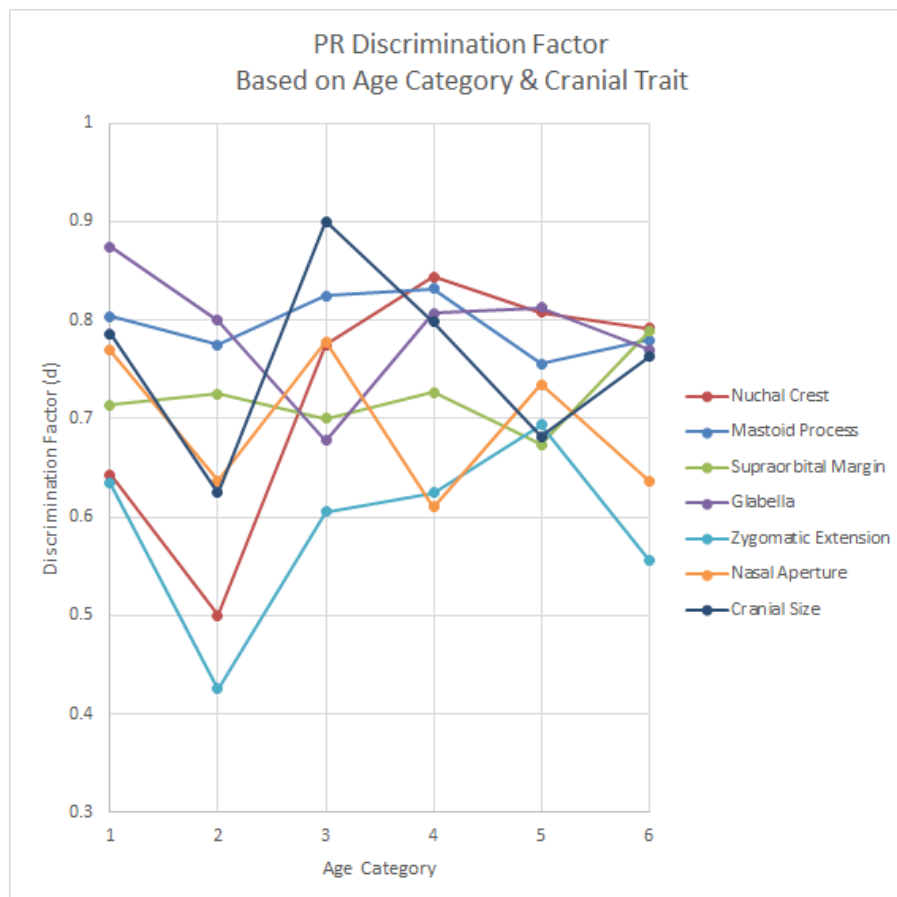
Figure 3.5: A line graph of the discrimination factor for each trait according to age category in the PR collection. This provides a visual comparison of each trait as an indicator of sex as well as a representation of how the discrimination factor of a trait changes according to age.
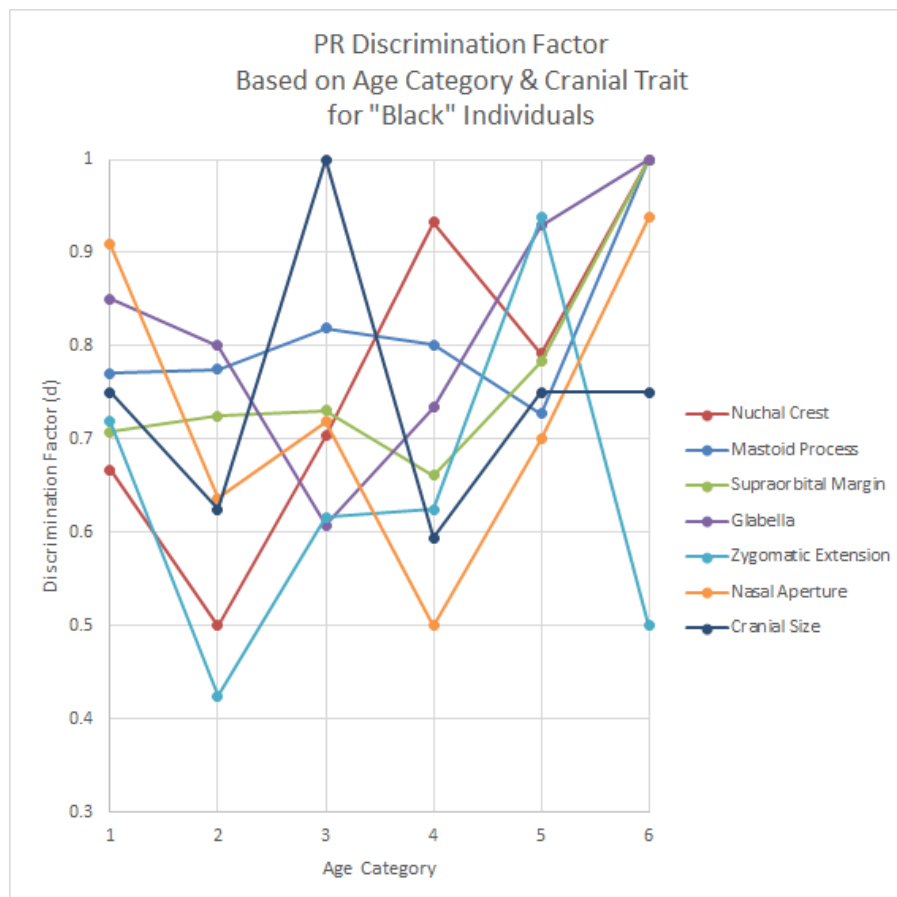
Figure 3.6: A line graph of the discrimination factor for each trait according to age category in "Black" individuals from the PR collection. This provides a visual comparison of each trait as an indicator of sex as well as a representation of how the discrimination factor of a trait changes according to age.
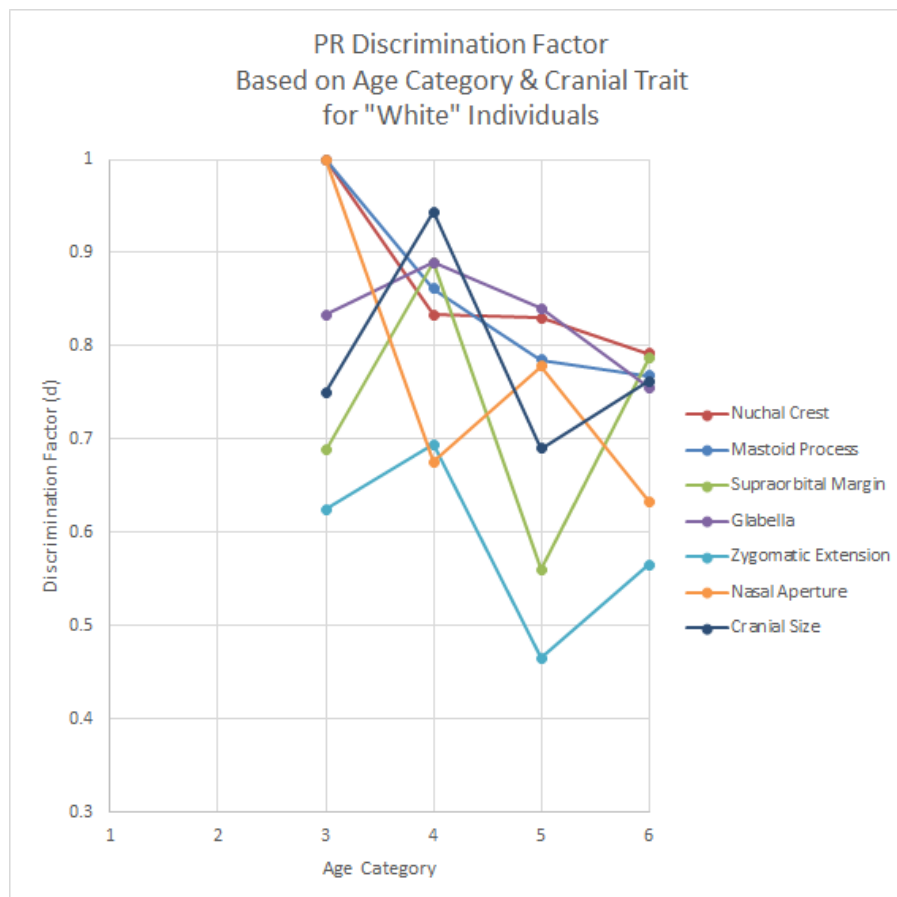
Figure 3.7: A line graph of the discrimination factor for each trait according to age category in "White" individuals from the PR collection. This provides a visual comparison of each trait as an indicator of sex as well as a representation of how the discrimination factor of a trait changes according to age.

Table 3.17: The median scores of each trait for PR males and females in each age category. Age categories that do not display a statistically significant difference in scoring distribution according to the Mann-Whitney statistical tests (i.e. $p < 0.05$) are greyed out.

Sex	Age Category	Nuchal Crest	Mastoid Process	Supraorbital Margin	Glabella	Zygomatic Extension	Nasal Aperture	Cranial Size
F	1	1.5	2.5	1.5	1.0	1.0	1.0	1.0
M		2.0	4.0	2.0	3.0	5.0	5.0	4.0
F	2	1.0	2.0	2.0	2.0	1.0	2.0	1.0
M		1.5	3.5	2.5	3.0	1.0	4.0	3.0
F	3	1.0	3.0	2.0	2.0	1.0	1.0	1.0
M		2.5	4.0	2.0	2.0	5.0	5.0	5.0
F	4	1.0	2.0	1.0	2.0	1.0	1.0	1.0
M		2.0	4.0	2.0	3.0	5.0	3.0	5.0
F	5	1.0	3.0	2.0	2.0	1.0	3.0	1.0
M		3.0	3.0	2.0	4.0	5.0	5.0	3.0
F	6	1.0	2.0	1.0	2.0	1.0	1.0	1.0
M		3.0	3.0	3.0	3.0	5.0	3.0	3.0

Table 3.18: The median scores of each trait for PR "Black" males and females in each age category. Age categories that do not display a statistically significant difference in scoring distribution according to the Mann-Whitney statistical tests (i.e. $p < 0.05$) are greyed out.

Sex	Age Category	Nuchal Crest	Mastoid Process	Supraorbital Margin	Glabella	Zygomatic Extension	Nasal Aperture	Cranial Size
F	1	1.5	2.5	1.5	1.0	1.0	1.0	1.0
M	1	2.0	3.5	2.0	3.5	5.0	5.0	3.0
F	2	1.0	2.0	2.0	2.0	1.0	2.0	1.0
M	2	1.5	3.5	2.5	3.0	1.0	4.0	3.0
F	3	1.0	3.0	2.0	2.0	1.0	1.0	1.0
M	3	2.0	4.0	2.5	2.0	5.0	3.0	5.0
F	4	1.0	2.0	2.0	2.0	1.0	1.0	1.0
M	4	3.0	4.0	2.0	3.5	5.0	1.0	3.0
F	5	1.0	3.0	2.0	2.0	1.0	3.0	1.0
M	5	2.5	3.0	3.0	3.0	5.0	4.0	3.0
F	6	1.0	2.0	1.5	1.0	3.0	1.0	1.0
M	6	2.0	4.5	3.0	3.5	1.0	3.0	4.0

Table 3.19: The median scores of each trait for PR "White" males and females in each age category. Age categories that do not display a statistically significant difference in scoring distribution according to the Mann-Whitney statistical tests (i.e. $p < 0.05$) are greyed out. Note that there were no individiuals in age category 2, so this category has been omitted.

Sex	Age Category	Nuchal Crest	Mastoid Process	Supraorbital Margin	Glabella	Zygomatic Extension	Nasal Aperture	Cranial Size
F	1	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
M		1.0	4.5	2.5	3.0	1.0	1.0	5.0
F	3	1.0	1.5	2.0	1.5	1.0	1.0	1.0
M		3.0	4.0	1.5	3.0	5.0	5.0	4.0
F	4	1.0	2.5	1.0	1.5	1.0	2.0	1.0
M		2.0	4.0	2.0	3.0	5.0	5.0	5.0
F	5	1.5	3.0	1.0	2.0	5.0	2.0	1.0
M		3.0	4.0	2.0	4.0	5.0	5.0	4.0
F	6	1.0	2.0	1.0	2.0	1.0	1.0	1.0
M		3.0	3.0	3.0	3.0	5.0	3.0	3.0

Discussion & Conclusion of the PR Collection Results

Overall, categorizing males and females had a poor accuracy when all traits were combined, with an accuracy ranging between 56.67% - 58.67%. Females, whether they were "Black" or "White", had a much better overall accuracy than males, with 96.88% of "Black" females and 76.74% of "White" females categorized correctly in the first round of assessment as opposed to 24.00% of "Black" males and 34.00% of "White" males categorized correctly. For the second round of assessment, a similar pattern was seen where 96.88% of "Black" females (same as the first round) and 79.07% of "White" females were categorized correctly in the second round, whereas 20.00% of "Black" males and 34.00% of "White" males were correctly categorized. More males were categorized incorrectly than females, with one "Black" male and nine "White" males wrongly categorized as females, whereas only one "White" female was wrongly categorized as a male for the first round; in the second round, two "Black" males and eight "White" males were categorized as females, and only one "White" female was categorized as a male. From these results, it is evident that it was more problematic to correctly categorize males than females, and that more "White" than "Black" individuals were categorized incorrectly.

Approximately a third of all individuals in the PR collection were categorized as indeterminate, with more "Black" males than "White" males being categorized as such (72.00% as opposed to 48.00% in the first round, respectively; 72.00% as opposed to 50.00% in the second round), and more "White" females than "Black" females being categorized as such (20.93% as opposed to 3.13%, respectively; 18.60% as opposed to 3.13% for the second round, respectively). This suggests that the combination of traits, as well as the levels of scoring and definitions of each score, used in this study are simply not discriminatory enough to categorize males and females, especially for "Black" males and "White" females. This is especially evident given the fact that no single trait had an overall discrimination factor above 0.800 when the entire PR collection was assessed.

In an attempt to draw conclusions that can be useful to practitioners working with South African individuals, Table 3.20 ranks all the traits in order of highest to lowest discrimination factor values for the overall PR collection, and then for "Black" and "White" individuals separately. This table will be useful when deciding which traits should be used to assess sex in South African individuals, based on whether "ancestry" is known or not. Examining the results

for each of these traits, it is evident that there is only one instance in which the discrimination factor is acceptable, which is the glabella for "Black" individuals. Otherwise, all other discrimination factor values range from 0.580 - 0.796. Despite the fact that all except one of the values are below the acceptable cut-off of 0.800, there are patterns in the results. The mastoid process, glabella, and nuchal crest are consistently in the top three, whereas the nasal aperture and zygomatic extension are consistently the lowest. It can be concluded that although most of the traits do not meet the acceptable cut-off value, the glabella and mastoid process are the best traits to use for the PR collection regardless of "ancestry", due to the fact that those values are greater than 0.780; and if "ancestry" is known/assessed to be "White", the nuchal crest can also be used in conjunction with these two traits. Conversely, the nasal aperture and zygomatic extension are established to be the worst traits to use in the PR collection, regardless of whether "ancestry" is taken into account.

Table 3.20: The ranking of traits based on discrimination factor values in the PR collection, for "Black" individuals, and for "White" individuals.

Rank	Overall	"Black" individuals	"White" individuals
1	mastoid process (0.793)	glabella (0.803)	mastoid process (0.795)
2	glabella (0.788)	mastoid process (0.796)	nuchal crest (0.786)
3	nuchal crest (0.775)	nuchal crest (0.752)	glabella (0.781)
4	cranial size (0.763)	supraorbital margin (0.737)	cranial size (0.774)
5	supraorbital margin (0.747)	cranial size (0.719)	supraorbital margin (0.767)
6	nasal aperture (0.653)	nasal aperture (0.671)	nasal aperture (0.656)
7	zygomatic extension (0.580)	zygomatic extension (0.609)	zygomatic extension (0.564)

Recent studies in South African forensic anthropology regarding sex assessment of the cranium focus mostly on metric methods, with a particular emphasis on GMM (Bidmos et al. 2010). To the researcher's knowledge, studies that have used morphological methods on South African crania are now outdated (i.e. they date back to the 1960's; e.g. De Villiers 1968), and have since been replaced by newer studies that focus on GMM instead. The only recent study of which the researcher is aware that uses the PR collection to assess sex from the cranium

is by Robinson and Bidmos (2009). The authors tested five discriminant functions previously defined by Steyn and Işcan (1998) on three regionally different skeletal collections in South Africa, one of which is the PR collection. These discriminant functions rely on craniometric measurements taken from the cranium, mandible, and humerus. Considering the functions that use craniometric points only from the cranium ("Functions 1, 2, and 5" in their study), the authors found that the rate of correct classification for individuals in the PR collection ranged from 66.3% - 84.7% (Robinson and Bidmos 2009). This correct classification rate is markedly better than what was achieved in this study, which was between 56.67% - 58.67%, although it must be remembered that discriminant function analyses force a categorization (i.e. individuals are never categorized as "indeterminate"). In this Ph.D. project, approximately a third of all individuals were categorized as indeterminate due to the nature of the scoring system; if it was necessary to force a categorization, the number of individuals correctly categorized would have increased, potentially making the results comparable to those achieved by Robinson and Bidmos (2009). Additionally, it would be prudent to discuss how or if the accuracy of the discriminant functions changed according to "ancestry", but unfortunately, Robinson and Bidmos (2009) do not provide this breakdown. If they did, it would have been interesting to see if "Black" individuals were categorized correctly at a higher rate than "White" individuals, which was the case in this project ("Black" individuals were categorized correctly with an accuracy ranging from 61.40% - 64.91%, whereas "White" individuals ranged from 53.76% - 54.84%). Additionally, since the sample size used by Robinson and Bidmos (2009) was smaller (49 females and 49 males), the effect of "ancestry" could have potentially influenced their results significantly.

In the study performed by Robinson and Bidmos (2009), the function that had the highest rate of classification ("Function 1") used the maximum length of the cranium, the bi-zygomatic breadth, nasal height, nasal breadth, basion-nasion length, and basion-bregma height. The function that performed the worst used the maximum length of the cranium, basion-nasion length, and maximum frontal breadth. The morphological equivalent of these traits are as follows: cranial size can be represented using the maximum length, the bi-zygomatic breadth, basion-nasion length, basion-bregma height, and maximum frontal breadth; nasal aperture can be represented by using the nasal height and nasal breadth. Indirectly, the basion-nasion length would be affected by the glabella, and so this length can be loosely attributed to glabella. Translating the results from the study by Robinson and Bidmos (2009), it can be inferred that the best correct classification rate was achieved when both cranial size and nasal aperture were taken

into account, whereas the worst classification rate was achieved when only cranial size was taken into account, although in both cases glabella loosely influenced the results through the inclusion of the basion-nasion length. Interestingly, the cranial size was mediocre in this study for indicating sex, with a discrimination factor ranging from 0.719 - 0.774, and nasal aperture was consistently the second worst indicator of sex with a discrimination factor ranging from 0.653 - 0.671. This situation is therefore an example where craniometric measurements can more accurately categorize individuals than morphological analyses, possibly because non-important information - which is incorporated for visual assessments - is discarded in craniometric measurements. In this case, since the assessment of cranial size includes overall gracility and robustness, it is possible that overall size information as provided by the craniometric measurements is more important than the visual assessment of ruggedness. Finally, it must be noted that the study by Robinson and Bidmos (2009) does not include any measurements that are related to the mastoid process, the nuchal crest, the supraorbital margin, or the zygomatic extension. It is therefore uncertain whether these traits would have performed well using discriminant function analyses, after translating them to their craniometric equivalent.

3.2   Discussion & Conclusion

Overall, using the combination of all traits, similar trends were seen in all four populations; the rate of correct categorization ranged from 50.67% (ML collection) - 60.96% (SB collection); the rate of incorrect categorization ranged from 6.00% (NU collection) - 11.33% (ML collection); and the rate of categorization individuals as indeterminate ranged from 28.88% (SB collection) - 40.00% (ML collection). The combination of the Buikstra and Ubelaker (1994) traits and the Williams and Rogers (2006) traits consistently performs poorly in the ML collection, since this collection had the lowest correct categorization rate, the highest incorrect categorization rate, and the highest rate of individuals categorized as indeterminate.

In all four skeletal collections, the mastoid process and the glabella were consistently among the most reliable indicators of sex. With the exception of the PR collection, these two traits had discrimination factor values above 0.800, and even in the PR collection these two traits had the highest values. Interestingly, these two traits were the only two recognized by both Buikstra and Ubelaker (1994) and Williams and Rogers (2006) as being good indicators of

sex. The results in this chapter therefore affirm the conclusions made by Buikstra and Ubelaker (1994) and Williams and Rogers (2006). In addition to the mastoid process and glabella, the nuchal crest was also a fairly good indicator of sex. The nuchal crest had acceptable discrimination factor values for the SB and NU collections, and was always ranked in the top three in all populations. In a practical context, knowing which traits are the most discriminatory in all populations is useful if the population/ancestry of an individual is unknown, in which case these three traits have been proven to be reliable regardless of population/ancestry.

In cases where the population/ancestry of an individual is known or can be assessed, the probability graphs for each traits can be used to help analysts interpret scoring. For example, an individual could be scored as a 3 for a given trait - this would normally be categorized as "indeterminate" and is therefore not very useful. With the probability graphs (and the associated $P(F)$ and $P(M)$ values used to generate them), however, the probability of a female in a given population being assigned a score of 3 can be established, as well as the probability of a male in the same population being assigned a score of 3. These probabilities are useful for investigators to decide if the score of 3 is more likely to be assigned to a male or a female, and, in the absence of any other kind of information, could state with an established degree of certainty what sex the unknown individual is likely to be (e.g. if $P(M)$ is greater than $P(F)$ for a score of 3, then it is more likely for a male to be given a score of 3 than a female; therefore, the individual is more likely to be male if they do in fact come from the population in question). Consequently, the use of the probability graphs, $P(F)$, $P(M)$, and d as established in this chapter allows an "indeterminate" score to be further interpreted for determining the individual's most probable sex, with the caveat that population can be assessed or is known. Furthermore, if the unknown individual's age is estimated, $P(F)$ and $P(M)$ can be calculated for their age category; if age is unknown, the overall $P(F)$ and $P(M)$ for that trait can be used instead. The limitation to using age, however, is the smaller sample sizes in the younger age categories, so it is recommended that analysts use the overall $P(F)$ and $P(M)$ values in conjunction with the age-specific values.

Intraobserver error was an issue in this research, with error rates ranging from 7.33% (zygomatic extension for males in the NU collection) - 54.73% (supraorbital margin for males in the PR collection). There were only two instances in which the intraobserver error was acceptable (i.e. $\leq 10\%$) - the cranial size for females in the SB collection (9.46%) and the zygomatic extension for males in the NU collection (7.33%). Scoring agreement was generally

in the categories of fair, moderate, and substantial agreement, and in no circumstance was the trait scoring less than chance, meaning that although the range of trait variation was generally well-recognized, it was somewhat problematic in determining which score should be assigned to different trait expressions. As for the high rates of intraobserver error, they are likely due to the researcher learning the range of trait variation within a population for the first round (which is more indicative of its scoring on a global scale) and then subconsciously adjusting for this fact during the second round (which is more indicative of scoring on a local scale). Interobserver error in the form of a second person also doing two rounds of assessment on the same population would substantiate this theory of a global versus local scoring bias.

In conclusion, the results of this chapter have proven to be extremely useful for investigating the range of trait variation in different populations, quantifying sexual dimorphism by defining the discrimination factor, and consequently allowing "indeterminate" scores to be interpreted in a more useful way. As far as the researcher is aware, no other published study has attempted the latter, making the research in this chapter the first mathematically robust study for doing so. In addition to this novel contribution to the field of skeletal sex assessment, the results from this chapter indicate traits that would be useful in supervised machine learning approaches, which is explored in Chapter 5. The results from this chapter also serve as a comparative reference for the results obtained from unsupervised machine learning, since it allows a comparison between the areas/traits that the program defines as sexually dimorphic and those that a human analyst finds sexually dimorphic.

Chapter 4

Examining the Properties of 3D Models for Research Purposes

A pilot test was performed, with two major aims: 1) to assess the quality of 3D models produced using the SLS, in terms of resolution, reproducibility, and reliability; and 2) to use these three qualities to compare two different methods of aligning the raw data generated from the SLS to create coherent 3D models. Currently, there are no studies that validate the quality of 3D models produced using structured light scanning, especially on a research-oriented data set. It is therefore necessary to examine the 3D data produced by structured light scanning in order to understand the limitations associated with using such data in further analyses. Resolution is defined as the distance between each neighbouring vertex for point clouds. For the purposes of this study, reproducibility refers to the degree of error, measured in millimetres (mm), between two 3D models generated from the same object using the same scanning and alignment parameters but obtained on two different days by the same analyst. Reliability is determined by whether or not different objects scanned and aligned with the same method have similar degrees of error.

This pilot test investigated two different methods of alignment - one given by DAVID 4, and another called CraniAlign that was created for the express purpose of facilitating the large number of scans that needed to be aligned for this PhD project. Currently, the DAVID 4 program is the industry standard for processing SLS scans by aligning and fusing scans together to create coherent 3D models in the form of meshes. Without access to the DAVID

SDK (Software Development Kit), however, these processes cannot be completely controlled by the user, and cannot be completely understood due to the use of proprietary algorithms. Furthermore, operations must be done manually in the DAVID 4 program which is not only time-consuming, but which also have proven to be unreliable since results vary greatly even when the same operations are repeated by the same user. It was therefore necessary to create an alternate program which could process scans reliably in order to create 3D models, use algorithms which are well-documented, can be understood by the user, and whose parameters can be controlled and replicated. CraniAlign was thus created to provide a transparent and controllable method for processing scans such that research can be reliably undertaken on the resulting 3D models. The automation of CraniAlign also allowed for scans to be aligned with minimal manual input, meaning that the program could be left running overnight or even parallel to other tasks such that the creation of coherent 3D models was as time-efficient as possible. This pilot test therefore determined and compared the resolution, reproducibility, and reliability between 3D models generated using the DAVID 4 program and the CraniAlign program.

The samples used in this pilot test consist of five crania, chosen from the St. Mary's skeletal collection at the University of Leicester. The individuals chosen were SMC206, SMC399, SMC417, SMC1142, and SMC1248, and were chosen based on the fact that they met the priority 1 criteria for inclusion (see Table 2.8), and were therefore representative of samples to be included into the ground-truth database. Following the scanning protocols listed above in Data Acquisition & Methodology, each crania was scanned twice, on separate days. On the first day of scanning, the scanning parameters (i.e. the shutter speed and projector brightness) for each crania were determined and recorded. These parameters were replicated on the second day when the crania were re-scanned. After the first day of scanning, the SLS was packed up such that on the second day, the SLS would need to be set up again and re-calibrated. This ensured that the maximum degree of unintentional error due to setting up and calibrating the equipment would be captured in the resulting 3D models. For both days, the windows in the lab were covered with an opaque screen to ensure that sunlight and external weather conditions would not affect the lighting in the lab. The lighting in the lab therefore remained as consistent as possible throughout the entire scanning process on both days. The result is that two sets of scans were created for each cranium, for a total of ten sets of scans.

Each set of scans underwent a two-step basic process in order to create a coherent 3D

model in the form of a point cloud or mesh: 1) coarse alignment, which refers to orienting scans roughly in position, and 2) fine alignment, which refers to minimizing the distance between two neighbouring scans such that they are as closely aligned to one another as possible. All ten sets of scans first underwent this processing with the DAVID 4 program, which involves manually determining the coarse and fine alignments. A third step was necessary when using the DAVID 4 program, termed "fusion" in the software, in which all of the scans were combined and presumably subsampled in some way to create one .obj file that represented the entire cranium. 3D models for ten crania - two for each individual - were therefore created by using the DAVID 4 program entirely. Due to issues with the CraniAlign program's ability to robustly align scans coarsely for full crania (see section 4.0.2) for a more detailed explanation of this issue), only the fine alignment was run using CraniAlign, using the coarsely-aligned data generated using the DAVID 4 program. The final point cloud was generated by a fusion method which used a well-defined algorithm (refer to section 4.0.2) to create an .obj file that represented the entire cranium. An additional ten 3D models of crania were therefore created using CraniAlign to govern the final alignment results. The resulting data generated consisted of four 3D models per individual - one generated from scans taken on day 1 and processed with DAVID 4; one generated from scans taken on day 2 and processed with DAVID 4; one generated from scans taken on day 1 and processed with CraniAlign; and one generated from scans taken on day 2 and processed with CraniAlign. For clarity, the term "pair" will henceforth refer to a set of 3D models generated using the same program (DAVID 4 or CraniAlign), but using scans of the same crania taken on different days. Refer to the flowchart given below in Figure 4.1.
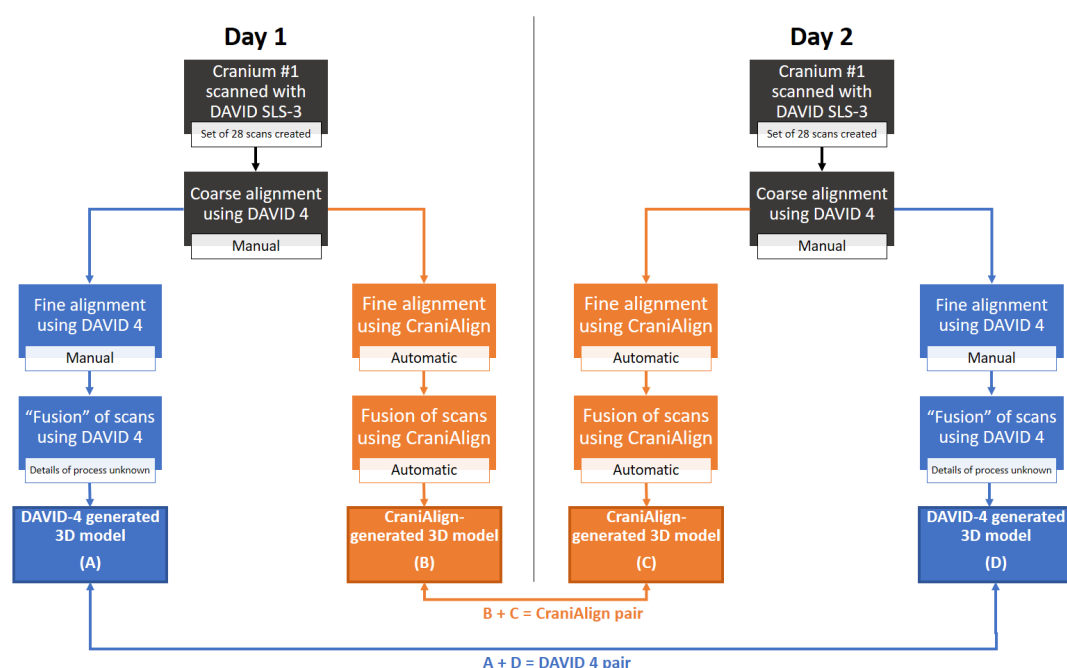
Figure 4.1: A flowchart demonstrating how data was acquired from a single cranium, which was scanned twice. The output is two pairs of 3D models - one pair generated using DAVID 4, and another pair generated using CraniAlign. The difference between the 3D models in a single pair is that they used different sets of scans of the same cranium - one obtained on Day 1 and the other obtained on Day 2. This process was repeated for each of the five crania.

4.0.1   The DAVID 4 Program

To create 3D models using the DAVID 4 software program, the following procedure was loosely followed to create a single 3D model: first, two neighbouring scans to be aligned to one another were chosen, usually using the "Around Y-Axis" parameter, without the use of texture. To align a scan in a different orientation (e.g. aligning the anterior side of a crania placed on its lateral side to the anterior side of a crania placed upright in anatomical position), the "Free" parameter was chosen instead. These alignment parameters allowed the DAVID algorithms to optimize the most probable overlap of the two scans. Once all scans were aligned in this manner, a global fine registration was performed which minimized the error between all scans instead of just between pairs. For the global fine registration, 30 iterations were performed and texture was often used with a weighting of 80. It should be noted, however, that alignment was not always successful, so monitoring of the resulting position of each scan was needed. Re-alignment was often necessary by repeating the same alignment procedure. The DAVID algorithms have a degree of randomness built in to allow for different possible positions to be obtained, so repeating the same alignment procedure would sometimes result in a better

alignment. The degree of randomness, however, cannot be controlled or even determined without access to the DAVID SDK, which is an expensive upgrade.
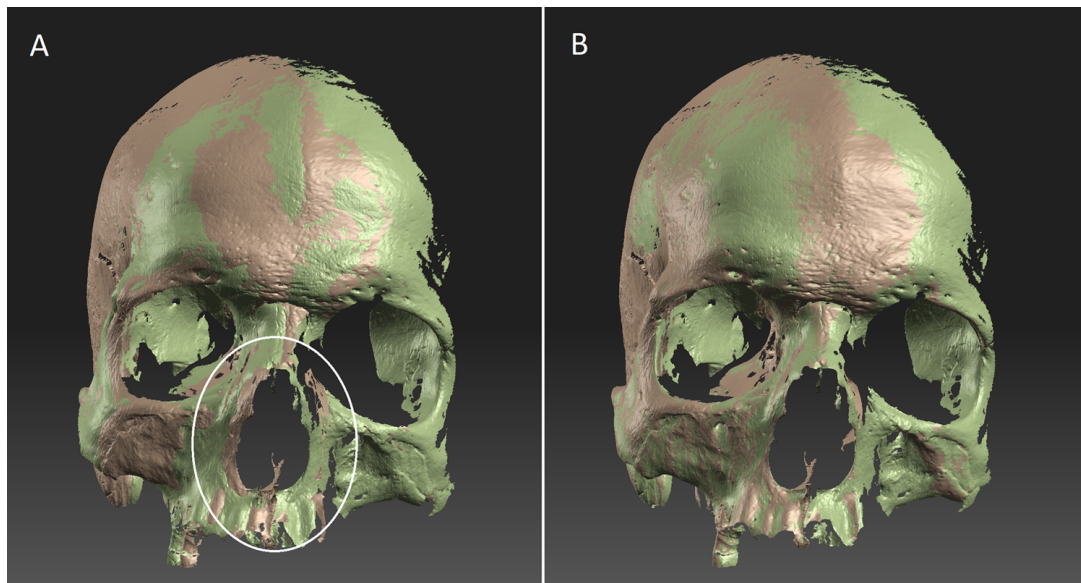
Figure 4.2: A) An example of how misalignment can occur using the DAVID 4 program, which is most evident in the nasomaxillary region (indicated by the white circle). Repeating the same alignment procedure fixed the misalignment, as seen in B). A degree of randomness built into the DAVID algorithms is therefore apparent, but cannot be controlled without access to the DAVID SDK.

Once aligned, the scans were fused. A resolution parameter of 2000 was chosen with a corresponding vertex spacing of 0.174 mm, and the sharpness setting was kept at the default value of 1. Although these parameters for resolution are chosen, that does not necessarily mean that the resulting 3D model, once aligned and fused, has the same vertex spacing. The actual vertex spacing was therefore determined once the 3D models were created.

The possible sharpness values given by the DAVID program are integer values ranging from -3 to +5. Decreasing numbers result in smoother 3D models. The trade-off to a smoothed 3D model, however, is possible loss of detail. The resolution given by the DAVID program is unitless and therefore arbitrary, although the given resolution corresponds to a specific vertex spacing. The possible resolution values and their corresponding vertex spacings are given in Table 4.1 below. For the purposes of the pilot test, it was determined that a resolution parameter of 2000 was the maximum achievable resolution for the crania (28 scans) with an i7-4700MQ dual-core CPU, each with a speed of 2.40 GHz, and 16.0 GB of RAM. The GPU was an NVIDIA GeForce GTX 765M with 2 GB of dedicated memory, which was important to facilitate the demands of the DAVID program. Nevertheless, the computational process of fusing 28 scans

at a resolution higher than 2000 was not possible with these computer specifications.

Table 4.1: Resolution parameters given by the DAVID 4 program and their corresponding vertex spacings.

Resolution Parameter	Vertex Spacing (mm)
250	1.39
500	0.695
700	0.497
1000	0.348
1500	0.232
2000	0.174
3000	0.116
4000	0.0869

4.0.2   The CraniAlign Program

The CraniAlign program was a unique opportunity for a PhD student in Archaeology to become directly involved with the creation and development of a software program that has widespread applications. Although the program is currently named "CraniAlign" and was developed for the alignment and processing of cranial scans, the algorithms and parameters used are meant to be applicable to scans of any object. Due to the intention of creating a program with such a widespread use, more technical knowledge in C++ programming was required than that possessed by a standard PhD student in Archaeology. Consequently, the creation and development of CraniAlign was jointly undertaken with Etienne Pillin, a PhD student in Mathematics with experience in C++ programming. The code itself was written by Pillin, whereas the author of this research project tested the parameters in the program on her data and indicated where the program could be improved. The source code for CraniAlign is not made available at this time since it is currently the intellectual property of Clotho AI, who have chosen not to release it as of yet.

Other than reading and writing files, the CraniAlign program consists of three main functions: coarse alignment, fine alignment, and fusion. The coarse alignment algorithm used was the Super 4PCS (4-Points Congruent Sets) algorithm developed by Mellado and colleagues (2014), and the fine alignment algorithm used was Sparse ICP by Bouaziz and colleagues

(2013). Fusion was a combination of algorithms from Super 4PCS and those that were written by Pillin.

For coarse alignment, Super 4PCS was chosen because it was well-documented by a team of researchers who are still actively developing it, meaning that it was likely that they would be available and keen to assist with the alignment required for this project, if necessary. Additionally, the algorithm reportedly performs robustly with noisy data, and can align scans even if there is a small amount of overlap, as well as on samples where there are little to no geometric features present (Mellado et al. 2014). The latter was particularly important for this PhD project because scans of the posterior and superior sides of the cranium contain no distinct features and consist only of a curved surface. In an example provided by Mellado and colleagues (2014), two scans of a very round bird, both with few features, were able to be roughly aligned even when outliers were added (see Figure 4.3). Ultimately, Super 4PCS was chosen due to its reported success as well as the fact that several other coarse alignment algorithms were investigated and did not generate results as quickly and as satisfactory as those of Super 4PCS.
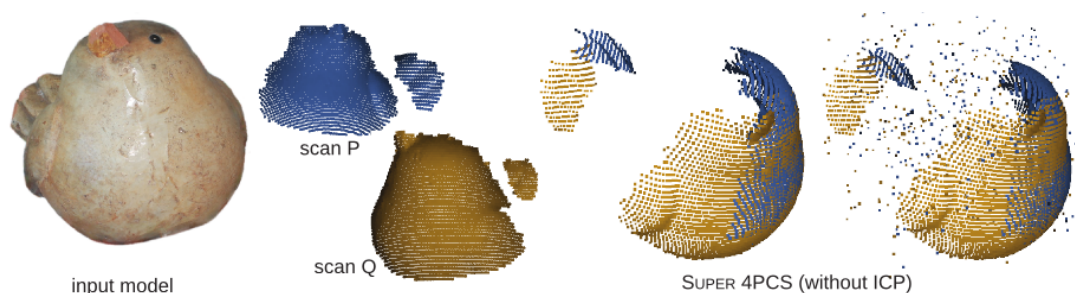
Figure 4.3: An example given by Mellado and colleagues (2014) of two scans with few distinct geometric features that were able to be coarsely aligned to each other using Super 4PCS, even with outliers. This situation is similar to aligning the posterior and superior sides of a cranium.

The output of Super 4PCS, and the resulting coarse alignment, is governed by several parameters: the number of points to be used for the alignment (n), the overlap parameter (o), and the delta parameter (δ). The overlap parameter is the expected overlap between two scans, and the delta parameter is a precision parameter used by Super 4PCS.

For fine alignment, Sparse ICP was chosen because it is the fine alignment algorithm suggested by Super 4PCS (Mellado et al. 2014), and is more robust at dealing with outliers than the traditional ICP algorithm (Bouaziz et al. 2013). The output of Sparse ICP was governed by four parameters: the number of iterations used by Sparse ICP, the precision to be obtained (ϵ)

when minimizing the distance between neighbouring scans, the p parameter which refers to the p-norm that is minimized by the algorithm (for a more technical overview of the p parameter, refer to Bouaziz et al. 2013), and the subsampling precision, which is a parameter added by Pillin in order to increase the speed at which scans were aligned. The subsampling precision refers to the fact that uniform-distance subsampled point clouds were used with Sparse ICP rather than the input point clouds. Two versions of Sparse ICP are available - point-to-point and point-to-plane. For the latter, the normals of the vertices are used for alignment whereas for the former, normals are not required to be calculated. Point-to-plane was used since this method was faster and gave more robust results.

Various implementations of the algorithms were tested during the development of CraniAlign, as well as different input parameters to achieve well-aligned scans. After months of testing, it became evident that coarsely aligning such a wide variety of point clouds robustly was extremely challenging, despite how promising Super 4PCS seemed to be initially. Even when Dr. Mellado from the Super 4PCS team was contacted for assistance, his suggestions were not able to fix the coarse alignment issues present in the cranial samples. No combination of parameters was robust enough to correctly coarsely align all sample crania in this pilot test. In an attempt to mitigate some scans that were not aligned well, CraniAlign needed to be improved.

The major improvement to CraniAlign was to make the alignment a context-agnostic method, meaning that no assumption is made about the order of scans or the intended order of alignment because both could be flawed due to human error in naming scans or rotating the turntable at an imprecise angle during the scanning process. Instead, CraniAlign determines which pairs of scans are most likely to result in a successful alignment, which is achieved by browsing each pair of scans, computing the associated overlap value by attempting to coarsely align each pair using Super 4PCS, and performing the alignment on pairs which have the highest overlap values above a given correctness threshold (c). As a result, if a scan was unable to be coarsely aligned with any other scan that resulted in a computed overlap value meeting or exceeding this threshold, that scan was discarded and not included in the final alignment of the cranium.

Unfortunately, these two features still did not generate results robust enough to be used on several samples; however, it performed quite well on autopsied samples. After months

of iteratively testing and developing CraniAlign, it was finally concluded that it was necessary to use the coarsely-aligned scans from the DAVID 4 program for fine alignment and fusion with CraniAlign for the purposes of the pilot test. For generating the samples needed for this PhD project, the coarse alignment performed well on most autopsied samples so there was value in developing the coarse alignment with Super 4PCS. The scans then underwent the fine alignment with CraniAlign, before the results were fused. Fusion involved merging all the aligned point clouds from each individual scan into one, and then subsampling the resulting point cloud using a uniform distance sampling method. The subsampling was performed with a precision of 0.15 so that the resulting point cloud had a similar number of points to those generated from the DAVID 4 program (roughly 12,000,000 - 15,000,000 points).

4.1   Methodology

In order to compare two paired 3D models to each other, and thus generate the data needed to analyze reproducibility, each 3D mesh first needed to be subsampled to create a point cloud. Using Cloud Compare, the number of points to be subsampled was set to 1,000,000. Normals were automatically generated at this point. Colour information was also added to the point cloud, if applicable. The resulting point cloud is seen below in Figure 4.4 A. Although the point cloud appears to be sparse, due to limiting the number of points so as to be manageable in subsequent analyses, the point size can be increased for visual purposes if needed. An example of a point cloud with an increased point size is seen in Figure 4.4 B.
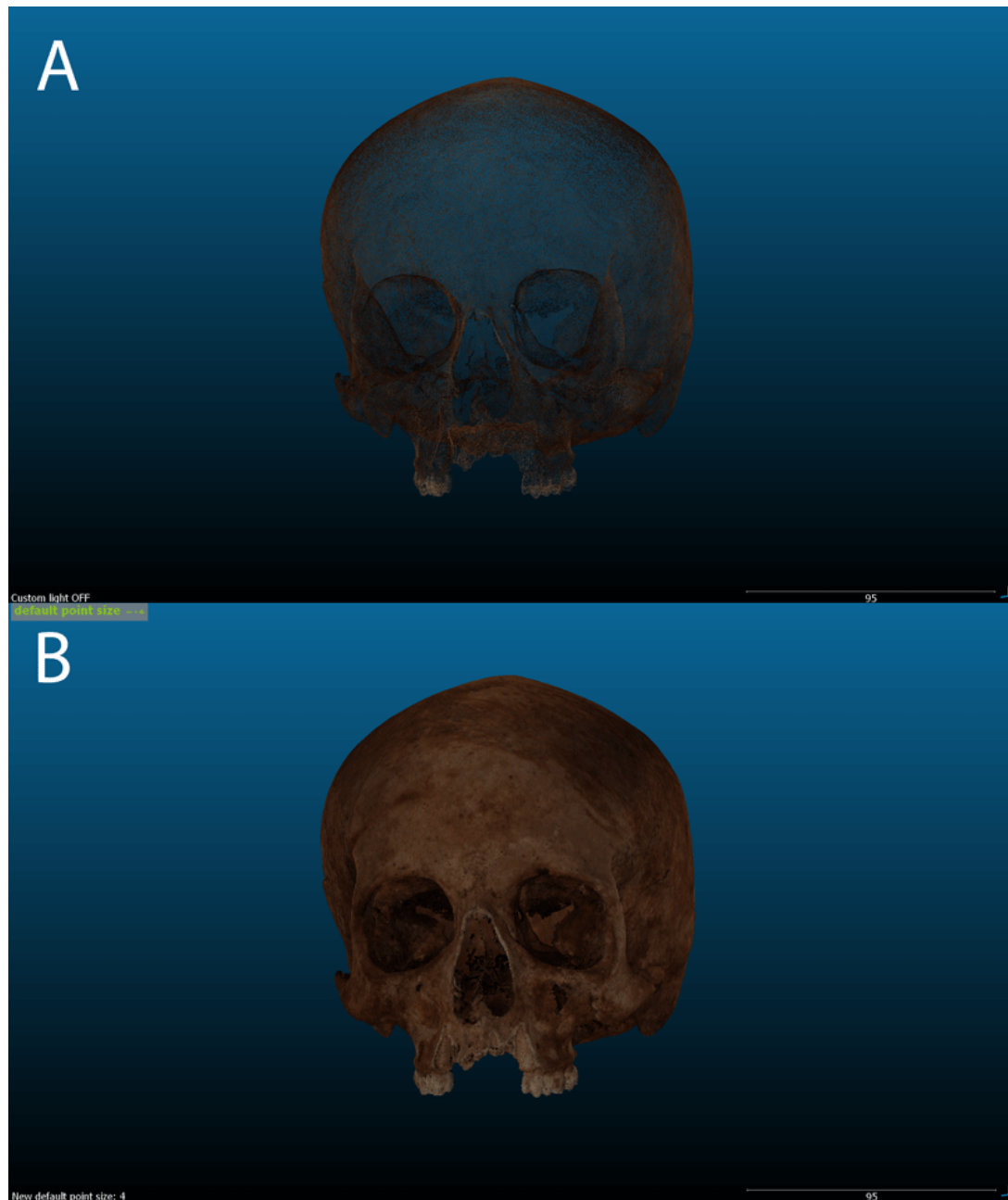
Figure 4.4: A: The point cloud generated from Cloud Compare after subsampling a mesh for 1,000,000 points. **B**: The point cloud after the point size has been increased to 4.

Each pair of point clouds then needed to be aligned to each other, a process which is referred to as "registration", and was therefore performed using the fine registration tool. In all samples, the first point cloud (i.e. the 3D model generated from scans taken on day 1) was used as the reference sample, meaning that it did not move, and the second point cloud was instead rotated and translated to align to the first. The RMSD (Root Mean Square Deviation, although in Cloud Compare it is simply referred to as RMS) was set to the default value of 10^{-5}, which is a precision parameter specific to Cloud Compare. The final overlap, which is an estimation of

how much overlap the two point clouds should have with each other, was set to 100% because the two point clouds should theoretically be identical. The farthest points removal option was enabled, which removed points that were likely to cause error in the registration computation. Consequently, outliers would be removed, such as random points in the background of the 3D model. In order to find the best transformation of the second point cloud to the first point cloud, rotation and translation in the X, Y, and Z axes were allowed. The random sampling limit was set to 100,000 - which is 10% of the number of points making up each subsampled point cloud - and refers to the number of points to be randomly sampled in order to align, or register, the two point clouds. A visual example of registration is given below in Figure 4.5.
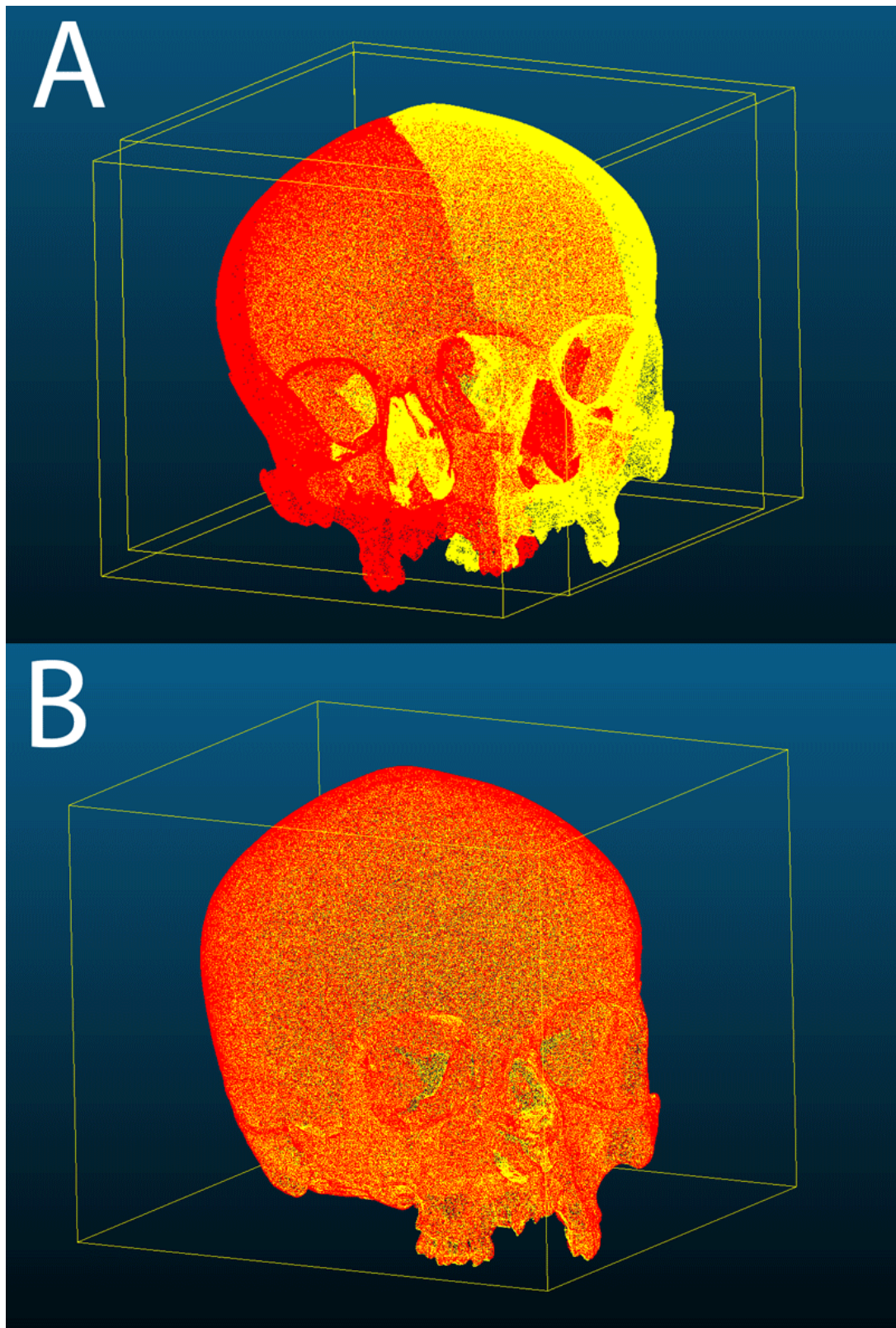
Figure 4.5: A: Two point clouds of the same cranium are loaded in Cloud Compare. The yellow point cloud corresponds to the first 3D model (i.e. created from scans taken on day 1 of data collection) whereas the red point cloud corresponds to the second 3D model (i.e. created from scans taken on day 2 of data collection). In order to determine the difference between the two models, they must be aligned to each other. **B**: The two point clouds have been aligned to each other using the fine registration tool.

After the registration of the two point clouds, the RMSD of the registration was calculated using the same 100,000 points sampled during the registration process. Values approaching 0 demonstrate little difference between the point clouds after the registration, and refer specifically to the average difference between the sampled points. It is important to note that values greater than 0 do not necessarily mean that the two point clouds are different from a practical point of view - it simply means that the scanned object was sampled using different points to generate the point cloud either during the scanning process, the subsampling of 1,000,000 points performed by Cloud Compare, the random sampling limit of 100,000 points for the RMSD calculation, or a combination of any of these three possibilities. It is important to note that due to the method in which Cloud Compare subsamples and calculates RMSD, the results of the RMSD as well as the amount of error/difference between two point clouds (termed "distance computation" in Cloud Compare) may be artificially inflated.

The distance computation simply calculates the distance between each corresponding point in the two registered point clouds, termed the C2C ("cloud to cloud") distance. If two scans are composed of identical points and theoretically registered with an RMSD of 0, then the C2C distance between the two point clouds should also be 0. It is necessary to note that values of 0 for both the RMSD and C2C distance are purely theoretical and are never achieved in practice due to limitations and assumptions associated with registration and distance computation algorithms. Using the distance computation tool, the mean C2C distance between the two point clouds were calculated, as well as the standard deviation, the maximum and minimum C2C distances, and the distribution. It is possible to visualize these distances by using the C2C absolute distances colour ramp feature. Areas of difference, according to increasing distance, are given by blue, green, yellow, orange, and red. An example is given below in Figure 4.6.
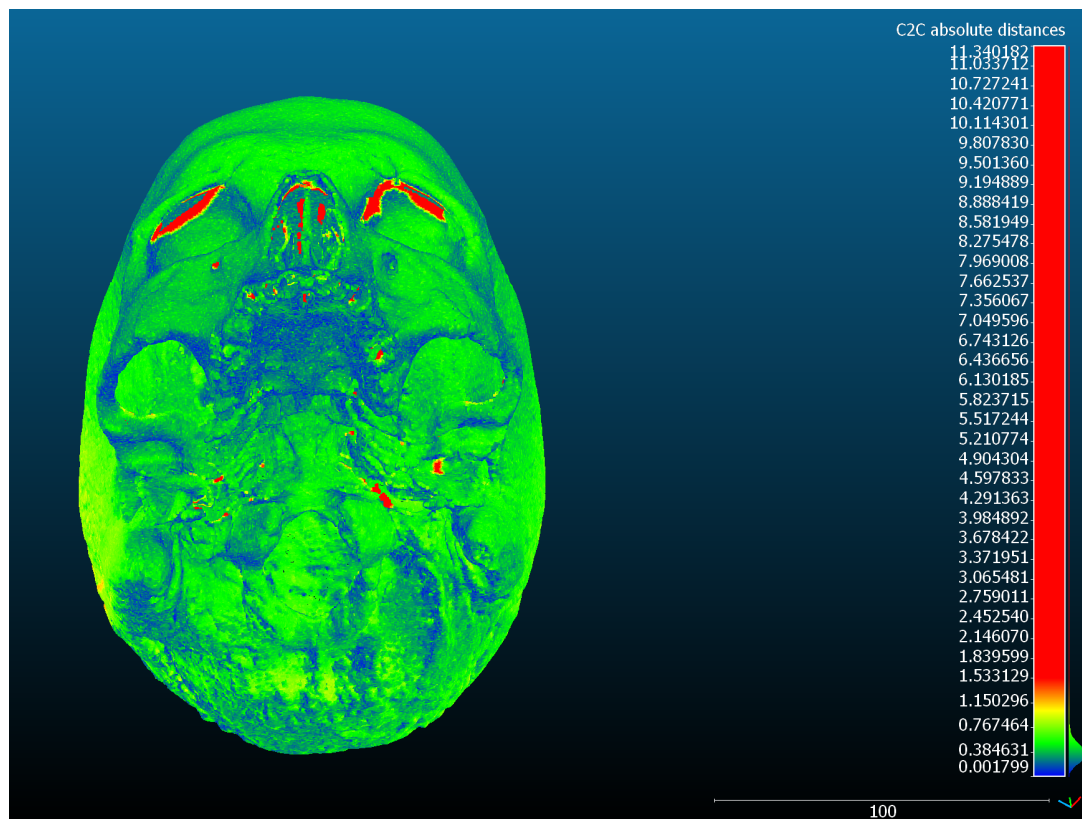
Figure 4.6: An example of a distance computation result. In this image, the C2C absolute distances are given by a colour ramp to the right, with the values and the scale given in millimeters. In this case, areas that differed the most between the two point clouds are the superior roof of the orbits, the interior of the nasal cavity, and in various other canals and foramina on the inferior aspect of the cranium.

4.2 Results

It is of interest in this study to test whether the results are consistent with the claims of DAVID Vision Systems - in particular, the advertised "precision" of the DAVID SLS-3 which has been stated to be 0.05% of the object size. It is unclear whether the term "precision" refers to resolution (explored in 4.2.1), or if it refers to the degree of error one can expect from 3D models generated using the SLS-3 (explored in 4.2.2). The stated 0.05% value will therefore be compared against both resolution and the error obtained in this study to determine what is most likely meant by "precision".

Furthermore, there is no clarification as to what is meant by object size (i.e. whether a maximum dimension is used to determine object size, or whether 'size' refers to 'volume' of the object). Thus, the calibration dimension used for the respective object will be used as

the reference object size. Since crania are calibrated with the 120 mm calibration panels, the maximum expected "precision" is 0.06 mm. Due to the vagueness of how precision is defined by DAVID Vision Systems, however, it is necessary to accept error values up to the same order of magnitude as what is expected. This shall be defined mathematically as

$$p \in [0.1e; 10e]$$

where $e =$ the calculated error given by the respective calibration dimension, and $p =$ "precision". This means that in order for either resolution or error to be considered consistent with the reported "precision", the achieved resolution or error should be 0.006 mm - 0.600 mm. Finally, it should be noted that although all calculations have been undertaken to the sixth decimal point, the results reported here have been rounded to the third decimal point to stay consistent with the same precision to which p is calculated.

4.2.1 Resolution of the 3D Models

Resolution was calculated two different ways in order to compare to the "precision" advertised by the DAVID SLS-3 - according to average vertex spacing ($\bar{R}$), and according to average surface density ($\bar{\rho}$). Vertex spacing (R) is defined as the distance between a point and its nearest neighbour. Surface density (ρ) is calculated by dividing the number of neighbours (N) by the neighbourhood surface (πR^2), and represents the number of points for a circular area with a radius equal to R. It is important to note that surface density and vertex spacing are not necessarily uniform throughout the entire point cloud, which is why the average is taken for both. The minimum surface density values are also reported, since they represent the lowest surface density achieved in this study, and by extension, the limitation of the resolution.

The resolution of the 3D models generated by DAVID 4 and CraniAlign are given below in Tables 4.2 and 4.3. Although only the DAVID 4 3D models are of interest to compare against the advertised "precision", the resolution parameters for CraniAlign are provided for transparency and interest, keeping in mind the fact that the CraniAlign results can be adjusted/improved according to the precision parameter used by the algorithm.

Table 4.2: Resolution-related results for cranial 3D models generated using DAVID 4.

Sample #	Day 1				Day 2			
	# of points	$\bar{R}$ (mm)	$\bar{\rho}$ (mm^{-2})	ρ_{min} (mm^{-2})	**# of points**	$\bar{R}$ (mm)	$\bar{\rho}$ (mm^{-2})	ρ_{min} (mm^{-2})
SMC206	13,060,087	0.157	12.987	3.805	12,068,495	0.164	11.905	3.031
SMC399	13,720,461	0.162	12.195	2.656	12,780,234	0.168	11.236	3.245
SMC417	12,483,730	0.162	12.195	2.592	13,144,141	0.160	12.500	3.245
SMC1142	12,712,477	0.157	12.987	1.707	13,196,521	0.155	13.333	3.788
SMC1248	15,193,593	0.150	14.085	3.596	13,553,039	0.159	12.658	3.731

Table 4.3: Resolution-related results for cranial 3D models generated using CraniAlign.

Sample #	Day 1				Day 2			
	# of points	$\bar{R}$ (mm)	$\bar{\rho}$ (mm^{-2})	ρ_{min} (mm^{-2})	**# of points**	$\bar{R}$ (mm)	$\bar{\rho}$ (mm^{-2})	ρ_{min} (mm^{-2})
SMC206	13,669,340	0.155	13.333	0.851	11,805,345	0.166	11.494	0.845
SMC399	14,180,055	0.162	12.195	0.851	11,616,068	0.176	10.309	0.791
SMC417	13,108,211	0.160	12.500	0.855	12,763,447	0.162	12.195	0.847
SMC1142	12,206,711	0.162	12.195	0.871	11,356,463	0.169	11.111	0.843
SMC1248	14,185,314	0.158	12.821	0.774	10,788,179	0.179	10.000	0.752

From the results given above, vertex spacing remains a possible candidate for what the DAVID SLS-3 refers to as "precision", given that all vertex spacing values for the DAVID-4 3D models fall within the established range of 0.006 mm - 0.600 mm.

4.2.2 Determining Reproducibility

To determine the reproducibility of the 3D models (i.e. the degree of error present between pairs of scans), each paired point cloud generated using either DAVID 4 or CraniAlign was compared to each other using Cloud Compare, and the degree of difference between the two 3D models was calculated. It is important to note that this constitutes the total error present, which encompasses the sum of errors from scanning, aligning, and fusing, as well as from the registration of the two models to each other in Cloud Compare (given by the RMSD), and from the C2C distance computation (i.e. comparison) itself between the two models. Note that all results reported from this point on were generated using the subsampled point clouds (i.e. each point cloud was limited to a maximum number of 100,000 points).

DAVID 4 Comparisons

This section reports the results obtained from comparing pairs of 3D models generated using DAVID 4. The RMSD values are reported as well as the C2C distances. Table 4.4 reports the RMSD values, as well as the number of points used in the calculation.

Table 4.4: RMSD results calculated from the registration of the 3D model pairs generated using DAVID 4.

Sample #	# of points used	Final RMSD from registration (mm)
SMC206	90,771	0.656
SMC399	92,327	0.675
SMC417	83,937	0.751
SMC1142	86,720	0.602
SMC1248	89,245	0.630

The distribution of the C2C distances for each pair aligned and fused with the DAVID 4 program are summarized in Table 4.5, and are also represented using a histogram that displays the data according to standard deviation, given in 4.7. Refer to Appendix G for sample-specific histograms displaying the data according to mean C2C distance.

Table 4.5: C2C distances for pairs of cranium 3D models generated using the DAVID 4 program for alignment.

Sample #	Mean (mm)	Standard Deviation (mm)	Median (mm)	Interquar-tile Range (IQR)	Maximum (mm)
SMC206	0.455	0.311	0.410	0.180	10.909
SMC399	0.423	0.374	0.379	0.224	11.386
SMC417	0.702	0.501	0.570	0.201	9.163
SMC1142	0.411	0.316	0.314	0.185	7.867
SMC1248	0.364	0.449	0.288	0.215	12.480
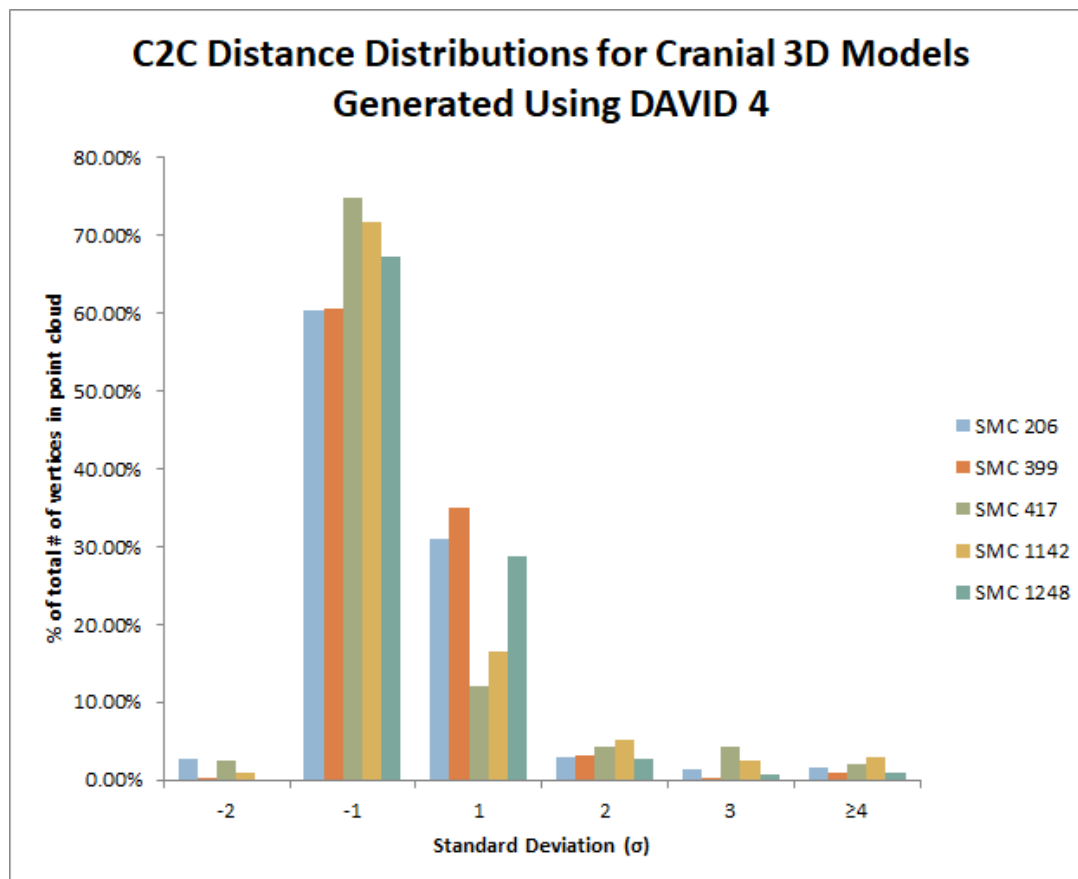
Figure 4.7: A histogram displaying the C2C distances for each cranial point cloud generated using DAVID 4, according to their standard deviation from the mean. Instead of using the absolute number of points that fall within specific deviation categories, the percentage of the total number of points in each point cloud was used in order to standardize the results across all samples.

As seen above in Figure 4.7, most of the points in the 3D models fall within one standard deviation (σ) of the mean C2C distance for each pair. Table 4.6 provides the exact distribution for each pair according to standard deviation.

Table 4.6: C2C distance distribution for pairs of cranial 3D models generated using the DAVID 4 program for alignment, expressed in percentage of total points per standard deviation.

Sample #	% of points within 1σ	% of points within 2σ	% of points within 3σ	% of points $\geq 4\sigma$
SMC206	91.31%	5.73%	1.42%	1.53%
SMC399	95.38%	3.46%	0.30%	0.87%
SMC417	86.83%	6.78%	4.39%	2.00%
SMC1142	88.16%	6.27%	2.57%	3.00%
SMC1248	95.77%	2.67%	0.69%	0.87%

The areas that displayed the most amount of difference between pairs generated using DAVID 4 were mostly features which were difficult to properly illuminate during the scanning procedure, such as various foramina on the inferior side of the crania, structures within the nasal cavity, and alveoli in which teeth are missing. Two notable exceptions exist - the occipital squama of SMC 417 and the posterior part of the cranium roughly around lambda on SMC 1142 - seem to exhibit high levels of difference (i.e. ≥ 4 standard deviations from the mean). Figures 4.8, 4.9, and 4.10 show the areas in which the majority of the differences occur.
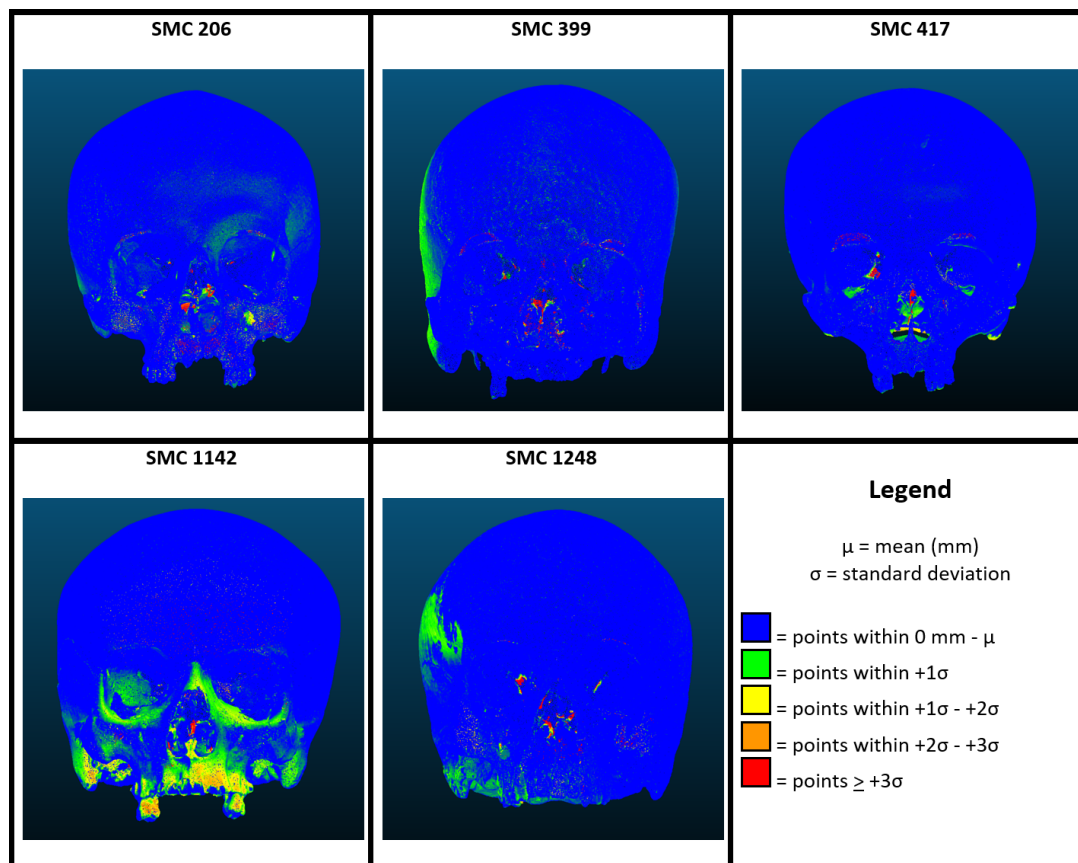
Figure 4.8: After the C2C distances for each cranial pair generated using DAVID 4 were computed, each point was coloured according to their deviation from the mean C2C distance. In the anterior view, areas that deviated the most are usually internal structures within the nasal cavity. SMC 1142 is an exception, where the inferior portion of the maxilla and teeth show great deviation, as well as the mastoid processes.
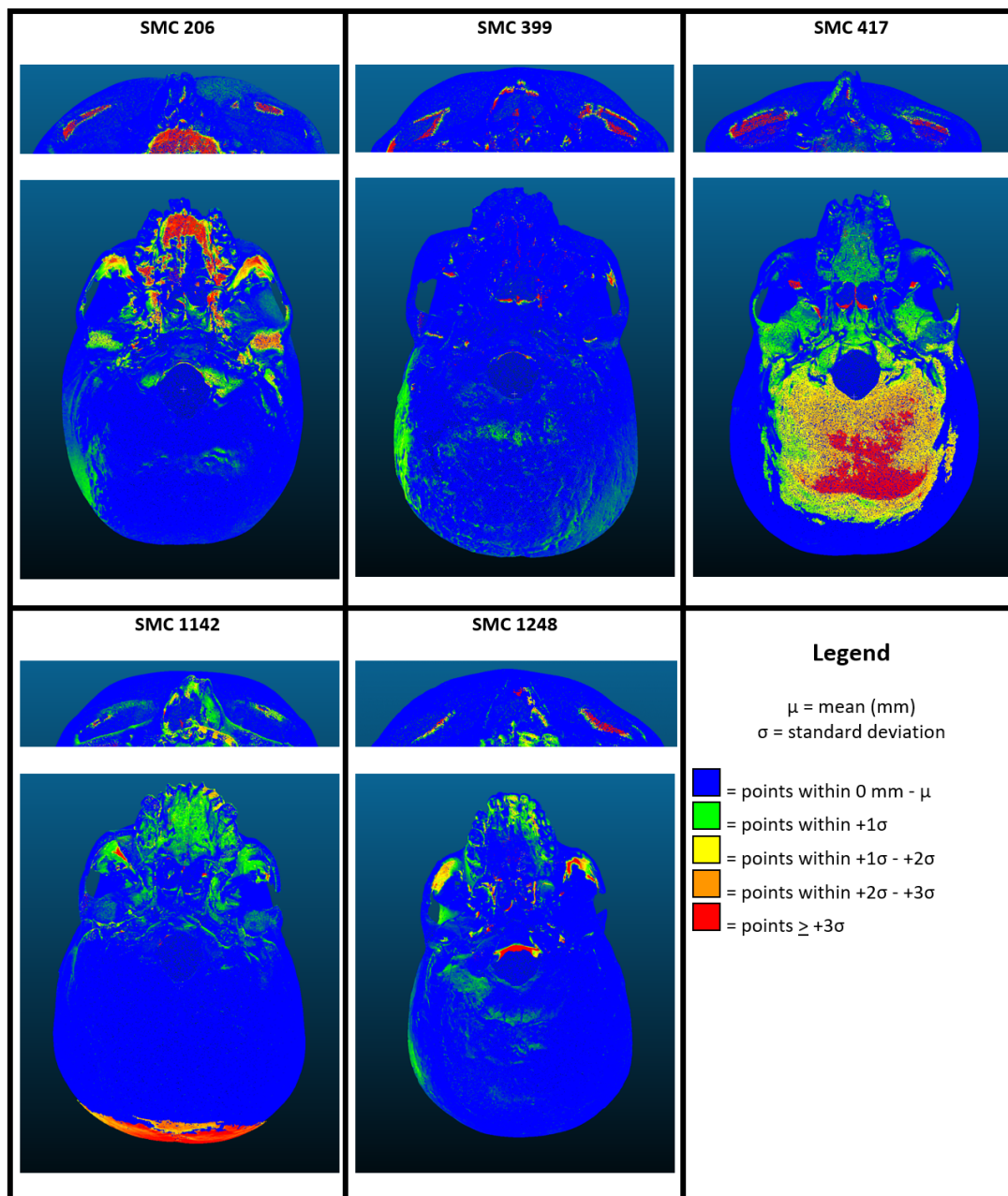
Figure 4.9: After the C2C distances for each cranial pair generated using DAVID 4 were computed, each point was coloured according to their deviation from the mean C2C distance. Along with the standard inferior view, the crania were tilted slightly so as to be able to visualize the superior roof of the orbital wall. Areas which consistently show high deviation are the superior roof of the orbits; the posterior aspect of the zygomatic arches; and within the alveolar cavities. SMC 417 and SMC 1142 show marked exceptions in the occipital squama and the posterior part of the cranium around lambda, respectively.
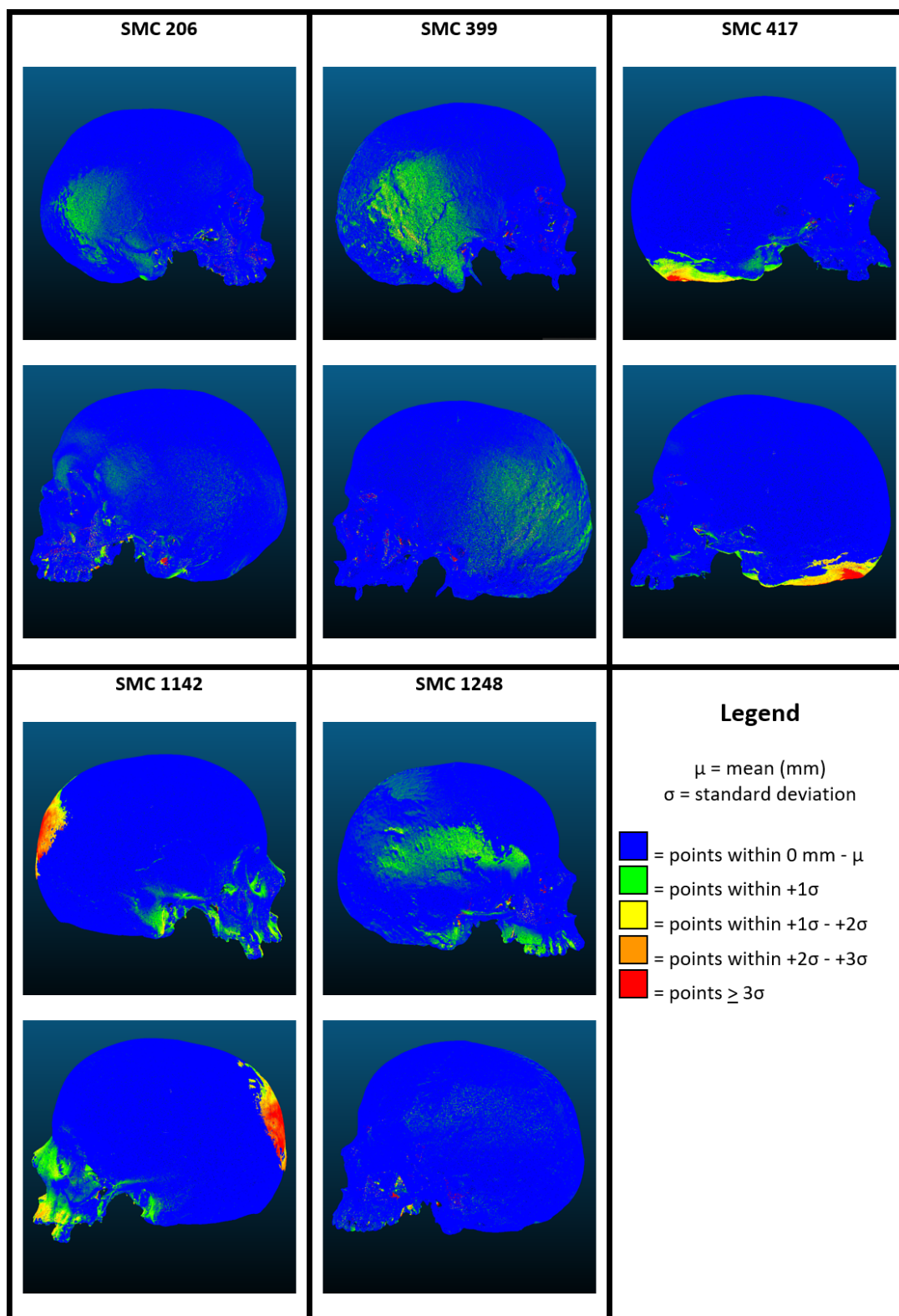
Figure 4.10: After the C2C distances for each cranial pair generated using DAVID 4 were computed, each point was coloured according to their deviation from the mean C2C distance. In the lateral view, the aforementioned anomalies in other views can be visualized - namely, the occipital squama of SMC 417 and the posterior part of the cranium of SMC 1142.

CraniAlign Comparisons

This section reports the results obtained from comparing pairs of 3D models generated using CraniAlign. The RMSD values are reported as well as the C2C distances. Table 4.7 reports the RMSD values, as well as the number of points used in the calculation.

Table 4.7: RMSD results calculated from the registration of the 3D model pairs generated using CraniAlign.

Sample #	# of points used	Final RMSD from registration (mm)
SMC206	96,059	0.785
SMC399	90,436	0.836
SMC417	86,773	0.711
SMC1142	71,627	0.664
SMC1248	84,243	0.739

The distribution of the C2C distances for each pair aligned and fused with the CraniAlign program are summarized in Table 4.8, and are also represented using a histogram that displays the data according to standard deviation, given in 4.11. Refer to Appendix I for sample-specific histograms displaying the data according to mean C2C distance.

Table 4.8: C2C distances for pairs of cranium 3D models generated using the CraniAlign program for alignment.

Sample #	Mean (mm)	Standard Deviation (mm)	Median (mm)	Interquartile Range (IQR)	Maximum (mm)
SMC206	0.417	0.264	0.371	0.269	11.066
SMC399	0.496	0.322	0.431	0.358	8.825
SMC417	0.432	0.354	0.337	0.289	11.225
SMC1142	0.589	0.617	0.355	0.363	8.842
SMC1248	0.403	0.310	0.348	0.241	12.167
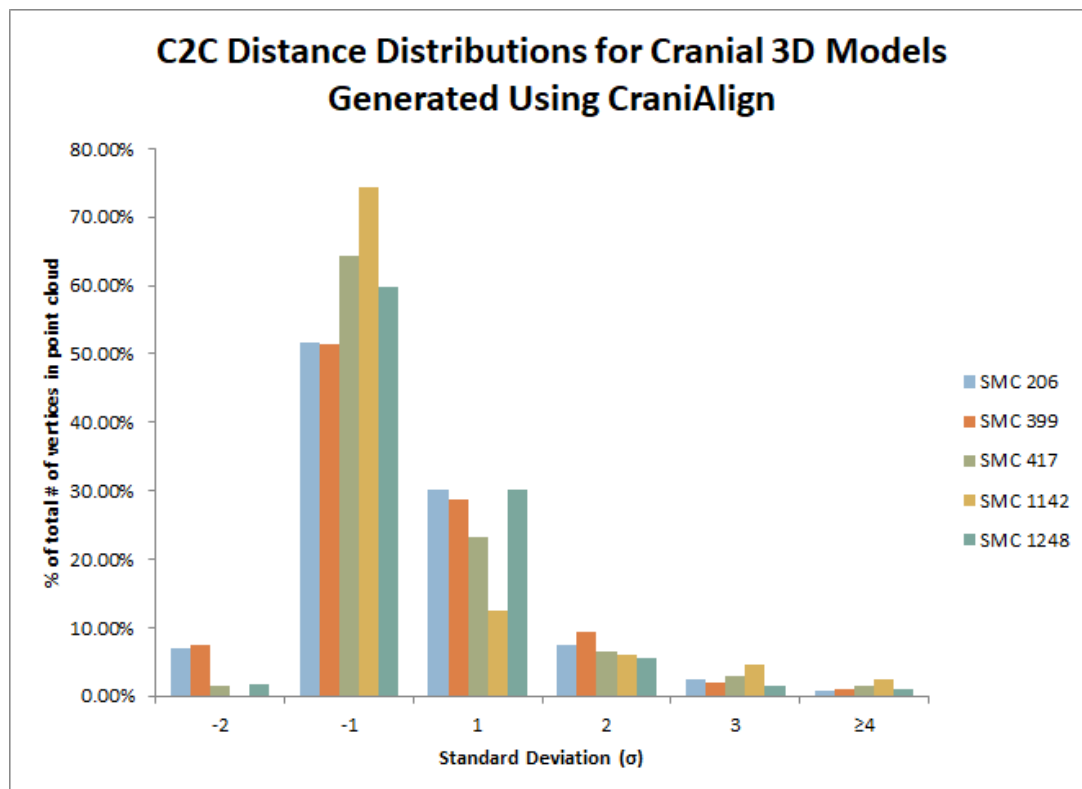
Figure 4.11: A histogram displaying the C2C distances for each cranial point cloud generated using CraniAlign, according to their standard deviation from the mean. Instead of using the absolute number of points that fall within specific deviation categories, the percentage of the total number of points in each point cloud was used in order to standardize the results across all samples.

As seen above in Figure 4.11, most of the points in the 3D models fall within one standard deviation (σ) of the mean C2C distance for each pair. Table 4.9 provides the exact distribution for each pair according to standard deviation.

Table 4.9: C2C distance distribution for pairs of cranial 3D models generated using the CraniAlign program for alignment, expressed in percentage of total points per standard deviation

Sample #	% of points within 1σ	% of points within 2σ	% of points within 3σ	% of points $\geq 4\sigma$
SMC206	82.08%	14.64%	2.40%	0.88%
SMC399	80.16%	17.00%	1.93%	0.91%
SMC417	87.67%	7.96%	2.92%	1.44%
SMC1142	86.81%	6.05%	4.66%	2.48%
SMC1248	90.12%	7.41%	5.62%	0.92%

The areas that displayed the most amount of difference between pairs generated using

CraniAlign were mostly the inferior surface, inside the orbits and nasal aperture, and the alveolar areas of the maxilla, as well as the teeth. SMC 417 and SMC 1142 also displayed notable difference in the superior surface of the crania. These differences are displayed in Figures 4.12, 4.13, and 4.14.
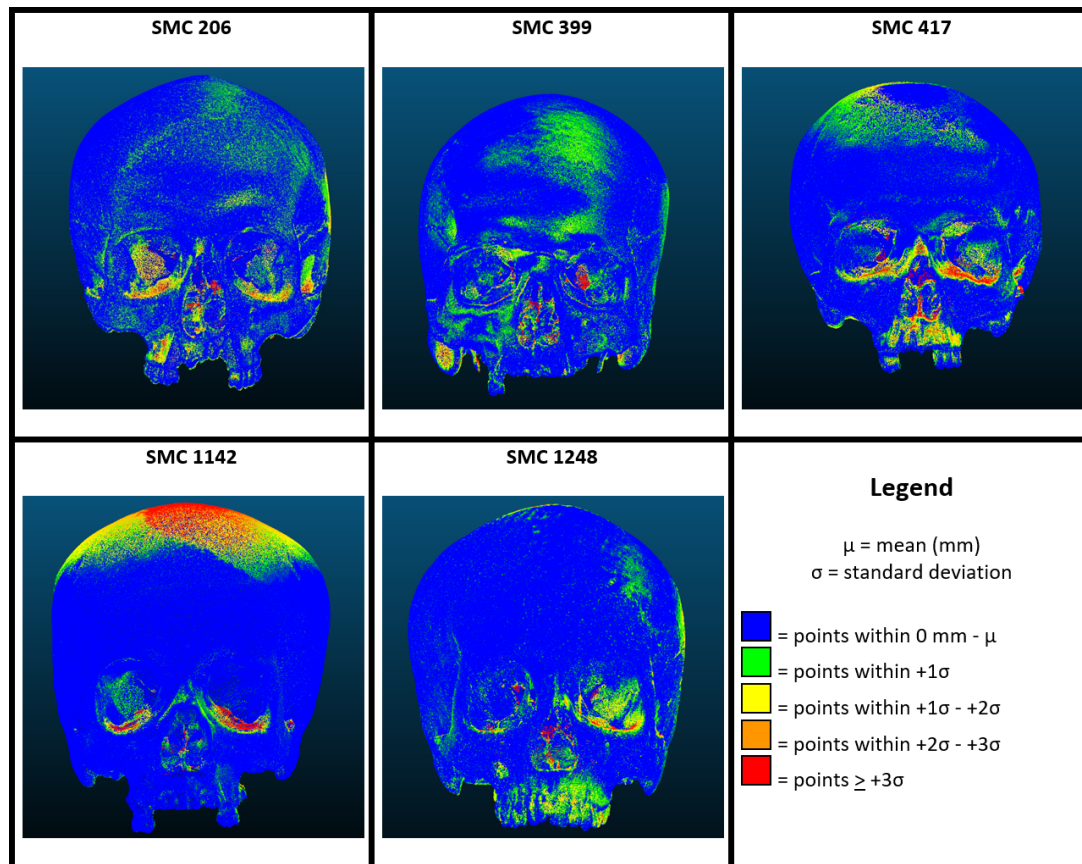
Figure 4.12: After the C2C distances for each cranial pair generated using CraniAlign were computed, each point was coloured according to their deviation from the mean C2C distance. In the anterior view, areas that deviated the most are usually internal structures within the nasal cavity and orbits. SMC 417 and SMC 1142 also show great deviation in the inferior portion of the maxilla and teeth.
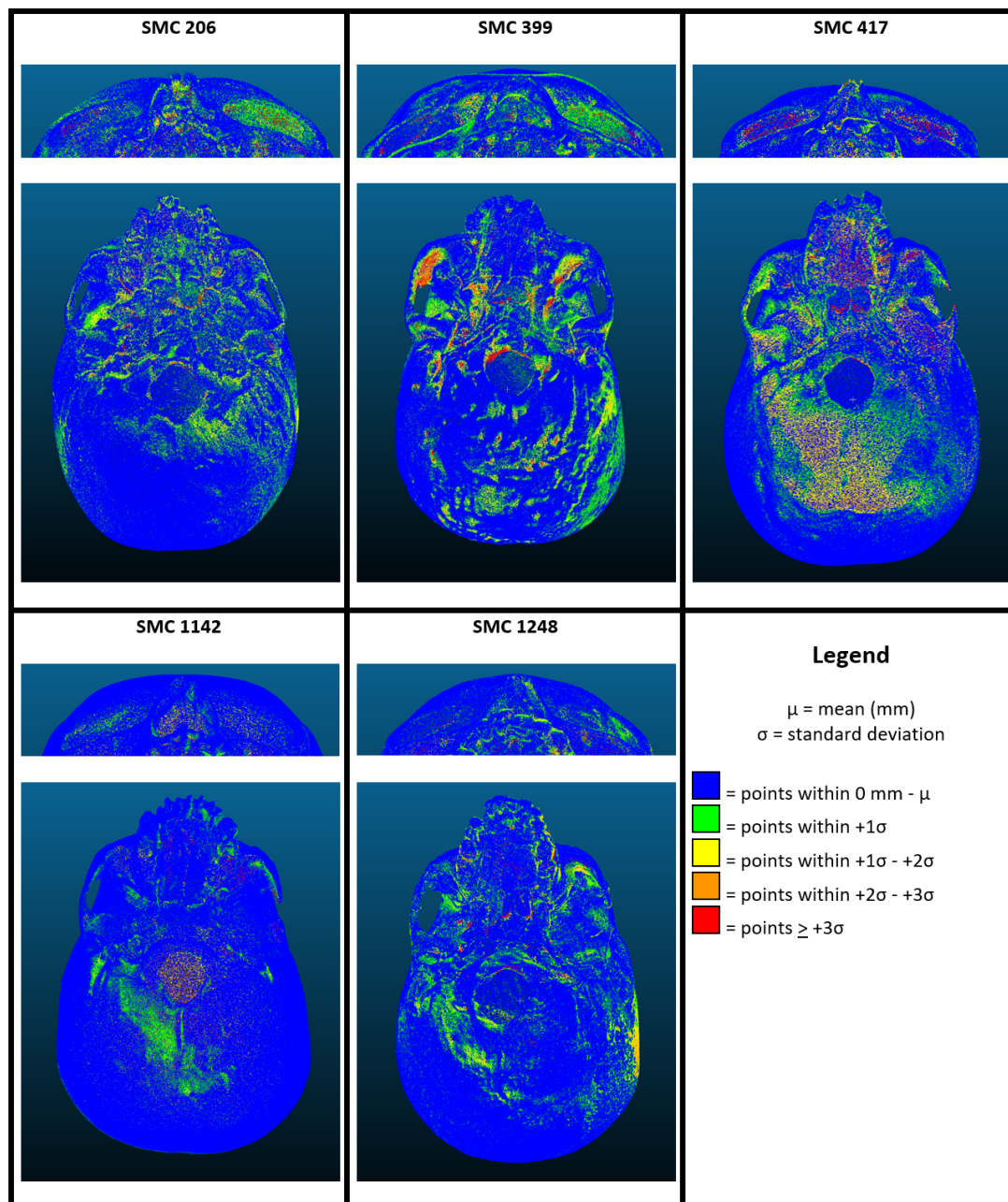
Figure 4.13: After the C2C distances for each cranial pair generated using CraniAlign were computed, each point was coloured according to their deviation from the mean C2C distance. Along with the standard inferior view, the crania were tilted slightly so as to be able to visualize the superior roof of the orbital wall. Areas which consistently show high deviation are the superior roof of the orbits; the posterior aspect of the zygomatic arches; and the various foramina on the basilar part of the cranium.
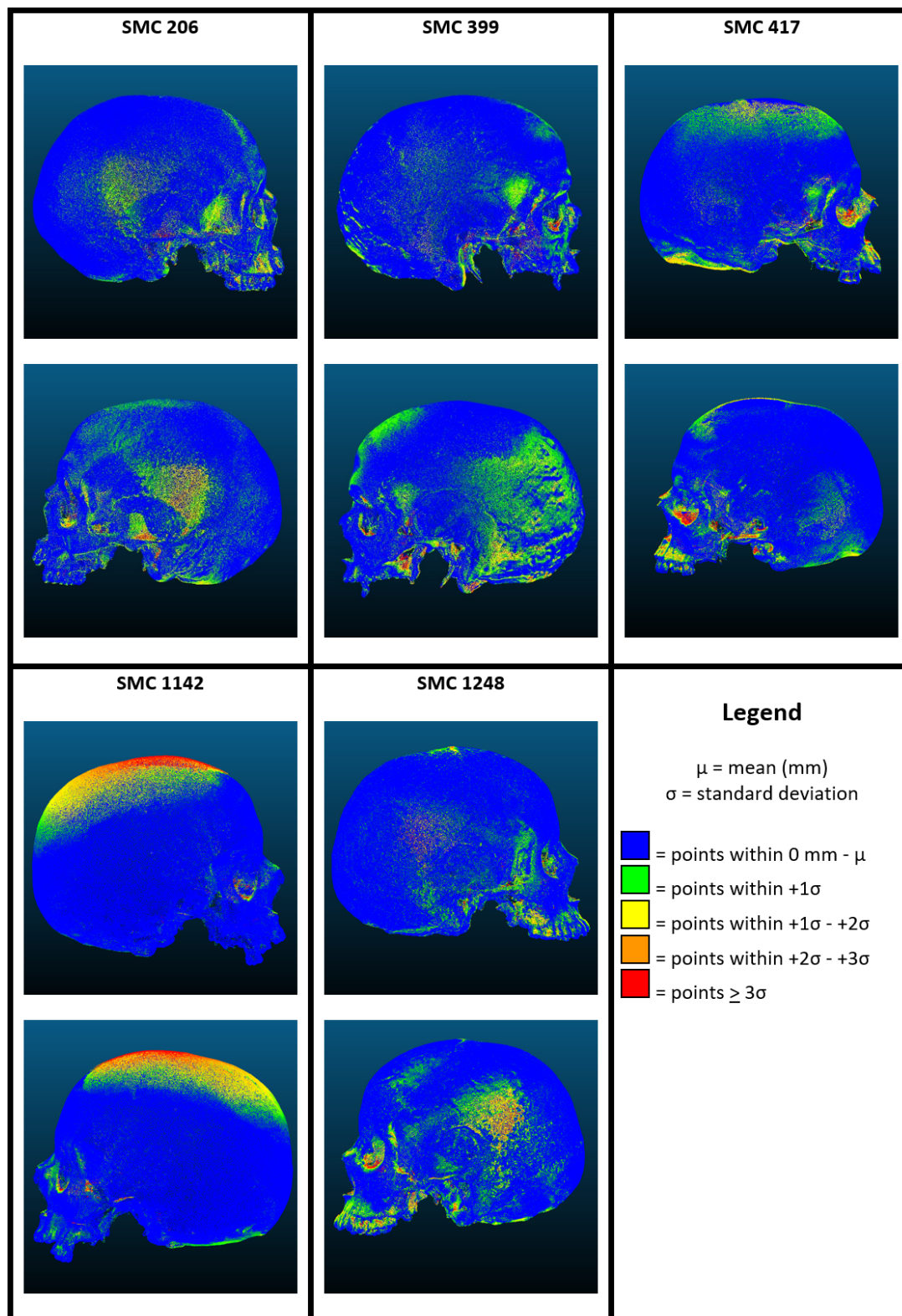
Figure 4.14: After the C2C distances for each cranial pair generated using CraniAlign were computed, each point was coloured according to their deviation from the mean C2C distance. In the lateral view, the aforementioned anomalies in other views can be visualized.

4.2.3 Determining Reliability

In this pilot test, reliability is defined as having similar C2C distances (i.e. error) for all pairs of point clouds within an alignment method (i.e. between all samples aligned with DAVID 4, or between all samples aligned with CraniAlign). Due to the extremely large number of data points for each sample ($\geq$ 1,000,000), statistical tests were not necessary since the use of such a large number of data points would only result in an increased sensitivity to differences between samples. Consequently, statistical significance would be achieved even if differences are inconsequential from a practical perspective. Instead, for the purposes of this study, the degree of reliability will be established by examining the range of error achieved within a single alignment method. The alignment method with the narrowest range of error is deemed more reliable.

Both the mean and median errors were used for this comparison, which were previously given in Tables 4.5 and 4.8. As seen in the histograms above in 4.2.2, the C2C distances are not normally distributed so the means are potentially affected by outliers or skewed data, which is why the medians are also reported. The non-normal distributions of the data are proven in Appendices H and J, in which the C2C distributions for each sample were fitted to 49 different probability distribution functions (PDF's) to determine which PDF best approximates the data.

The median error for DAVID 4 3D models ranged from 0.288 mm - 0.570 mm (a difference of 0.282 mm), whereas the median error for CraniAlign 3D models ranged from 0.337 mm - 0.431 mm (a difference of 0.094 mm). The mean error for DAVID 4 3D models ranged from 0.364 mm - 0.702 mm (a difference of 0.338 mm); for CraniAlign models, the mean error ranged from 0.403 mm - 0.589 mm (a difference of 0.186 mm). It can therefore be concluded that CraniAlign produces results that are more reliable than those from DAVID 4.

4.3 Discussion

Resolution

For DAVID 4 samples, the average vertex spacing was consistent with the target vertex spacing provided by DAVID 4 based on their resolution parameter (i.e. Table 4.1), since the

achieved average vertex spacing was actually slightly less than the target. This means that the DAVID 4 3D models had a slightly higher resolution than what was expected. It should be noted, however, that a side project that scanned, aligned, and fused mandibles with the DAVID 4 resulted in average vertex spacing values that were consistently twice what was expected by DAVID 4. Although the results for this pilot test were consistent with what was expected, they are apparently not supported or reproduced by other tests that used the DAVID 4. It is possible that DAVID 4 is not actually fusing the scans to the correct resolution parameter and the results of this pilot test were a coincidence, or Cloud Compare and DAVID 4 calculate vertex spacing differently (and, again, the results of this pilot test were coincidentally consistent). It is impossible to determine precisely where the discrepancy lies without more transparent documentation from DAVID 4.

CraniAlign samples achieved average vertex spacing values and average surface density values that were very similar to what was achieved with DAVID 4. The major difference between the two alignment methods in terms of resolution is the values obtained for the minimum surface density, which is a measure of the resolution limitation. CraniAlign samples had minimum surface densities ranging from 0.752 mm^{-2} - 0.871 mm^{-2}, whereas DAVID 4 had values ranging from 1.707 mm^{-2} - 3.805 mm^{-2}. Essentially, this means that DAVID 4 samples had a greater minimum resolution (i.e. more points per circular area) than CraniAlign. It is possible that this discrepancy between the two alignment methods is due to the fact that CraniAlign uses a uniform distance sampling method (as opposed to a random sampling method, although this can be undertaken by CraniAlign as well) during fusion. An example of the difference between the two sampling methods is given below in Figure 4.15. The uniform distance sampling method would therefore discard points that are too close together such that a given distance or area is represented by one point. As the name suggests, the resulting 3D point cloud has points that are uniformly distributed. Conversely, a random sampling method would mean that no such distribution of points is guaranteed, and points that are very close together in a given area could potentially be used in the fused point cloud. The higher values from DAVID 4 3D models could therefore be outliers. Again, however, it is impossible to determine if the sampling method during fusion is indeed the cause of the difference seen between CraniAlign and DAVID 4 models, since there is no such documentation on the DAVID 4 fusion process.
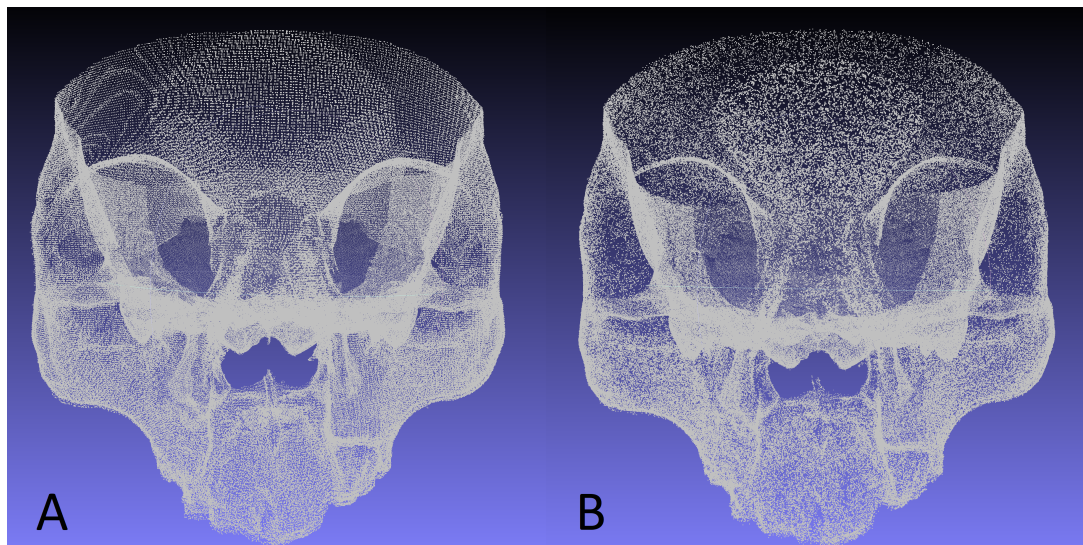
Figure 4.15: A) An example of a point cloud that has been fused with a uniform distance sampling method using CraniAlign. Note the even distribution of the points that represent the geometry of the autopsied cranium. B) An example of the same point cloud that has been fused with a random sampling method using CraniAlign. The result is a geometry that is less "smooth" than seen in A).

Given the similar results for average vertex spacing and average surface density between the two alignment methods, it can be concluded that CraniAlign successfully meets the industry standard set by DAVID 4 in terms of resolution. The resulting point clouds generated by CraniAlign are therefore acceptable for research purposes due to the transparency of the algorithms used and the quality of the results in terms of resolution. Finally, the "precision" advertised by the DAVID SLS-3 is most consistent with the vertex spacing values achieved in this study, so it is likely that "precison" actually refers to resolution.

Reproducibility & Reliability

Reproducibility, or the degree of error, was determined by calculating the total error (given by the C2C distances), what could have contributed to the total error, and where these errors occurred in the cranial 3D models. The range of error for each alignment method was therefore used to establish which method gave more reliable results. It was determined that the CraniAlign 3D models had a smaller range of error (medians differed by 0.094 mm; means differed by 0.186 mm) than DAVID 4 3D models (medians differed by 0.282 mm; means differed by 0.338 mm). Consequently, the use of CraniAlign produces results that differ less across all samples, and is therefore more reliable than the results produced by DAVID 4, which vary more.

The error from the scanning protocol, the scanning procedure itself, the alignment/fusion method, the subsampling of points, the registration of the two 3D models in Cloud Compare, and the process of calculating the RMSD is summed up by the RMSD values themselves. The RMSD values for DAVID 4 3D models ranged from 0.602 mm - 0.751 mm, and from 0.664 mm - 0.836 mm for CraniAlign 3D models. The difference in values between the two alignment methods alludes to the error that is attributable to the alignment/fusion method, and the subsampling of points (which is random in Cloud Compare) used for the registration of the models and the calculation of the RMSD. Without testing Cloud Compare's capabilities directly, the degree of error that can be attributed to the subsampling and calculation of RMSD cannot be isolated.

The RMSD values are larger than the mean and median error values for both DAVID 4 and CraniAlign 3D models; however, it should be remembered that RMSD is calculated differently than C2C distance. RMSD, as per the name, is the square root of the sum of all squared error whereas C2C distance is simply the distance between two points in a point cloud, and is calculated for every point in a 3D model. As seen in Tables 4.5 and 4.8, the maximum C2C distance values are two orders of magnitude greater than the mean and median values. If these outlier values were used in the RMSD calculations, the RMSD values would have been inflated because RMSD calculations are sensitive to outliers. This would explain why the RMSD values are higher than the mean and median C2C distance values.

The areas where the total errors occurred varied according to sample, although in general the areas that were most reproducible were the anterior and lateral sides, regardless of alignment method. The areas that had the most error, and were therefore less reproducible, tended to be the inferior surface, particularly the areas portraying the foramina and canals in the CraniAlign samples (see Figure 4.13). It is possible that the same amount of error is present in the DAVID 4 samples, but due to the wider range of error, these areas are still coloured blue since they fall within one standard deviation of the mean (see Figure 4.9).

SMC 417 in particular produced interesting results in the inferior surface. In both the CraniAlign and DAVID 4 samples, the entire occipital showed significant error (i.e. errors greater than 2 standard deviations, given by orange and red colours). The fact that this same area is highlighted in both methods suggests that this particular error is due to the scanning procedure and/or protocol. It is possible that the occipital bone became loosened at some point between the two days of scanning, causing the position of the occipital to be slightly different

on the second day of scanning compared to the first. Since the Saint Mary's Church collection at the University of Leicester is a teaching collection, it is possible that the samples were handled after scanning on the first day and before scanning on the second day, causing small modifications/damage to the bone which is reflected in the results of this pilot test.

SMC 1142 also showed significant error in both the CraniAlign and DAVID 4 samples, although in different regions. In the DAVID 4 samples, the areas that exhibit the most error are the postero-medial aspects of the two parietal bones. In the CraniAlign samples, the problematic area is the entire superior surface of the cranium. The size and location of each of these two problematic areas are consistent with the size and location of a single scan. It is therefore likely that in each of the alignment methods, a single scan was misaligned for one of the 3D models in the pair, which explains the high amount of error in this area. As previously mentioned, the superior surface of the cranium is the most problematic to align due to the lack of geometric features, so the chance of misaligning scans in this area is high. Furthermore, since coarse alignment was undertaken by DAVID 4 for both alignment methods, the final coarse alignment is determined by a visual inspection. Due to the lack of geometric features, it is possible that the scans appeared to be aligned well but actually were not. The success of CraniAlign's fine alignment is predicated on the assumption that scans are already coarsely aligned properly, so this misalignment issue would not have been completely fixed by CraniAlign. This explains why the error is present in both the DAVID 4 and CraniAlign samples, albeit due to different scans.

4.4   Conclusion

In conclusion, this pilot test was successful in facilitating the creation of CraniAlign, and establishing the fact that CraniAlign's performance is on par with the industry standard set by DAVID 4. Performance was evaluated by determining the resolution, the reproducibility, and the reliability of the resulting 3D models. CraniAlign was able to produce 3D models with average vertex spacing and average surface density values comparable to those from DAVID 4, while providing a more transparent means of fusion. Both DAVID 4 and CraniAlign 3D models displayed similar results in the areas of the cranium that were most or least reproducible, but CraniAlign produced 3D models with a narrower range of error and lower mean and median errors. CraniAlign therefore has a higher degree of reproducibility and reliability than DAVID 4.

Additionally, the automation of CraniAlign is a huge advantage over the manual input needed for DAVID 4, especially for research purposes in which a large number of scans and samples need to be processed. There are limitations to the CraniAlign program in its current version, however, but these are discussed in Chapter 5.

In the context of this research project, the point clouds that were ultimately used in the machine learning algorithm consisted of 2500 points, and not 100,000 as was used in this pilot test. This is because point clouds that consist of exactly 2500 points needed to be used in order to ensure that the resulting neural network fit on the memory of the Graphics Processing Unit (GPU) (Qi et al. 2016), so the number of points was a limiting factor for the machine learning analysis. The use of 2500 points in the pilot test, however, was not appropriate or justifiable because it would severely limit the ability to examine how the different steps of creating, aligning, and fusing meshes affect the resolution. Nevertheless, the results of this pilot test are still applicable to the overall goal of this research project because they establish the inherent limitations and sources of error in the data that were to be used in the machine learning algorithms, before any analyses took place.

This pilot test was also the first study of its kind, as far as the researcher is aware, because it successfully evaluated the properties of 3D models in order to validate their use in research, as well as in forensic science analyses. The comparison between a proprietary software such as DAVID 4 and an open-source program like CraniAlign stresses the continued need for transparent methods of analyses for research and forensic science. Black box algorithms face issues of admissibility in court, and do not abide by the Daubert standards (Daubert v. Merrell Dow Pharmaceuticals 1993); neither do they allow for an informed approach to research analyses. This pilot test has therefore successfully demonstrated the limitations of data interpretation when using black box algorithms, since some conclusions cannot be drawn with any degree of certainty unless further information is provided on the algorithms used in proprietary software. If higher education and research are to advance, it is necessary for good scientific protocols to be followed, which includes transparency in the methodologies used for analysis.

Chapter 5

Exploring Cranial Sexual Dimorphism in Different Populations Using Deep Learning

Supervised deep learning, which has been established as a powerful tool for supervised learning in particular (Goodfellow et al. 2016), is used in this study as a proof of concept to classify cranial point cloud data using sex and population information as the associated labels. The goal is to establish three models, each with an acceptable accuracy (i.e. $\geq$ 80% correct classification), that can be reliably used on forensic or archaeological samples - one for classifying sex regardless of ancestry/population, another for classifying ancestry/population regardless of sex, and a third for classifying based on both sex and ancestry/population. By establishing these models, three methods of sex and ancestry assessment will be created that meet the Daubert criteria (1993), can be used by any researcher or analyst without inter- or intraobserver error, and will be the first to establish the use of deep learning to point cloud data in forensic anthropology and osteology. For this study to be considered successful, a neural network must achieve an accuracy of at least 80% when tested on both the training dataset and the evaluation dataset. The results of the models for sex classification and both sex and population classification are compared to the sex assessment results from Chapter 3 (Cranial Sexual Dimorphism in Various Populations) to determine whether the neural networks' parameters provide a higher or lower rate of classification, and what can be improved in both the

visual assessments and the training of the neural networks. The result of the neural network for population classification are discussed according to their theoretical implications on the idea of assessing ancestry skeletally.

5.1 Methodology

A program was created for this project in Python to facilitate the deep learning analysis required for this study (see https://bitbucket.org/JessicaFrances/workspace/projects/POIN for the repository). The program first browses a folder path for the point cloud data and associated labels and allocates a fraction of the samples as holdout or evaluation samples (in this case, 20% of the total number of individuals were designated to be the holdout samples). Next, the program creates the neural network according to sex, ancestry, or sex and ancestry depending on which model is to be generated. The neural network creation was done by modifying the deep learning algorithm given by PointNet (Qi et al. 2016) in order to recognize the holdout and evaluation samples. There were two outputs: 1) in the Linux terminal, the training and evaluation accuracies obtained by the neural network were printed; and 2) in a designated file folder, a .csv file was generated with the results of how each individual in both the training and evaluation datasets were classified.

PointNet (Qi et al. 2016) was the algorithm used in this project, which can perform two types of tasks - classification and segmentation. Classification refers to the problem of identifying the group to which a sample belongs, whereas segmentation refers to the problem of identifying the different parts of a given sample (refer to Figure 5.1 for an example of segmentation). Classification - which is the main task of interest in this study - is a task for which deep learning has been recognized to be the best tool to use due to its excellent performance in object recognition (e.g. Krizhevsky et al. 2012; Ioffe and Szegedy 2015; Goodfellow et al. 2016). PointNet (Qi et al. 2016) was deemed as the most suitable and promising choice for use in this study mainly because it both utilizes deep learning for classification tasks and directly accepts point cloud data as an input. Although PointNet does compute probability densities as part of its algorithm (Qi et al. 2016), it reports the category for which the maximum probability was calculated for a given sample. This is useful in a study such as this one where it is of interest whether an individual was correctly categorized, i.e. a binary result is desirable.
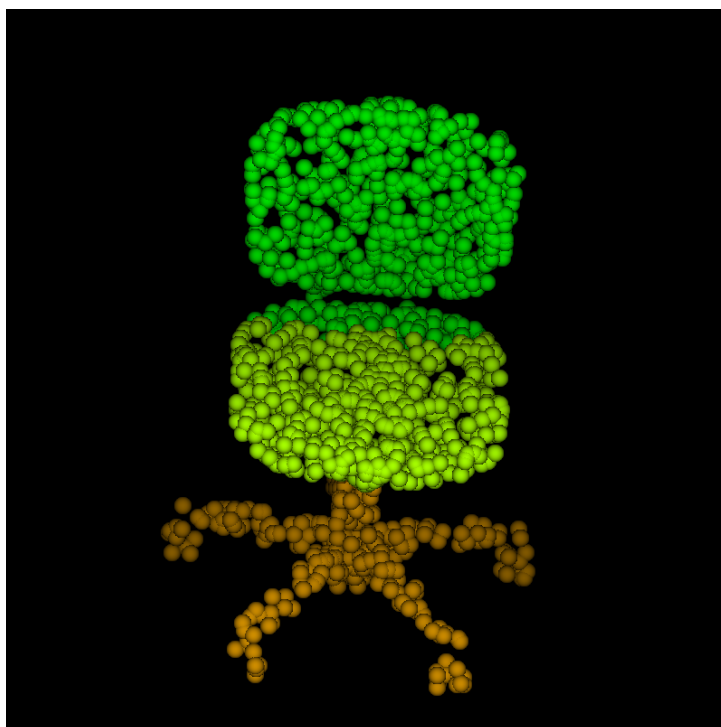
Figure 5.1: An example of a segmentation problem which has been successfully applied to a point cloud of a chair. The three different colours represent the three parts of a chair recognized by PointNet (Qi et al. 2016) (the feet, the seat, and the back of the chair). Source: https://github.com/fxia22/pointnet.pytorch (PointNet, Qi et al. 2016)

The data used in this study are the point cloud data of crania created by structured light scanning and processed with CraniAlign, or a combination of DAVID 4 and CraniAlign. Due to the issues with coarse alignment using Super 4PCS (discussed in Chapter 4 (Examining the Properties of 3D Models)), full crania needed to first be coarsely aligned with DAVID 4 before undergoing fine alignment and fusion with CraniAlign. The majority of samples from the SB and PR collection were full crania and thus underwent this process. Due to time constraints, however, not all the samples were able to be processed in this manner since manually aligning scans in DAVID 4 is a slow and tedious process. In addition, although coarse alignment was generally successful with autopsied crania, it was not robust enough to work for all such samples. The number of individuals that have an associated point cloud is therefore less than the number of individuals that were assessed visually (the results of which are in Chapter 3 (Cranial Sexual Dimorphism in Various Populations)); however, there are still enough samples to provide a solid proof of concept, which is the aim of this study. The breakdown of the point cloud dataset is given below in Table 5.1. Note that 20% of females and males from each population were designated for the evaluation dataset, with the rest used in the training dataset, for a total

of 253 individuals in the training dataset and 63 individuals in the evaluation dataset. Given that PointNet requires point clouds to be composed of exactly 2500 points, the point cloud data underwent an additional subsampling process. Using CraniAlign's fusion and subsampling process, 2500 points were randomly selected to generate the point cloud data required for PointNet. Random sampling was used instead of uniform distribution sampling because the former allows an exact number of points to be specified, whereas the latter does not.

Table 5.1: The breakdown of individuals with associated point cloud data that were suitable for use in deep learning for classification.

	SB	NU	ML	PR	Total
Females	40	28	33	33	**134**
Males	30	65	37	50	**182**
Total	**70**	**93**	**70**	**83**	**316**

For the creation of each of the three models (sex, population, and both sex and population), four parameters were established to govern the way in which the neural network was trained: number of epochs, batch size, learning rate, and momentum. Number of epochs and batch size have been defined and discussed in Chapter 1.3 (3D Methods of Analyzing Bone). The learning rate is the initial rate at which information from the training dataset is acquired, whereas the momentum is the decay rate of the learning curve.

In order to establish the values given to the four parameters, validation curves were created for each model. A validation curve is a scatterplot of model accuracy vs. the parameter in question, and plots the accuracy for both the training and evaluation results. In this way, the maximum accuracy achieved for a given parameter can be established. Additionally, the performance of the model on both the training and evaluation datasets can be compared - if the accuracy for the training dataset is much higher than that of the evaluation dataset, it is a sign that the model is overfitted and not able to be generalized.

The minimum number of epochs was established by plotting the accuracy after a set number of epochs, and determining at what number of epochs the accuracy for both the training and evaluation datasets begin to converge (i.e. there is little to no change in accuracy despite increasing numbers of epochs). When convergence occurs, it is a sign that increasing the number of epochs will not yield any improvements to the neural network; this number was therefore

set as the minimum number of epochs. The actual number of epochs used in subsequent tests was therefore set to a higher value than what was determined, in order to account for stochastic noise. The accuracies were recorded for each epoch, so it was safer to set a high epoch value, rather than risk not having enough epochs.

The batch size determines how the dataset is truncated; thus, both the training and the evaluation datasets must be a multiple of the batch size. Any samples that remain are discarded. This is not common practice in deep learning, which usually uses k-cross validation to ensure that at some point during the iterations, all samples are included in both the training and evaluation dataset (Raschka 2015). Because it was of interest in this study to make the results of the models comparable to the visual assessment results, however, it was necessary to ensure that the training and evaluation samples were never mixed. Instead, in this study the samples were randomly selected for inclusion into either dataset, but in order to ensure that roughly 20% of all collections (as opposed to 20% of the entire dataset, regardless of which collection the samples came from) were represented in the evaluation dataset, several seed values were tested for each batch size. The seed value is the random initializer for the selection and governs how and which samples are chosen. Seed values of 5 - 200 were tested in intervals of 5 to determine which seed value gave the best distribution for each batch size tested. Ideally, the number of SB samples should be roughly equal to the number of ML samples (usually a difference of 1 was permissible); the number of PR samples should be greater than that of SB and ML; and the number of NU samples should be greater than any of the other collections. There were times when several seed values gave proportions that were either the same, or very similar. In these cases, the seed in which the highest accuracy was obtained for the evaluation dataset was chosen.

The last two parameters were established by reporting the highest accuracy obtained for the evaluation dataset, as well as the corresponding accuracy obtained for the training dataset. The accuracy for the evaluation dataset was prioritized over that of the training dataset, because it is always a possibility that a high accuracy on the training dataset is a result of the model being overfitted to the data. The accuracy of the evaluation dataset is therefore more relevant and applicable. Once the first parameter (i.e. number of epochs) was established, it was kept constant for the second parameter; once the second parameter was established, the first two parameters were kept constant for the subsequent parameter, and so on. The order in

which the parameters were investigated and established was therefore strictly followed (number of epochs, batch size/seed value, learning rate, and momentum) for the creation of all three models.

Using all available samples, the accuracy achieved from all four established parameters was then reported for each model. A list of individuals and how they were classified was also generated for each model so that the individuals that were misclassified could be investigated. For the sex classification model, this list also allowed an "interobserver error" comparison between the results of the visual assessment in Chapter 3 (Cranial Sexual Dimorphism in Various Populations) and the results given by PointNet. This was done by calculating the kappa statistic which is also defined in Chapter 3. For the purpose of calculating the kappa statistic, all individuals classified as "indeterminate" for the visual analyses were eliminated from the calculation.

5.2 Results

The results are divided into three models - one that classifies according to sex and is population-agnostic; one that classifies according to population and is sex-agnostic; and one that classifies according to both sex and population. The starting default parameters used were 30 for batch size (which was chosen because this is roughly 10% of the sample size, rounded to the nearest five) with an associated default seed value of 42; 0.01 for learning rate and 0.9 for momentum, which were both defaults from PointNet.

Although the visual assessment results in Chapter 3 (Cranial Sexual Dimorphism in Various Populations) are also broken down by age, it was not possible to do the same for the machine learning analysis using deep learning since many of the younger age categories had too few individuals to accommodate both an evaluation and training dataset. Furthermore, it is much more accurate to use regression methods to estimate the age of an individual rather than to attempt to classify individuals into age categories whose ranges are technically arbitrarily determined (i.e. not determined by an algorithm or a mathematical model).

5.2.1 Sex Classification Model

To create the model for classifying by sex regardless of population, the minimum number of epochs was determined to be 50. As seen below in Figure 5.2, the accuracy for both datasets have begun to converge, with the evaluation dataset converging even earlier at 40. The number of epochs in subsequent testing for the creation of the sex classification model was set to 75 to ensure that the deep learning algorithm was allowed to run until convergence.
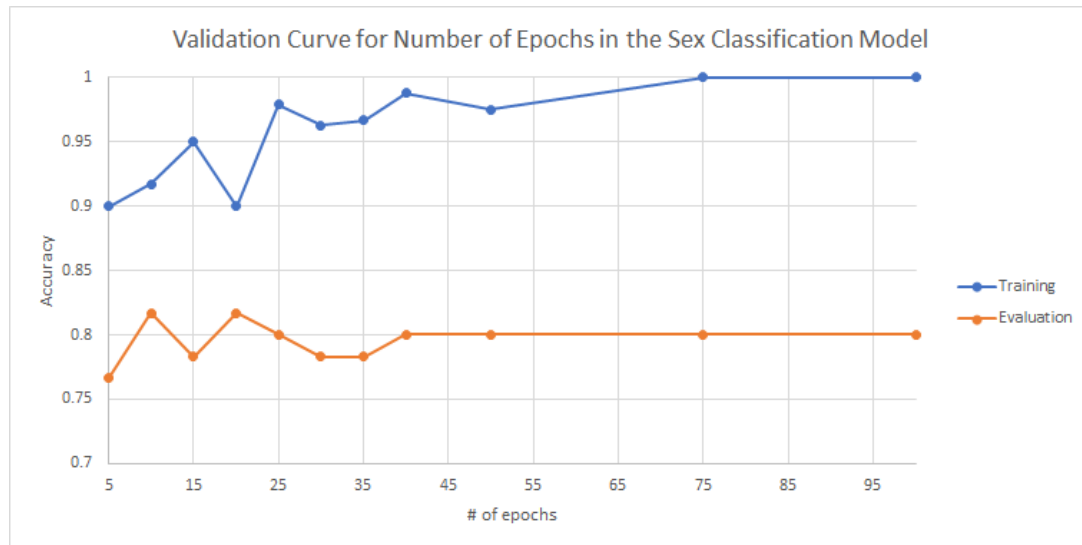
Figure 5.2: A validation curve to establish the minimum number of epochs, which is given by when both datasets show convergence. In this case, convergence occurs for both datasets when the epoch number is 50. These results are obtained with the default values for all parameters (batch size of 30; learning rate of 0.01; momentum of 0.9).

Next, the optimal batch size and associated seed was established. At the 19th epoch, a batch size of 40 with an associated seed value of 45 gave the best results, with a training accuracy of 100.0% and an evaluation accuracy of 90.0%. Due to this batch size, the evaluation and training datasets could only be a multiple of 40; therefore, the total number of individuals in the training dataset was 240, and the number of individuals in the evaluation dataset was 40. The breakdown of each dataset is given below in Tables 5.2 and 5.3. Learning rate and momentum were kept to the default values. The validation curve for batch size is given in Appendix K, Figure K.1.

Table 5.2: The breakdown of individuals in the training dataset used to create the sex classification model.

	SB	NU	ML	PR	Total
Females	27	21	29	27	**104**
Males	26	49	25	36	**136**
Total	**53**	**70**	**54**	**63**	**240**

Table 5.3: The breakdown of individuals in the evaluation dataset used to test the sex classification model.

	SB	NU	ML	PR	Total
Females	5	5	2	4	**16**
Males	3	9	5	7	**24**
Total	**8**	**14**	**7**	**11**	**40**

Establishing the learning rate gave two viable candidates - with a learning rate of 0.01, the training accuracy was 100.0% and the evaluation accuracy was 90.0%; with a learning rate of 0.0175 the evaluation accuracy was higher at 92.5% but the training accuracy was low at 86.7% (see Figure K.2 in Appendix K). Both learning rates were therefore investigated with momentum in order to determine which learning rate, and at which momentum value, the best results were obtained. It was established that a learning rate of 0.01 gave better results, and actually resulted in 3 models with different momentum values (0.25 at epoch 65; 0.3 at epoch 29; and 0.95 at epoch 43) that gave identical accuracies for both the training and evaluation datasets (100.0% and 92.5%, respectively) (see K.3 in Appendix K).

The three models were investigated and compared in terms of their performance on the evaluation dataset only. The model with a momentum of 0.25 shall be denoted as S_1; the one with a momentum of 0.3 shall be denoted as S_2; and the one with a momentum of 0.95 shall be denoted as S_3. In S_1, one male from the NU collection and one "White" male from the PR collection were misclassified; in S_2, the misclassification instead occurred with one SB male and one ML male. In both S_1 and S_2, the same SB female was misclassified. The difference between S_1 and S_2 is therefore the applicability of these models to classify males in different populations. As for S_3, the accuracy was better for males than females, since only one male was misclassified (from the SB collection, but he was a different individual than the one

misclassified in S_2). Out of the two females that were misclassified in S_3, one was the same SB individual misclassified in the other two models, and the second female that was misclassified was a "Black" individual from the PR collection.

The performance of the artificial neural networks (ANN's) were then compared to the results achieved from the visual assessments performed on the same individuals in Chapter 3 (Cranial Sexual Dimorphism in Various Populations). Tables 5.4 and 5.5 respectively summarize the results of the training and evaluation datasets compared to the two rounds of visual assessment results, and Table 5.6 reports the degree of agreement between each of the neural network models and the two rounds of visual assessment.

Table 5.4: The performance of the sex classification models on the training dataset compared to the results of the two rounds of visual assessment. Note that an indeterminate result is not applicable to the neural network models and is therefore greyed out.

	Correct	**Incorrect**	**Indeterminate**
Visual Assessment *Round 1*	131/240 *(54.6%)*	15/240 *(6.2%)*	94/240 *(39.2%)*
Visual Assessment *Round 2*	140/240 *(58.3%)*	18/240 *(7.5%)*	82/240 *(34.2%)*
ANN's (S_1, S_2, S_3) *Training Dataset*	240/240 *(100.0%)*	0/240 *(0.0%)*	

Table 5.5: The performance of the sex classification models on the evaluation dataset compared to the results of the two rounds of visual assessment. Note that an indeterminate result is not applicable to the neural network models and is therefore greyed out.

	Correct	**Incorrect**	**Indeterminate**
Visual Assessment *Round 1*	24/40 *(60.0%)*	7/40 *(17.5%)*	9/40 *(22.5%)*
Visual Assessment *Round 2*	20/40 *(50.0%)*	9/40 *(22.5%)*	11/40 *(27.5%)*
ANN's (S_1, S_2, S_3) *Evaluation Dataset*	37/40 *(92.5%)*	3/40 *(7.5%)*	

Table 5.6: A matrix of kappa values for the sex classification ANN's and the two rounds of visual assessment in order to indicate which models resulted in similar classifications.

		Visual Assessment Round 1	Visual Assessment Round 2
S_1	**Training**	0.847	0.819
	Evaluation	0.869	0.756
S_2	**Training**	0.847	0.819
	Evaluation	0.771	0.756
S_3	**Training**	0.847	0.819
	Evaluation	0.738	0.651

5.2.2 Population Classification Model

To create the model for classifying by population regardless of sex, the minimum number of epochs was determined to be 35. As seen below in Figure 5.3, the accuracy for both datasets have begun to converge, with the training dataset converging even earlier at 20. The number of epochs in subsequent testing for the creation of the population classification model was set to 60 to ensure that the deep learning algorithm was allowed to run until convergence.
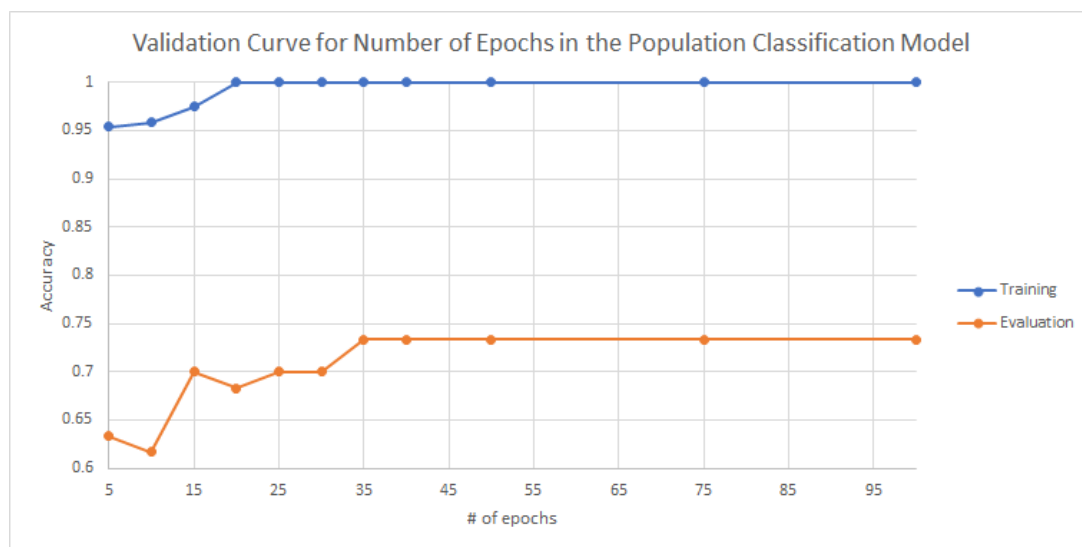
Figure 5.3: A validation curve to establish the minimum number of epochs, which is given by when both datasets show convergence. In this case, convergence occurs for both datasets when the epoch number is 35. These results are obtained with the default values for all parameters (batch size of 30; learning rate of 0.01; momentum of 0.9).

The optimal batch size was determined to be 40, with a seed of 135. The validation curve for batch size is given in Appendix L, Figure L.1. Similar to the sex classification model, the

training dataset therefore consisted of 240 individuals while the evaluation dataset contained 40; however, due to the different seed, the composition of both datasets are slightly different than that of the sex classification model. The composition of the datasets are given below in Tables 5.7 and 5.8

Table 5.7: The breakdown of individuals in the training dataset used to create the population classification model.

	SB	NU	ML	PR	Total
Females	32	23	26	27	**108**
Males	20	46	26	40	**132**
Total	**52**	**69**	**52**	**67**	**240**

Table 5.8: The breakdown of individuals in the evaluation dataset used to test the population classification model.

	SB	NU	ML	PR	Total
Females	5	3	3	3	**14**
Males	4	9	6	7	**26**
Total	**9**	**12**	**9**	**10**	**40**

When testing values for learning rate, it was discovered that many different values gave the same result - a training accuracy of 100.0% and an evaluation accuracy of 97.5% (39/40). Investigating momentum values gave even more instances in which the same accuracies were achieved. For the purpose of this chapter, which is to provide a proof of concept, three models were randomly selected as examples and presented here - P_1 represents a model with a learning rate of 0.01 and a momentum of 0.75, with 15 epochs; P_2 represents a model with a learning rate of 0.01 and a momentum of 0.4, with 10 epochs; and P_3 represents a model with a learning rate of 0.005 and a momentum of 0.9, with 17 epochs. All three models used a batch size of 40 and a seed of 135. For the sake of completion, the validation curves for learning rate and momentum are still provided in Appendix L. In all three models, the only error was the same ML female who was misclassified as belonging to the PR collection.

5.2.3 Sex & Population Classification Model

To create the model for classifying according to both sex and population, the minimum number of epochs was determined to be 65. As seen below in Figure 5.4), the accuracy for both datasets have begun to converge, with the training dataset converging even earlier at 50. The number of epochs in subsequent testing for the creation of the population classification model was set to 100 to ensure that the deep learning algorithm was allowed to run until convergence.
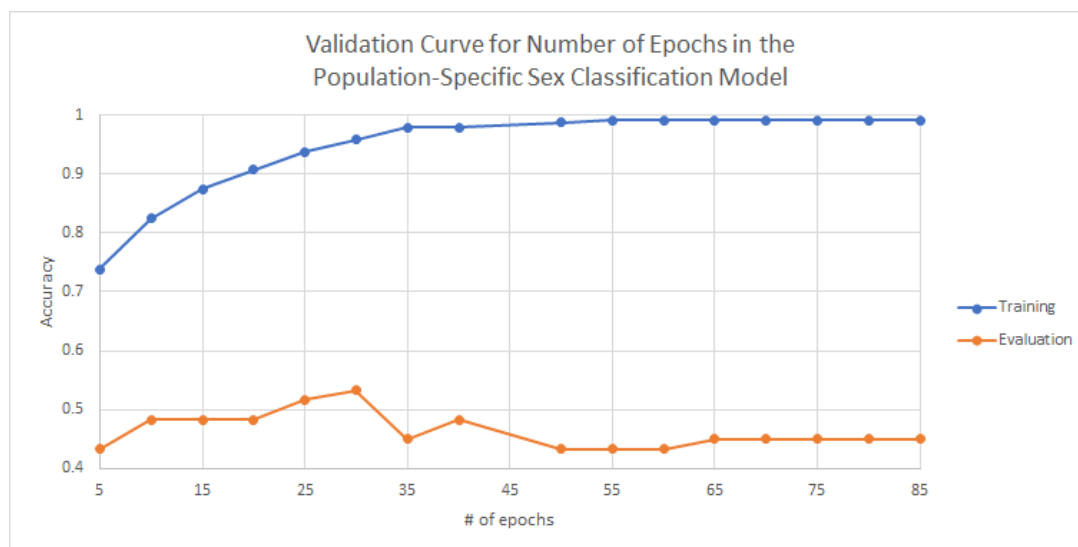
Figure 5.4: A validation curve to establish the minimum number of epochs, which is given by when both datasets show convergence. In this case, convergence occurs for both datasets when the epoch number is 65. These results are obtained with the default values for all parameters (batch size of 30; learning rate of 0.01; momentum of 0.9).

The optimal batch size was determined to be 40, with a seed of 45. The validation curve for batch size is given in Appendix M, Figure M.1. Due to the fact that the same batch size and seed were used as those in the sex classification model, the composition of both the evaluation and training datasets are the same as those given above Tables 5.2 and 5.3.

The optimal learning rate was determined to be 0.0175 with a momentum of 0.9, which gave a training accuracy of 97.1% (235/240) and an evaluation accuracy of 87.5% (35/40) at the 44th epoch. The validation curves for learning rate and momentum are provided in Appendix M. The misclassification in the training dataset was purely due to sex, meaning that all individuals were correctly categorized into their respective collections. All incorrectly classified individuals were female - four from the ML collection and one "Black" from the PR collection. Similarly, all individuals in the evaluation dataset were properly classified by collection, but the

five misclassifications were due to sex - one NU female, one SB female, one PR "White" male, one PR "Black" male, and one SB male.

5.3   Discussion & Conclusion

Three sex classification models were produced, all with the same accuracy of 100%. When testing these models on the holdout samples, the models were able to correctly classify 92.5% (37/40) of the individuals, although each model misclassified different individuals. Both S_1 and S_2 performed slightly worse than S_3 at classifying males, although the individuals mis-classified in S_2 were all Caucasian/European males whereas in S_1 the misclassified individuals were non-European. Further testing is needed to establish whether S_1 is truly better suited for classifying European males than S_2. Conversely, S_3 performed slightly better than the other two models at classifying males, since only one SB male was misclassified, though there does not seem to be a population-specific bias in S_3. All three models therefore show great promise as universal sex assessment methods, due to the high model accuracy and test accuracy.

In all three sex classification models - as well as in both rounds of visual assessments described in Chapter 3 - the same SB female was misclassified as a male. Upon investigating the identity of this individual, it was discovered that the context number assigned to her is one number off from a male family member, whose cranium was too fragmented to assess for this research project. It therefore begs the following question: did commingling occur, or was there a mix-up during the recovery and documentation of the remains such that the cranium thought to belong to the female actually belongs to her male relative? Upon reviewing photographs of the mandible of the male, who died 41 years prior to the woman, it is evident that the soil staining of the mandible is much darker than that of the cranium, which is consistent with the cranium being interred a significant amount of time after the mandible. The mandible and the cranium therefore do not belong to the same individual due to the different degree of soil staining between the two. It is possible that the cranium belongs to the woman, who was 61 years old at death and who may have begun to exhibit more male-like features due to her advanced age.

Interestingly, although the accuracy obtained from the visual assessment results are much lower than the accuracy obtained by the three sex classification models when tested on holdout samples (92.5%), the kappa statistics showed that for those individuals who were clas-

sified as either male or female in the visual assessments, there was a high degree of agreement with the sex classification models (ranging from 0.651 - 0.869). It must be remembered that the calculation of the kappa statistics only took into account those who were actually classified as either male or female. This means that when the combination of morphological traits were discriminate enough to classify an individual, they were sufficient to categorize an individual correctly. The results therefore confirm that the traits identified by Buikstra and Ubelaker (Buikstra and Ubelaker 1994) and Williams and Rogers (Williams and Rogers 2006) are appropriate traits to use in skeletal sex assessment, but the high number of indeterminate individuals suggests that these traits are not enough on their own to account for the overlap in sexual dimorphism that exists. Conversely, the fact that the three neural networks created in this study were able to achieve such high accuracies suggests that there are other combinations of geometric characteristics that can decrease the number of individuals that fall into the indeterminate category and increase classification rates. Further research is required to determine exactly what geometric characteristics are useful in increasing classification rates, but if successfully identified and incorporated into existing sex assessment methods, the usefulness of these methods (e.g. as quantified by the discrimination factor d in Chapter 3) would increase. A suggestion for how future research projects could address this problem is discussed in Chapter 6 (Directions for Future Research).

Several population classification models were produced, all with an accuracy of 100%. Testing these models on the holdout samples gave an accuracy of 97.5% (39/40). Examining three of these models, the same ML female was incorrectly classified as belonging to the PR collection, so there is no tangible way to compare the practical difference between all three models. The fact that several models were produced with the same training and evaluation accuracies, regardless of learning rate and momentum values, highly suggests that the task of classifying according to population is a fairly "easy" task for deep learning. In the same mindset, traditional skeletal ancestry assessment methods (with broad categories such as "Caucasian", "Asian", and "African") are starting to be abandoned in favour of identifying population-specific traits. The results of the population classification models produced in this study support this endeavour because it provides a solid proof of concept for population classification even between groups belonging to the same broad category (ML, SB, and even "White" PR individuals for example, would have been classified together as "Caucasian" instead of being distinguished). Further investigation is required to determine whether the models produced from this study are

robust enough to distinguish individuals coming from very mixed populations, such as in North America, or whether such individuals would be classified according to their genetic/ancestral population. Such an investigation would yield information as to whether secular changes or genetic background more prominently contribute to traits typically used for ancestry/population classification.

The population-specific sex classification model produced in this study has an accuracy of 97.1% (235/240), and when tested on the holdout sample, had an accuracy of 87.5% (35/40). All individuals regardless of whether they were in the training dataset or the evaluation dataset were correctly classified according to population, but the errors occurred due to a misclassification according to sex. In the training dataset, all of the misclassified individuals were female, whereas a mix of both male and female individuals were misclassified in the evaluation dataset. Four out of the five incorrectly classified individuals in the training dataset were from the ML collection. Despite the fact that the model produced does not account as well for ML females as it does for individuals in other collections, there did not seem to be a population bias when tested on the evaluation dataset because the misclassified individuals in the holdout sample were quite evenly distributed across collections. Further testing on external samples is required in order to establish whether a population bias truly does or does not exist.

In conclusion, three types of models were created using deep learning on cranial point clouds consisting of 2500 points - a global sex classification model, a population classification model, and a population-specific sex classification model. According to the criteria set for the purposes of this study, all three models were successful in that both the training and evaluation accuracies were well over 80.0%. Although normally a model with a training accuracy of 100.0% is a result met with trepidation due to the implications of overfitting, the results in this study have shown that the models were still applicable to a holdout sample that was not involved in the creation of the model at all. The difference in accuracies between the training and evaluation datasets were 2.5% (in the population classification models), 7.5% (in the sex classification models), and 9.4% (in the population-specific sex classification model). This difference in accuracies can be minimized further by using k-cross validation, which is a standard practice in deep learning (Raschka 2015), in order to ensure that throughout the many different iterations/epochs, all individuals at some point are used in the training and evaluation dataset. The use of k-cross validation in the training of a neural network therefore increases the chance

that the resulting model will be able to be generalized to other data (Raschka 2015). Furthermore, the models can be improved by optimizing all parameters (batch size, sampling seed, learning rate, and momentum) in a multi-dimensional way rather than testing them one by one as was done in this study. Testing different seed values for the model itself (not the sampling seed) to influence the initialization of the nodes may also result in the creation of a model with higher accuracies.

In traditional methods of assessing bone, there is a reported accuracy for the method given by the researcher(s) who developed the method. This reported accuracy is cited and used as a baseline when testing the method, but in reality the reported accuracy is probably the best-case scenario. Even in cases where a researcher both developed and tested their method on a holdout sample, their reported accuracies for the training and evaluation datasets tend to be higher than what can be achieved by other researchers due to interobserver error (a topic that is discussed in Lam et al. 2016). The training accuracy of a neural network is therefore analogous to the reported accuracy associated with a given method, and the evaluation accuracy is representative of the method's actual performance, similar to when other researchers test a given method on a different dataset. The two major advantages with neural networks is that 1) the method's performance is computed simultaneously with the creation of the model, so there is no need to wait for other researchers to test the model in order to have an idea of the model's applicability to other datasets; and 2) the performance of the neural network will not vary according to the researcher.

One major limitation of the use of neural networks is that the performance of the neural network only indicates the probability that a sample belongs to a given category. This is in contrast to the posterior probability (discussed in Chapter 1 (Introduction & Background)) which indicates the probability that a given sample was correctly categorized. At the moment, there is no way to calculate the posterior probability for a neural network, meaning that the suitability of the neural network to classify a given sample cannot be evaluated. This is an open research problem in the field of computer science that currently does not have a well-established solution. Consequently, the use of neural networks in forensic applications is limited, since forensic analyses require the ability to interpret whether a given sample is appropriately analyzed by a given method. Despite this limitation, what would strengthen the results of using a neural network is to modify the output such that the probability density for a given, unknown sample is

reported. In this manner, the probability that an unknown individual belongs to each available category can be provided along with the performance of the neural network so that the resulting categorization is more transparent and can be further interpreted by an analyst. Therefore, the fact that neural networks can provide a reliable quantification as to its performance is a major step in being recognized in court, since it satisfies one of the Daubert criteria (1993).

No researcher has attempted to use deep learning on point cloud data of crania for sex classification, so the only comparable studies to this one are those that fall under the category of landmark-based GMM studies identified in Chapter 1 (Introduction & Background). As previously discussed, these studies use craniometric landmarks and measurements for their analyses which create a more abstract geometric shape for analysis. In this study, 2500 points were used in the analysis - two orders of magnitude greater than any landmark-based study. The advantage to using such a high number of points is that the true geometric shape of crania is better represented and the shape information is better preserved. There is also the limitations of the mathematical tools used by other studies to consider. Some methods, such as FORDISC (Jantz and Ousley 2005) use made-to-order discriminant functions, which are not complex enough to create a robust mathematical model encompassing the variation between samples belonging to the same category. It is also common practice in osteology/archaeology to use Principal Components Analysis (PCA) for data clustering such that classifications are made based on a given sample's distance from the center of the cluster (e.g. Luo et al. 2013; Chovalopoulou and Bertsatos 2018). The intended purpose of PCA, however, is not as a classification method. Instead, PCA is a method that attempts to explain the variance in a given dataset, by identifying how certain factors explain the variance in the dataset (Jolliffe 2002). For example, in two-dimensional PCA analyses, factor 1 and factor 2 are two abstract factors and it is up to the researcher to interpret what those factors are given their knowledge of the data. PCA is therefore - and should be used as - an investigative tool that can help direct further analyses. PCA is appropriate for establishing the distance between different clusters, but this distance is not a direct way to determine whether something belongs to a given category. Conversely, deep learning is a method that specifically addresses the problem of classification, and has been well-established as a suitable tool for such a task (Goodfellow et al. 2016). Applying deep learning to classification problems, such as sex and population classification, is therefore a much more appropriate method of analysis than those commonly used in biological anthropology studies.

Although population classification was not the focus of this PhD project, the creation of three population classification models yielded results previously unseen in the literature - a model with a theoretical accuracy of 100.0% and 97.5% accuracy in practice. Even Ances-Trees (Navega et al. 2015), a decision trees machine learning method tested on European and African samples, did not perform as well. When testing AncesTrees on a dataset that consisted of individuals from six ancestral groups, only 75.0% of African individuals and 79.2% of European individuals were correctly categorized; when a model was created that only included these two groups, the performance increased to 93.8%. The results in this PhD project are therefore much better than those of AncesTrees, since an accuracy of 97.5% was achieved using double the number of population groups for population classification only. Considering the population-specific sex classification, the achieved accuracy in this project was 87.5%, although 100% of individuals were correctly classified by population. To put these results into context, the AncesTrees model using only European and African individuals had a 50% chance of correctly classifying an individual whereas in this project, there was a 25% chance of correct classification for the population model, and 12.5% chance of correct classification for the population-specific sex classification model. The ability to achieve a test accuracy comparable and even higher than what was achieved by AncesTrees is therefore not a small feat. There are two possible reasons for the huge discrepancy between the accuracies from AncesTrees and the models from this study - the first is the simple fact that deep learning is much better suited for classification tasks than decision trees (Goodfellow et al. 2016), and the second is that the models produced in this project used 2500 points that represent the entire geometry of the cranium whereas AncesTrees used 23 craniometric measurements that do not capture the entire geometric shape.

The results of this study are applicable to forensic anthropology because they present three new models with unprecedented accuracies and performance which can aid in skeletal identification. These models will need to be investigated more in order to test the models' true performance on external samples, but the fact that an evaluation accuracy is provided means that a known error rate has been established, meeting one of the criteria of the Daubert standards (Daubert v. Merrell Dow Pharmaceuticals 1993). Testing and developing these models on larger datasets from even more populations would also improve the ability of the models to be generalized while maintaining high performance. Finally, providing an output of what geometric information the model is actually using will help analysts improve current existing

techniques of both morphological and metric assessments of the cranium.

Chapter 6

Directions for Future Research

The purpose of this PhD project was to create a new method of sex assessment for crania using 3D point cloud data and machine learning, and to compare this new method to existing morphological assessments. The output of this project has actually created three methods of assessment instead of the original one - a global method of sex assessment, a method for assessing ancestry/population (which was a byproduct of this project and not the focus), and a population-specific sex assessment method. In fact, three models for sex assessment, three models for ancestry/population, and one model for sex assessment have been produced, all with accuracies and performances that exceeded the initial expectations. This PhD project has therefore successfully provided a proof of concept as to how deep learning can be a valuable tool for addressing fundamental research questions, such as those pertaining to the creation of a biological profile in bioarchaeology and in forensic anthropology. Its applicability to forensic anthropology is further established by meeting several of the requirements regarding the fourth Daubert criteria (Daubert v. Merrell Dow Pharmaceuticals (1993)) pertaining to the scientific methodology - the models created from deep learning are testable and have been tested in this project; there is a known and potential error rate, given by the accuracy of the training dataset (i.e. the model itself) and the accuracy of the evaluation dataset (i.e. the performance), respectively; and the testing was subject to proper standards and controls through the use of holdout samples that were not involved in the creation of the models. Therefore, what remains in order to completely satisfy the requirements of the fourth Daubert criteria are: subjecting these three models to peer review and publication, such that the true performance of the model can be

tested further; and to have the models, as well as the application of deep learning for sex and population classification, be accepted by the scientific community.

Two methods exist in order to make the models accessible for peer review or even for use by an analyst. The first method is to simply export the state of the trained neural networks into a repository. The end user would then access the repository and run it with their data, which should be transformed in the same manner as what was done in this project, i.e. with 2500 points randomly subsampled. The output would then be a classification of their data. The major limitation to exporting the state of the models, however, is that the file size is large (i.e. between 3 - 6 GB), and therefore cannot be easily accommodated by free online repositories due to upload and download speeds as well as the size limitations imposed by the online repository itself (e.g. Github imposes a 100 MB/file limit and BitBucket's Large File Storage (LFS) free account has a 1 GB overall storage limit). The second method is therefore to only export the training data and the training parameters such that the end user can generate the state of the model themselves. However, the end user would need to do so on a hard drive that has sufficient space to store the model locally. The user would also have access to all of the training data which would have ethical implications since the data would include the 3D representation of skeletal remains and the associated personal information such as age, sex, and population. It can be argued that the first method of exporting the state of the neural network poses a similar ethical problem because all of this sensitive information is provided in an aggregated format, though less readily accessible/readable to a human browsing the file. Without a carefully laid-out plan for commercializing and/or disseminating the output from this research project - which at the very least must include a plan for funding the data storage, ensuring the security of the data, and addressing ethical issues surrounding the dissemination of the data - it will not be possible to subject the neural networks to peer review.

The use of a computer program to assess skeletal samples is an invaluable tool, but a program should not replace a human analyst. Therefore, an important aspect of this research project was to create a computer program that can offer an improved performance pertaining to current osteological assessments, as well as to provide a tool for improving human analysts themselves in performing osteological assessments. The latter is especially important when a quick but accurate assessment needs to be undertaken in circumstances that preclude the ability to scan a sample and use a program to compare it - for example, in a forensic context

when human remains are found in the field, police will need an immediate answer as to whether or not the remains are consistent with the suspected missing person, or if it is another individual that could possibly be involved. Sex, ancestry, and age therefore need to be quickly but accurately assessed by the consulting forensic anthropologist/archaeologist. It is therefore vital to continue to use computers as tools to improve current methods and not as replacements for them, and this mindset was maintained throughout this research project. Unfortunately, deep learning at the moment does not return an output that allows an analyst to understand what makes a model successful - namely, the output is limited to the solution and the state of the model, which does not allow scrutiny into what geometric features/combination of geometric features were found to be useful for the classification task. This inability to provide an output that is open to further scrutiny is a major obstacle for deep learning to be accepted and widely used in forensic anthropology - the output, similar to what is provided by 3D-ID (Ross et al. 2010) and AncesTrees (Navega et al. 2015), is merely a result of whether an individual has been classified as male or female, and/or into which population. There is no ability to interpret the results or to gauge the possibility whether that individual falls within the small percentage of incorrectly classified individuals. A wrong classification affects the biological profile created for an individual, and has drastic implications in both bioarchaeology and forensic anthropology. Computer programs should therefore not be used on their own, and their results should always be interpreted jointly with the results of an analyst's own assessments.

In order for the creation of the three methods of sex and population classification to be possible to begin with, it was necessary to establish the 3D ground-truth database of cranial point clouds. It is the hope of this PhD researcher that with proper permissions and ethical considerations, such data will become more accessible to researchers in the form of a proper database. A proper database is not simply a collection of data with associated information, but is actually a formalized structure by which data can be stored, accessed, and modified. The creation and maintenance of a proper database belongs to the realm of computer science and is beyond the scope of this project; as a result, the concept of a database will not be discussed further (*N.B.* for a discussion on the proper implementation of a research-oriented database containing sensitive material, refer to the thesis by Pillin (2019)). There is difficulty in creating a collaborative database of human skeletal material due to ethical concerns - namely, how to prevent misuse of such material and ensure restricted access. With computer programs, however, access to such a database can be restricted such that the information can be accessed by a

program but cannot be visualized by the researcher. For example, in this project a program was created in Python that automatically browsed a file folder, found all of the .obj files, browsed and stored the associated sex and population information from a .csv file, and was able to return the results of PointNet's performance in a Linux terminal. None of these steps required the files to be opened or any of the cranial data to be seen by the researcher. In this way, the misuse of the visual representation of human remains can be eliminated, while allowing researchers globally to access a 3D database with immense research potential. Programs, such as the one that was created as part of this project, can be written by independent analysts to verify existing analyses or to create new ones. By sharing virtual ground-truth data with researchers and building one large database while respecting ethical boundaries, research potential will increase drastically and new methods with more robust statistical conclusions can be created.

With the precedent set by studies such as this one where high accuracies (i.e. above 90%) are obtained for classification methods, it is prudent to reconsider the 80% threshold that has traditionally been utilized as an indicator of an acceptable method in forensic anthropology. It is clear that the utility of machine learning can far surpass this threshold, so it follows that the performance of such methods utilized for legal purposes also be held to higher standards. With the development of machine learning research, and with studies that test the applicability of its numerous algorithms, the expectations for methods applied to forensic anthropology should be made more rigorous. One important direction which this PhD project highlights is the issue of deep learning algorithms returning results that currently cannot be linked to tangible factors. To do so, it would first be necessary to create a program that can automatically recognize craniometric points on the cranium (which would be best accomplished by only using those points which can be defined metrically - e.g. euryon, which is defined as the points on the cranium that give the widest cranial breadth, could be metrically determined by defining the maximum distance of a point cloud in a given plane). A given radius surrounding certain craniometric points could define the boundaries of the geometric shape corresponding to the morphological traits used in this project (e.g. defining a set radius around opisthocranion, which is the most posterior point of the cranium that is not on the external occipital protuberance, would be equivalent to assessing the shape of the occipital protuberance itself, given that the radius includes this feature).

Once the automatic detection of craniometric landmarks and metrically-defined morpho-

logical features are established, two approaches could be taken. First, a supervised machine learning approach, in which the data are labelled according to sex and population, would allow the determination of which craniometric points and/or morphological features are most informative regarding correct classification because the traits that resulted in correct classification could be ranked according to performance. This approach would provide insight into what characteristics are actually useful for sex and ancestry classification, allowing a better understanding of which traits are globally useful for sex assessment and which ones are population specific. The results would also provide insight into whether the successful performances in the models produced in this PhD project were due to the fact that deep learning was used, or because 2500 points that represent the geometry of the cranium were used, rather than the standard craniometric points which number far less. Secondly, an unsupervised machine learning approach could be undertaken to create grouped clusters from the data. It would then be possible to understand the combinations of craniometric points and morphological traits, and how these combinations are actually useful in sex and ancestry classification. The clustering based on these craniometric points and traits could be compared to the correct classification rates, which would allow insight into how the cranial traits are applicable and useful for identifying sex and ancestry in different populations. For example, if two clusters are identified by the program which correspond to a male and female group, it can be inferred that the combination of traits used for this clustering are not population-specific. The results of such a model could be compared to the global sex classification models produced in this project to determine if it is possible that the same traits were weighted more heavily in this project. Conversely, if multiple clusters are created which correspond to specific sex and ancestry groups (e.g. Italian males vs. Japanese females), the traits used can be inferred to be useful and affected by both sex and population. The results of the unsupervised machine learning could then be compared to the population-specific sex classification results achieved in this project to provide insight, once again, into which traits were weighted most heavily in the neural network models. In conclusion, the supervised machine learning approach could determine which traits are useful and to what extent, whereas the unsupervised approach would investigate how these traits are useful, and in what combination. Combining the results of unsupervised and supervised machine learning would therefore be a powerful analytical approach to understanding and interpreting sexual dimorphism in different populations, and would be the next step in building upon - and understanding the neural networks created from - the research in this PhD project.

Another direct extension of this PhD project is to test the performance of the population-agnostic sex assessment models produced in this project on samples that do not belong to any of the four populations represented in the dataset. The performance of classifying sex on individuals that are unlike those used to train the neural networks would indicate how robust the population-agnostic sex assessment models are to external samples, and would also provide insight into whether the "global" sex assessment models truly can be considered "global". Testing the models on North American individuals, especially those who are descendants from the geographical populations represented in this project, would provide results that could be used to address the question of whether or not the environment has impacted cranial morphology in a manner that distinguishes North American individuals from their ancestrally genetic counterparts. Finally, the acquisition of samples from populations that are not represented in this project could be incorporated into the training of the neural network to increase the model's ability to be generalized to other populations, provided that the samples have associated ground-truth information. Consequently, the models could be further improved both in terms of performance and generalization.

This PhD project used PointNet's classification algorithm, which has now been successfully applied to cranial point cloud data, and sets the groundwork for future research to investigate the applicability of PointNet's other algorithm which focuses on segmentation. The segmentation task in deep learning creates a neural network in which geometric shapes are recognized and classified even if the input is incomplete. This has tremendous potential in both bioarchaeology and in forensic anthropology, because bone fragments are often found in both contexts due to trauma the individual could have sustained in life, post-mortem damage, or excavation damage. If a neural network was created to recognize all the different bones in the human skeleton, bone fragments that are otherwise visually unidentifiable could potentially be identified through such a geometric analysis. Furthermore, if a hypothetical neural network were trained on different human bones from different age, sex, and population categories with ground-truth information, the ability to identify a bone fragment could also return the most likely age, sex, and/or population category to which the fragment belongs. The ability of deep learning to be able to classify bones that are visually indeterminate has been established in this PhD project, so it is quite possible that it can perform well when applied to identifying biological characteristics from bone fragments. The success of such a hypothetical neural network, however, would be highly dependent on creating a large enough dataset with associated

ground-truth information, which would be difficult especially for bone fragments. Nevertheless, the segmentation algorithm provided by PointNet should be explored.

Expanding on the premise of applying machine learning to fragmentary remains, it should be noted that both autopsied and non-autopsied crania were present in all of the different skeletal collections, and there were individuals whose crania had damaged or missing sections. The high performance of the neural networks on both the evaluation and training dataset allows for an important conclusion to be made: the information that is required for sex and population classification was acceptably represented by 2500 points, and the information in the missing parts was not essential. This conclusion is very promising for the application of the neural networks generated in this research project to fragmentary remains. It would therefore be interesting to test the neural networks' ability to classify even more fragmentary remains, perhaps both with and without PointNet's segmentation algorithm.

In conclusion, this PhD project has established a quantitative method of comparing point cloud data, and has also been the first to apply deep learning algorithms to point cloud data representing the entire geometry of crania in order to classify individuals into sex, population, and population-specific sex categories. As a result, this PhD project adds to the wide array of tools and methods that already exist for establishing the biological profile of a skeletonized individual. From discriminant function analysis to PCA to decision trees; from craniometric landmarks to point cloud data, it is clear that there is not one catch-all method that is the best. The wide range of analyses that researchers use in physical anthropology is a testament to the creative ways in which different individuals decide to use different tools, which allow the knowledge of the field to grow and expand. It is the hope of this PhD researcher that the current players in the field of bioarchaeology, biological anthropology, and forensic anthropology continue to keep an open mind to new applications and new methods such that the field can continue to improve while staying rooted in sound scientific methods.

Chapter 7

Bibliography

Adams, B. J. and Byrd, J. E. (2002). Interobserver variation of selected postcranial skeletal measurements. *Journal of Forensic Sciences*, 47:1193–1202.

Adams, D. C., Rohlf, F. J., and Slice, D. E. (2004). Geometric morphometrics: ten years of progress following the "revolution". *Italian Journal of Zoology*, 71:5–16.

Afrianty, I., Nasien, D., Kadir, M. R., and Haron, H. (2014). Backpropagation neural network for sex determination from patella in forensic anthropology. *Advances in Computer Science and its Applications*, 279:723–728.

Algee-Hewitt, B. F. B. (2016). Population inference from contemporary American craniometrics. *American Journal of Physical Anthropology*. 10.1002/ajpa.22959.

Ascadi, G. Y. and Nemeskeri, J. (1970). *History of human span and mortality*. Akademiai Kiado, Budapest.

ASTM (1997). *E 1441-97. Standard Guide for Computed Tomography (CT) Imaging*.

Barrier, P., Dedouit, F., Braga, J., Joffre, F., Rougé, D., Rousseau, H., and Telmon, N. (2009). Age at death estimation using multislice computed tomography reconstructions of the posterior pelvis. *Journal of Forensic Sciences*, 54:773–778.

Bass, W. M. (2005). *Human osteology: A labaratory and field manual*. Missouri Archaeological Society, Columbia, 5th edition.

Bass, W. M. and Driscoll, P. A. (1983). Summary of skeletal identification in tennessee: 1971-1981. *Journal of Forensic Sciences*, 28:159–168.

Berrizbeitia, E. L. (1989). Sex determination with the head of the radius. *Journal of Forensic Sciences*, 34:1206–1213.

Đuric, M. (2005). The reliability of sex determination of skeletons from forensic context in the balkans. *Forensic Science International*, 147:159–164.

Bidmos, M. A., Gibbon, V. E., and Štrkalj, G. (2010). Recent advances in sex identification of human skeletal remains in South Africa. *South African Journal of Science*, 106.

Bogin, B. (1999). *Patterns of Human Growth*. Cambridge University Press, Cambridge.

Bouaziz, S., Tagliasacchi, A., and Pauly, M. (2013). Sparse iterative closest point. *Computer Graphics Forum (Symposium on Geometry Processing)*, 32(5):1–11.

Braz, V. S. (2009). Anthropological estimation of sex. In Blau, S. and Ubelaker, D. H., editors, *Handbook of Forensic Anthropology and Archaeology*, pages 201–207. Left Coast Press, Inc., Walnut Creek, California.

Buikstra, J. E. and Ubelaker, D. H. (1994). *Standards for Data Collection from Human Skeletal Remains*. Arkansas Archaeological Survey Research Series No. 44, Fayetteville.

Bulut, O., Petaros, A., Hizliol, I., Wärmländer, S. K., and Hekimoglu, B. (2016). Sexual dimorphism in frontal bone roundness quantified by a novel 3D-based and landmark-free method. *Forensic Science International*. http://dx.doi.org/doi:10.1016/j.forsciint.2016.01.028.

Bulygina, E., Mitteroecker, P., and Aiello, L. (2006). Ontogeny of facial dimorphism and patterns of individual development within one human population. *American Journal of Physical Anthropology*, 131:432–443.

Buschang, P. H., Baume, R. M., and Nass, G. (1983). A craniofacial growth maturity gradient for males and females between 4 and 16 years of age. *American Journal of Physical Anthropology*, 61:373–382.

Byers, S. N. (2002). *Introduction to forensic anthropology*. Allyn and Bacon Publishers, Boston.

Calce, S. E. (2012). A new method to estimate adult age-at-death using the acetabulum. *American Journal of Physical Anthropology*, 148:11–23.

Calce, S. E. and Rogers, T. L. (2011). Evaluation of age estimation technique: Testing traits of the acetabulum. *Journal of Forensic Sciences*, 56:302–311.

Cappella, A., Cummaudo, M., Arrigoni, E., Collini, F., and Cattaneo, C. (2016). The issue of age estimation in a modern skeletal population: are even the more modern curent aging methods satisfactory for the elderly? *Journal of Forensic Sciences*, 62:12–17.

Cartmill, M. (1999). The status of the race concept in physical anthropology. *American Anthropologist*, 100:651–660.

Cavalli, F., Lusnig, L., and Trentin, E. (2017). Use of pattern recognition and neural networks for non-metric sex diagnosis from lateral shape of calvarium: an innovative model for computer-aided diagnosis in forensic and physical anthropology. *International Journal of Legal Medicine*, 131:823–833.

Challis, J., Robinson, J., Ruark, D. W., and Thorburn, G. (1976). The development of endocrine function in the human fetus. In Roberts, D. and Thomson, A. M., editors, *The biology of human fetal growth*, pages 149–194. Taylor and Francis, London.

Chovalopoulou, M.-E. and Bertsatos, A. (2018). Exploring the shape variation of the human cranium. a geometric morphometrics study on a modern greek population sample. In Rissech, C., Lloveras, L., Nadal, J., and Fullola, J., editors, *Geometric Morphometrics. Trends in Biology, Paleobiology and Archaeology*, pages 25–39. Universitat de Barcelona.

Christensen, A. M. (2004). The impact of Daubert: implications for testimony and research in forensic anthropology (and the use of frontal sinuses in personal identification). *Journal of Forensic Sciences*, 49:1–4.

Christensen, A. M., Passalacqua, N. V., and Bartelink, E. J. (2014). *Forensic Anthropology Current Methods and Practice*. Elsevier Inc., United States.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Coquerelle, M., Bayle, P., Bookstein, F. L., Braga, J., Halazonetis, D. J., Katina, S., and Weber, G. W. (2010). The association between dental mineralization and mandibular form: a study combining additive conjoint measurement and geometric morphometrics. *Journal of Anthropological Sciences*, 88:129–150.

Coquerelle, M., Bookstein, F. L., Braga, J., Halazonetis, D. J., Weber, G. W., and Mitteroecker, P. (2011). Sexual dimorphism of the human mandible and its association with dental development. *American Journal of Physical Anthropology*, 145:192–202.

Curate, F., Umbelino, C., Perinha, A., Nogueira, C., Silva, A., and Cunha, E. (2017). Sex determination from the femur in portuguese populations with classical and machine-learning classifiers. *Journal of Forensic and Legal Medicine*, 52:75–81.

Daubert v. Merrell Dow Pharmaceuticals (1993). Inc., 509 u.s. 579, 589.

DAVID-4 (2007-2017). *Documentation for DAVID 4*. DAVID Group.

De Villiers, H. (1968). *The skull of the South African Negro: A biometrical and morphological study*. Wits University Press, Johannesburg.

Decker, S. J., Davy-Jow, S. L., Ford, J. M., and Hilbelink, D. R. (2011). Virtual determination of sex: Metric and nonmetric traits of the adult pelvis from 3D computed tomography models. *Journal of Forensic Sciences*, 56:1107–1114.

DiGangi, E. A. and Hefner, J. T. (2013). Ancestry estimation. In DiGangi, E. A. and Moore, M. K., editors, *Research Methods in Human Skeletal Biology*, pages 117–149. Elsevier Inc.

Edgar, H. J. H. and Hunley, K. L. (2009). Race reconciled? How biological anthropologists view human variation. *American Journal of Physical Anthropology*, 139:1–4.

Enlow, D. H. (1982). *Handbook of facial growth*. W. B. Saunders Company, Toronto, 2 edition.

Ferrant, O., Rouge-Maillart, C., Guittet, L., Papin, F., Clin, B., Fau, G., and Telmon, N. (2009). Age at death estimation of adult males using coxal bone and CT scan: a preliminary study. *Forensic Science International*, 186:14–21.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications Ltd, London, 4 edition.

France, D. L. (1998). Observation and metric analysis of sex in the skeleton. In Reichs, K. J., editor, *Forensic osteology: advances in the identification of human remains*, pages 163–168. Charles C. Thomas, Springfield, Illinois.

Frayner, D. W. and Wolpoff, M. H. (1985). Sexual dimorphism. *Annual Review of Anthropology*, 14:429–473.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57:453–476.

Frelich, L. and Hunt, D. R. (2007). Morphological characteristics of ancestry in the fetal/newborn human skeleton. In *Proceedings of the 59th annual meeting of the American Academy of Forensic Sciences*, San Antonio, Texas. American Academy of Forensic Sciences.

Furmanová, K., Urbanová, P., and Kozlikova, B. (2017). Anthrovis: Visual analysis of 3d mesh ensembles for forensic anthropology. In *Proceedings of the 33rd Spring Conference on Computer Graphics*, pages 1–9.

Galdames, I. C. S., Russo, P. P., Matamala, D. A. Z., and Smith, R. L. (2009). Sexual dimorphism in the foramen magnum dimensions. *International Journal of Morphology*, 27:21–23.

Galdames, I. C. S., Zavando, M. D. A., and Smith, R. L. (2008). Evaluating accuracy and precision in morphologic traits for sexual dimorphism in malnutrition human skull: A comparative study. *International Journal of Morphology*, 26:877–881.

Gapert, R., Black, S., and Last, J. (2009a). Sex determination from the foramen magnum: discriminant function analysis in an eighteenth and nineteenth century British sample. *International Journal of Legal Medicine*, 123:25–33.

Gapert, R., Black, S., and Last, J. (2009b). Sex determination from the occipital condyle: discriminant function analysis in an eighteenth and nineteenth century British sample. *American Journal of Physical Anthropology*, 138:384–394.

Gapert, R., Black, S., and Last, J. (2013). Test of age-related variation in the craniometry of the adult human foramen magnum region: implications for sex determination methods. *Forensic Science, Medicine, and Pathology*, 9:478–488.

Garvin, H. M., Sholts, S. B., and Mosca, L. A. (2014). Sexual dimorphism in human cranial trait scores: Effects of population, age, and body size. *American Journal of Physical Anthropology*.

Gill, G. W. and Gilbert, R. (1990). Race identification from the midfacial skeleton: American Blacks and Whites. In Gill, G. W. and Rhine, S., editors, *Skeletal Attribution of Race*, pages 47–53. Maxwell Museum of Anthropology: Anthropology Papers No. 4, New Mexico.

González, P. N., Bernal, V., Perez, S. I., and Barrientos, G. (2007). Analysis of dimorphic structures in the human pelvis: its implications for sex estimation in samples without reference collections. *Journal of Archaeological Science*, 34:1720–1730.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gowland, R. and Chamberlain, A. (2002). A Bayesian approach to aging perinatal skeletal material from archaeological sites: Implications for the evidence for infanticide in Roman Britain. *International Journal of Osteoarchaeology*, 21:82–91.

Grabherr, S., Cooper, C., Ulrich-Bochsler, S., Uldin, T., Ross, S., Oesterhelweg, L., Bolliger, S., Christe, A., Schnyder, P., Mangin, P., and Thali, M. J. (2009). Estimation of sex and age of "virtual skeletons" - a feasibility study. *European Radiology*, 19:419–429.

Grumbach, M. M. and Kaplan, S. L. (1974). Fetal pituitary hormones and the maturation of the central nervous system regulation of anterior pituitary function. In Gluck, L., editor, *Modern perinatal medicine*, pages 247–272. Year Book Medical Publishers, Chicago.

Haglund, W. D. and Reay, D. T. (1993). Problems of recovering partial human remains at different times and locations: concerns for death investigators. *Journal of Forensic Sciences*, 38:69–80.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Science and Business Media, LLC, New York, 2 edition.

Hefner, J. T. (2009). Cranial nonmetric variation and estimating ancestry. *Journal of Forensic Sciences*, 54:985–995.

Hollimon, S. E. (2011). Sex and gender in bioarchaeological research - theory, method, and interpretation. In Agarwal, S. C. and Glencross, B. A., editors, *Social Bioarchaeology*, pages 149–182. Blackwell Publishing Ltd.

Holton, N. E., Alsamawi, A., Yokley, T. R., and Froehle, A. W. (2016). The ontogeny of nasal shape: An analysis of sexual dimorphism in a longitudinal sample. *American Journal of Physical Anthropology*. 10.1002/ajpa.22941.

Humphrey, L. T. (1998). Growth patterns in the modern human skeleton. *American Journal of Physical Anthropology*, 105:57–72.

Humphries, A. L. and Ross, A. H. (2011). Craniofacial sexual dimorphism in two Portuguese skeletal samples. *Anthropologie*, 49:13–20.

Işcan, M. Y. and Steyn, M. (2013). *The Human Skeleton in Forensic Medicine*. Charles C Thomas Pub Ltd, Springfield, Illinois.

Işcan, M. Y., Yoshino, M., and Kato, S. (1995). Sexual dimorphism in modern japanese crania. *American Journal of Human Biology*, 7:459–464.

Ioffe, S. and Szegedy, C. (2015). *Batch normalization: Accelerating deep network training by reducing internal covariate shift*.

Jantz, R. L. and Jantz, L. M. (2016). The remarkable change in Euro-American cranial shape and size. *Human Biology*, 88:56–64.

Jantz, R. L. and Ousley, S. D. (2005). *FORDISC*. Forensic Anthropology Center, University of Tennessee, Knoxville, 3.1 edition.

Johansen, A. C. (2014). *Validating the Use of 3D Photogrammetric Measurements on Human Crania*. Unpublished Honour's Bachelor of Science Thesis, University of Toronto.

Jolliffe, I. T. (2002). *Principal Component Analysis Series: Springer Series in Statistics*. Springer, New York, 2 edition.

Joyce, R. A. (2005). Archaeology of the body. *Annual Review of Anthropology*, 34:139–158.

Jung, H. and Woo, E. J. (2016). Evaluation of mastoid process as sex indicator in modern white Americans using geometric morphometrics. *Journal of Forensic Sciences*. 10.1111/1556-4029.13079.

Kakaliouras, A. M. (2014). When remains are "lost": thoughts on collections, repatriation, and research in American physical anthropology. *Curator: The Museum Journal*, 57:213–223.

Kanchan, T., Gupta, A., and Krishan, K. (2013). Estimation of sex from mastoid triangle - a craniometric analysis. *Journal of Forensic and Legal Medicine*, 20:855–860.

Keen, J. A. (1950). A study of the differences between male and female skulls. *American Journal of Physical Anthropology*, 8:65–79.

Kennedy, K. A. R. (1995). But professor, why teach race identification if races don't exist? *Journal of Forensic Sciences*, 40:797–800.

Kilroy, G. S. and Tallman, S. D. (2019). Secular change in macromorphoscopic trait frequencies in modern European Americans. In *88th Annual Meeting of the American Association of Physical Anthropologists, poster presentation*, Cleveland, Ohio. American Association of Physical Anthropologists.

Kimmerle, E. H., Ross, A., and Slice, D. (2008). Sexual dimorphism in America: geometric morphometric analysis of the craniofacial region. *Journal of Forensic Sciences*, 53:54–57.

Klepinger, L. L. (2006). *Fundamentals of forensic anthropology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Komar, D. (2004). Reassociating commingled remains separated by distance and time: the tale of Simon and Steven. In *Proceedings of the 56th annual meeting of the American Academy of Forensic Sciences*, Dallas, Texas. American Academy of Forensic Sciences.

Komar, D. A. and Potter, W. E. (2007). Percentage of body recovered and its effect on identification rates and cause and manner of death determination. *Journal of Forensic Sciences*, 52.

Krishan, K., Chatterjee, P. M., Kanchan, T., Kaur, S., Baryah, N., and Singh, R. K. (2016). A review of sex estimation techniques during examination of skeletal remains in forensic anthropology casework. *Forensic Science International*. http://dx.doi.org/10.1016/j.forsciint.2016.02.007.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification w withh deep convolutional neural networks. In *Proceedings of NIPS 2012*.

Krogman, W. M. and İşcan, M. Y. (1986). *The Human Skeleton in Forensic Medicine*. C. C. Thomas, Springfield, Illinois.

L'Abbé, E. N., Loots, M., and Meiring, J. H. (2005). The Pretoria bone collection: A modern South African skeletal sample. *Journal of Comparative Human Biology*, 56:197–205.

Lahr, M. M. (1996). *The Evolution of Modern Human Diversity: A Study of Cranial Variation*. Cambridge University Press, Cambridge.

Lam, J. F. (2014). *Using Photogrammetry to Generate 3D Models for Morphological Skeletal Age Estimation*. Unpublished Master's Thesis, University of Toronto.

Lam, J. F., Johansen, A. C., and Rogers, T. L. (2016). An evaluation of the Calce method for age estimation. *Journal of Forensic Sciences*, 61:1319–1321. doi: 10.1111/1556-4029.13134.

Larsen, C. S. and Walker, P. L. (2004). The ethics of bioarchaeology. In Turner, T., editor, *Ethical Issues in Biological Anthropology*, pages 111–122. State University of New York Press.

Liscio, E. (2014). Lecture notes 3b - laser scanners.

Loth, S. R. and Henneberg, M. (1996). Mandibular ramus flexure: a new morphologic indicator of sexual dimorphism in the human skeleton. *American Journal of Physical Anthropology*, 99:473–485.

Loth, S. R. and Henneberg, M. (1998). Mandibular ramus flexure is a good indicator of sexual dimorphism. *American Journal of Physical Anthropology*, 105:91–92.

Lovejoy, C. O., Meindl, R. S., Pryzbeck, T. R., and Mensforth, R. P. (1985). Chronological metamorphosis of the auricular surface of the ilium: A new method for the determination of adult skeletal age at death. *American Journal of Physical Anthropology*, 68:15–28.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision*, 2:1150–1157.

Luo, L., Wang, M., Tian, Y., Duan, F., Wu, Z., Zhou, M., and Rozenholc, Y. (2013). Automatic sex determination of skulls based on a statistical shape model. *Computational and Mathematical Methods in Medicine*.

Macaluso Jr., P. J. (2011). Metric sex determination from the basal region of the occipital bone in a documented French sample. *Bulletins et mémoires de la Société d'anthropologie de Paris*, 23:19–26.

Manthey, L., Jantz, R. L., Vitale, A., and Cattaneo, C. (2018). Population specific data improves fordisc's performance in italians. *Forensic Science International*, 292:263.e1–263.e7.

Marjanovic, M. (2007). A process for converting a set of image slices into a segmented 3d surface mesh. In *Dynamic Modeling of the Oral, Pharyngeal and Laryngeal Complex for Biomedical Applications*. http://artisynth.org/pmwiki.php?n=OPAL.MarkoMarjanovic.

Mays, S. (1993). Infanticide in Roman Britain. *Antiquity*, 67:883–888.

McCulloch, W. S. and Pitts, W. (1998). A logical calculus of the ideas immanent in nervous activity. In *Neurocomputing: Foundations of Research*, pages 15–27. MIT Press Cambridge, Johannesburg.

McKeown, A. H. and Schmidt, R. W. (2013). Geometric morphometrics. In DiGangi, E. A. and Moore, M. K., editors, *Research Methods in Human Skeletal Biology*, pages 325–359. Elsevier Inc.

Mellado, N., Mitra, N., and Aiger, D. (2014). Super 4pcs: Fast global pointcloud registration via smart indexing. *Eurographics*, 33:235–247.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Mitteroecker, P. and Gunz, P. (2009). Advances in geometric morphometrics. *Evolutionary Biology*, 36:235–247.

Mo, D. (2005). *A study of juvenile nasal morphology oriented toward ancestry differences and elucidation of ontogenetic patterns*. Master's thesis, University of Toronto.

Moore, M. K. (2013). Sex estimation and assessment. In DiGangi, E. A. and Moore, M. K., editors, *Research Methods in Human Skeletal Biology*, pages 91–116. Elsevier Inc.

Navega, D., Coelho, C., Vicente, R., Ferreira, M. T., Wasterlain, S., and Cunha, E. (2015). AncesTrees: ancestry estimation with randomized decision trees. *International Journal of Legal Medicine*, 129:1145–1153.

Nikita, E. (2014). Age-associated variation and sexual dimorphism in adult cranial morphology: Implications in anthropological studies. *International Journal of Osteoarchaeology*, 24:557–569.

Pickering, R. B. and Bachman, D. C. (1997). *The use of forensic anthropology*. CRC Press, Boca Raton.

Pillin, E. J. A. J. (2019). *A Holistic Approach to Fingerprint Identification*. Unpublished Doctoral Thesis, University of Leicester.

Pitzer, B. (2015). *Automatic Reconstruction of Textured 3D models*. KIT Scientific Publishing, Karlsruhe, Germany.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2016). PointNet: Deep learning on point sets for 3D classification and segmentation. *CoRR*, abs/1612.00593.

Ramsthaler, F., Kettner, M., Gehl, A., and Verhoff, M. A. (2010). Digital forensic osteology: morphological sexing of skeletal remains using volume-rendered cranial ct scans. *Forensic Science International*, 195:148–152.

Raschka, S. (2015). *Python Machine Learning*. PACKT Publishing, Birmingham.

Relethford, J. H. (2009). Race and global patterns of phenotypic variation. *American Journal of Physical Anthropology*, 139:16–22.

Rhine, S. (1990). Non-metric skull racing. In Gill, G. W. and Rhine, S., editors, *Skeletal Attribution of Race*, pages 9–20. Maxwell Museum of Anthropology: Anthropology Papers No. 4, New Mexico.

Roberts, C. and Manchester, K. (2005). *The Archaeology of Disease*. Cornell University Press, New York, 3 edition.

Robinson, M. S. and Bidmos, M. A. (2009).   The skull and humerus in the determination of sex: Reliability of discriminant function equations. *Forensic Science International*, 186:86.e1–86.e5.

Robling, A. G. and Ubelaker, D. H. (1997). Sex estimation from the metatarsals. *Journal of Forensic Sciences*, 42:1062–1069.

Rogers, T. L. (1999). A visual method of determining the sex of skeletal remains using the distal humerus. *Journal of Forensic Sciences*, 44:57–60.

Rogers, T. L. (2005). Determining the sex of human remains through cranial morphology. *Journal of Forensic Sciences*, 50:493–500.

Rogers, T. L. (2009). Skeletal age estimation. In Blau, S. and Ubelaker, D. H., editors, *Handbook of Forensic Archaeology and Anthropology*, pages 208–221. Left Coast Press, Inc.

Rogers, T. L. and Allard, T. T. (2004). Expert testimony and positive identification of human remains through cranial suture patterns. *Journal of Forensic Sciences*, 49:203–207.

Rosas, A. and Bastir, M. (2002). Thin-plate spline analysis of allometry and sexual dimorphism in the human craniofacial complex. *American Journal of Physical Anthropology*, 117:236–245.

Ross, A. H., Baker, L. E., and Falsetti, A. (2003). Sexual dimorphism a proxy for environmental sensitivity? A multitemporal view. *Journal of the Washington Academy of Sciences*, 89:1–12.

Ross, A. H., Slice, D. E., and Williams, S. E. (2010). Geometric morphometric tools for the classification of human skulls. Technical report, U.S. Department of Justice. Document No. 231195.

Rusu, R. B. and Cousins, S. (2011). 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China.

Saly, A. (2014). *Examination of Three-Dimensional Images for Metric Analyses of Ancestry and a New Method for Determining Ancestry from the Femur*. Unpublished Master's Thesis, University of Toronto.

Schladitz, K. (2011). Quantitative micro-ct. *Journal of Microscopy*, 243:111–117.

Schwartz, J. H. (1995). *Skeleton keys*. Oxford University Press, Oxford.

Shearer, B. M., Sholts, S. B., Garvin, H. M., and W^{f}arml^{f}ander, S. K. T. S. (2012). Sexual dimorphism in human browridge volume measured from 3d models of dry crania: A new digital morphometrics approach. *Forensic Science International*, 222:400.e1–400.e5.

Sholts, S. B., Walker, P. L., Kuzminsky, S. C., Miller, K. W., and W^{f}arml^{f}ander, S. K. (2011). Identification of group affinity from cross-sectional contours of the human midfacial skeleton using digital morphometrics and 3d laser scanning technology. *Journal of Forensic Science*, 56:333–338.

Sholts, S. B., W^{f}arml^{f}ander, S. K., Flores, L. M., Miller, K. W. P., and Walker, P. L. (2010). Variation in the measurement of cranial volume and surface area using 3D laser scanning technology. *Journal of Forensic Science*, 55:871–876.

Sidler, M., Jackowski, C., Dimhofer, R., Vock, P., and Thali, M. (2007). Use of multislice computed tomography in disaster victim identification – advantages and limitations. *Forensic Science International*, 169:118–128.

Slice, D. E. (2005). Modern morphometrics. In Slice, D. E., editor, *Modern Morphometrics in Physical Anthropology*, pages 1–24. Kluwer Academic/Plenum Publishers, New York.

Slice, D. E. (2007). Geometric morphometrics. *Annual Review of Anthropology*, 36:261–281.

Slice, D. E. and Ross, A. (2014). *3D-ID. Geometric Morphometric Classification of Crania for Forensic Scientists.* NCSU Forensic Analysis Laboratory, North Carolina State University, 2014-11-1 edition.

Spradley, M. K. and Jantz, R. L. (2011). Sex estimation in forensic anthropology: Skull versus postcranial elements. *Journal of Forensic Sciences*, 56:289–296.

St. Hoyme, L. E. and Işcan, M. Y. (1989). Determination of sex and race: accuracy and assumptions. In Işcan, M. Y. and Kennedy, K. A. R., editors, *Reconstruction of life from the skeleton*, pages 53–93. Wiley-Liss, New York.

Stevenson, J. C., Mahoney, E. R., Walker, P. L., and Everson, P. M. (2009). Technical note: Prediction of sex based on five skull traits using decision analysis (CHAID). *American Journal of Physical Anthropology*, 139:434–441.

Steyn, M. and Işcan, M. Y. (1998). Sexual dimorphism in the crania and mandibles of South African Whites. *Forensic Science International*, 98:9–16.

Stinson, S., Bogin, B., Huss-Ashmore, R., and O'Rourke, D. (2012). *Human Biology: An Evolutionary and Biocultural Perspective*. Wiley-Blackwell, Hoboken, New Jersey.

Stojanowski, C. M. and Schillaci, M. A. (2006). Phenotypic approaches for understanding patterns of intracemetary biological variation. *Yearbook of Physical Anthropology*, 49:49–88.

Sutter, R. C. (2003). Nonmetric subadult skeletal sexing traits: I. a blind test of the accuracy of eight previously proposed methods using prehistoric known-sex mummies from Northern Chile. *Journal of Forensic Sciences*, 48:1–9.

Telmon, N., Gaston, A., Chemla, P., Blanc, A., Joffre, F., and Rougé, D. (2005). Application of the Suchey-Brooks method to three-dimensional imaging of the pubic symphysis. *Journal of Forensic Sciences*, 50:1–6.

Thayer, Z. A. and Dobson, S. (2010). Sexual dimorphism in chin shape: implications for adaptive hypotheses. *American Journal of Physical Anthropology*, 143:417–425.

Trentin, E., Lusnig, L., and Cavalli, F. (2018). Parzen neural networks: Fundamentals, properties, and an application to forensic anthropology. *Neural Networks*, 97:137–151.

Tsurumoto, T., Saiki, K., Okamoto, K., Imamura, T., Maeda, J., Manabe, Y., and Wakebe, T. (2013). Periarticular osteophytes as an appendicular joint stress marker (JSM: Analysis in a contemporary Japense skeletal collection. *PLOS One*, 8:e57049.

Ubelaker, D. H. (2008). Forensic anthropology:' methodology and diversity of applications. In Katzenberg, M. A. and Saunders, S. R., editors, *Biological Anthropology of the Human Skeleton*, pages 41–69. John Wiley & Son, Inc., 2 edition.

Veroni, A., Nikitovic, D., and Schillaci, M. A. (2010). Brief communication: Sexual dimorphism of the juvenile basicranium. *American Journal of Physical Anthropology*, 141:147–151.

Vidarsdottir, U. S. (1999). *Changes in the form of the facial skeleton during growth: a comparative morphometric study of modern humans and Neanderthals*. Unpublished PhD Thesis, University of London.

Vidarsdottir, U. S. and O'Higgins, P. (2001). Development of sexual dimorphism in the facial skeleton of anatomically modern *Homo sapiens*. *American Journal of Physical Anthropology*, 114:144.

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37:360–363.

Villa, C., Buckberry, J., Cattaneo, C., and Lynnerup, N. (2013). Technical Note: Reliability of Suchey-Brooks and Buckberry-Chamberlain methods on 3D visualizations from CT and laser scans. *American Journal of Physical Anthropology*, 151:158–163.

Walker, P. L. (1995). Problems of preservation and sexism in sexing: some lessons from historical collections for palaeodemographers. In Saunders, S. R. and Herring, A., editors, *Grave reflections: portraying past through cemetery studies*, pages 31–48. Canadian Scholars' Press Inc., Toronto.

Walker, P. L. (2008a). Bioarchaeological ethics: A historical perspective on the value of human remains. In Katzenberg, M. A. and Saunders, S. R., editors, *Biological Anthropology of the Human Skeleton*, pages 3–40. John Wiley & Son, Inc., 2 edition.

Walker, P. L. (2008b). Sexing skulls using discriminant function analysis of visually assessed traits. *American Journal of Physical Anthropology*, 136:39–50.

Weinberg, S. M., Putz, D. A., Mooney, M. P., and Siegel, M. I. (2005). Evaluation of non-metric variation in the crania of black and white perinates. *Forensic Science International*, 151:177–185.

White, T. D. (1991). *Human Osteology*. Academic Press, Cambridge, 3 edition.

White, T. D., Black, M. T., and Folkens, P. A. (2012). *Human Osteology*. Academic Press, San Diego.

White, T. D. and Folkens, P. A. (2005). *The Human Bone Manual*. Elsevier Inc., San Diego.

Wilkinson, C. (2004). *Forensic Facial Reconstruction*. University Press, Cambridge.

Williams, B. A. and Rogers, T. L. (2006). Evaluating the accuracy and precision of cranial morphological traits for sex determination. *Journal of Forensic Sciences*, 51. 10.1111/j.1556-4029.2006.00177.x.

Wood, C. (2015). The age-related emergence of cranial morphological variation. *Forensic Science International*, 251:220.e1–220.e20.

APPENDIX A: Ethical Approval

UNIVERSITY OF LEICESTER University Ethics Sub-Committee for Science and Engineering and Arts Humanities

27/02/2016

Ethics Reference: 5474-jfl6-scharchaeolgy&anchist

TO:
Name of Researcher Applicant: Jessica Lam
Department: Archaeology & Ancient History
Research Project Title: Using Novel 3D Comparative Techniques to Assess Skeletal Remains

Dear Jessica Lam,

RE: Ethics review of Research Study application

The University Ethics Sub-Committee for Science and Engineering and Arts Humanities has reviewed and discussed the above application.

1.   Ethical opinion

The Sub-Committee grants ethical approval to the above research project on the basis described in the application form and supporting documentation, subject to the conditions specified below.

2.   Summary of ethics review discussion

The Committee noted the following issues:
We are approving the application on the understanding that the personal information collected would be age, sex, the population to which the individual belongs, and any disease/pathology and/or trauma that could affect the subsequent analysis, and that no recording of names of individuals will take place. Instead each individual will be assigned an individual or sample number.

3.   General conditions of the ethical approval

The ethics approval is subject to the following general conditions being met prior to the start of the project:

As the Principal Investigator, you are expected to deliver the research project in accordance with the University's policies and procedures, which includes the University's Research Code of Conduct and the University's Research Ethics Policy.

If relevant, management permission or approval (gate keeper role) must be obtained from host organisation prior to the start of the study at the site concerned.

4.   Reporting requirements after ethical approval

You are expected to notify the Sub-Committee about:
- Significant amendments to the project
- Serious breaches of the protocol
- Annual progress reports
- Notifying the end of the study

5.   Use of application information

Details from your ethics application will be stored on the University Ethics Online System. With your permission, the Sub-Committee may wish to use parts of the application in an anonymised format for training or sharing best practice. Please let me know if you do not want the application details to be used in this manner.

Best wishes for the success of this research project.

Yours sincerely,

Prof. Paul Cullis
Chair

APPENDIX B: Data Sheets

\<Name of Skeletal Collection\>

Visual Sex Assessment
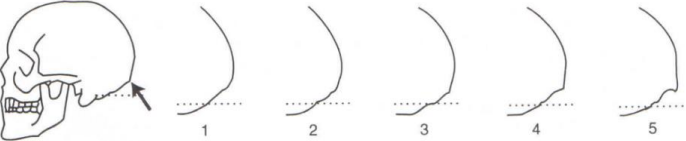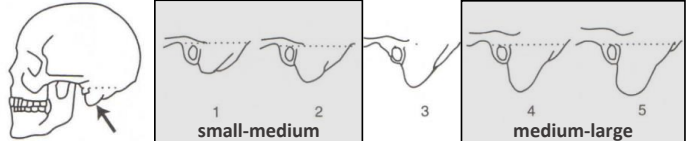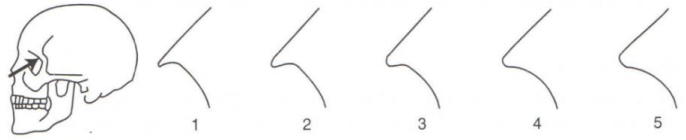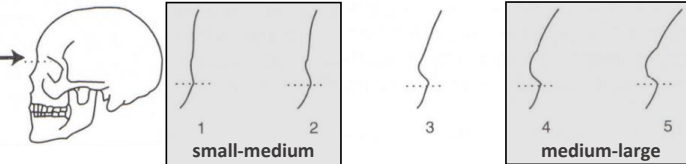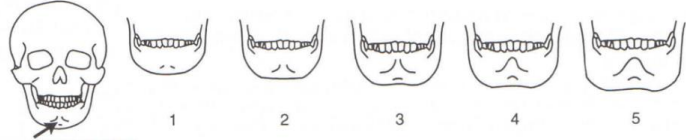
Jessica Frances Lam
jfl6@le.ac.uk
University of Leicester
INTREPID Forensics Programme

Individual Reference #: _____

Date of Assessment: _____
(MM/DD/YY)

Photo # Range: _____

Sex Assessment – Buikstra & Ubelaker's *Standards* and Williams & Rogers' Traits

Trait	Visualization & Score (circle)
Nuchal Crest *(Lateral profile)* Rugosity associated to attachment of nuchal musculature; <u>ignore contour of underlying bone</u>	 **1** = smooth, no bony projections visible in lateral profile **5** = massive nuchal crest that projects a considerable distance; well-defined bony ledge or hook
Mastoid Process **(Assess R & L)** Compare size with surrounding structures (e.g. EAM & zygomatic process); most important variable is <u>volume of mastoid process</u>, not length	 **1** = very small process; projects a small distance below inferior margin of EAM & digastric groove **5** = length and width several times that of EAM
Supra-Orbital Margin **(Assess R & L)** Hold finger against margin of orbit at lateral aspect of supraorbital foramen; hold edge of orbit between fingers to determine thickness	 **1** = extremely sharp bolder, e.g. slightly dulled knife **2** = thick, rounded margin with curvature approximating a pencil
Supra-Orbital Ridge/ Glabella *(Lateral profile)* Compare with diagrams	 **1** = smooth contour of frontal, little or no projection at midline **5** = massive glabellar prominence, rounded loaf-shaped projection (well-developed)
Zygomatic Extension **(Assess R & L)**	**1** = does not extend past EAM **5** = extends past EAM
Nasal Aperture	**1** = lower, wider, rounded margins **3** = intermediate **5** = high, thin, sharp margins
Size & Architecture	**1** = small/smooth **3** = intermediate **5** = big/rugged
Mental Eminence Hold mandible with thumbs on either side of mental eminence; move thumbs medially until they delimit the lateral borders	 **1** = little or no projection of mental eminence above surrounding bone **5** = massive mental eminence; occupies most of the anterior portion of mandible

Comments: _____

_____

pg. _____ / _____

<Name of Skeletal Collection>

Visual Sex Assessment
(Intraobserver Error)
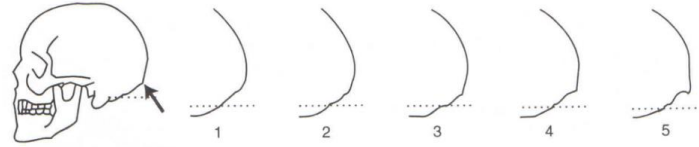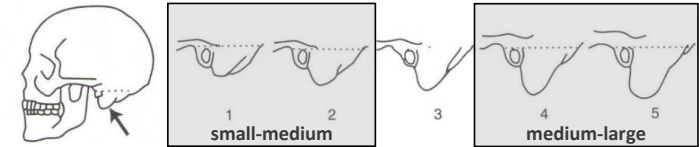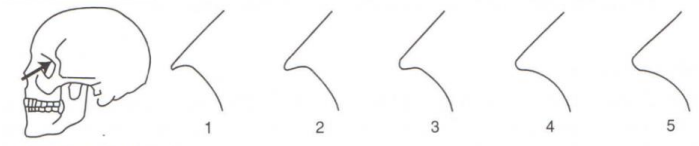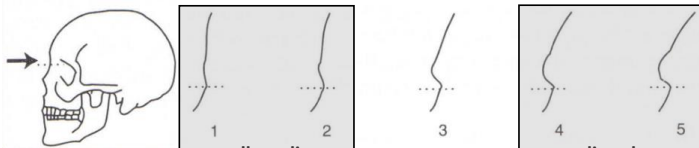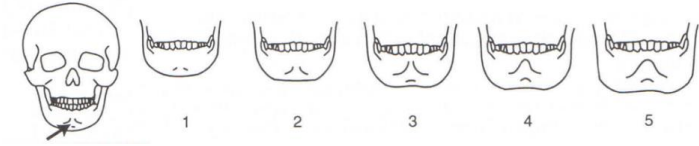
Jessica Frances Lam
jfl6@le.ac.uk
University of Leicester
INTREPID Forensics Programme

Individual Reference #: _____

Date of Assessment: _____
(MM/DD/YY)

Sex Assessment – Buikstra & Ubelaker's *Standards* and Williams & Rogers' Traits

Trait	Visualization & Score (circle)
Nuchal Crest *(Lateral profile)* Rugosity associated to attachment of nuchal musculature; <u>ignore contour of underlying bone</u>	 **1** = smooth, no bony projections visible in lateral profile **5** = massive nuchal crest that projects a considerable distance; well-defined bony ledge or hook
Mastoid Process **(Assess R & L)** Compare size with surrounding structures (e.g. EAM & zygomatic process); most important variable is <u>volume of mastoid process</u>, not length	 small-medium medium-large **1** = very small process; projects a small distance below inferior margin of EAM & digastric groove **5** = length and width several times that of EAM
Supra-Orbital Margin **(Assess R & L)** Hold finger against margin of orbit at lateral aspect of supraorbital foramen; hold edge of orbit between fingers to determine thickness	 **1** = extremely sharp bolder, e.g. slightly dulled knife **2** = thick, rounded margin with curvature approximating a pencil
Supra-Orbital Ridge/ Glabella *(Lateral profile)* Compare with diagrams	 small-medium medium-large **1** = smooth contour of frontal, little or no projection at midline **5** = massive glabellar prominence, rounded loaf-shaped projection (well-developed)
Zygomatic Extension **(Assess R & L)**	**1** = does not extend past EAM **5** = extends past EAM
Nasal Aperture	**1** = lower, wider, rounded margins **3** = intermediate **5** = high, thin, sharp margins
Size & Architecture	**1** = small/smooth **3** = intermediate **5** = big/rugged
Mental Eminence Hold mandible with thumbs on either side of mental eminence; move thumbs medially until they delimit the lateral borders	 **1** = little or no projection of mental eminence above surrounding bone **5** = massive mental eminence; occupies most of the anterior portion of mandible

Comments: _____
_____

pg. _____ / _____

<Skeletal Collection>

3D Models:
List of Scans & Parameters

Jessica Frances Lam
jfl6@le.ac.uk
University of Leicester
INTREPID Forensics Programme

1 = complete (~100%)   *2 = fairly complete (≥ 75%)*   *3 = incomplete (<75%) / fragmented*

Individual #	Cranium *(circle)*			Scanning Parameters	Mandible *(circle)*			Scanning Parameters
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:
	1	2	3	Exposure: Brightness:	1	2	3	Exposure: Brightness:

10 samples per side   pg. _____ / _____

# APPENDIX C: Trait Distribution Graphs for the SB Collection

C.1 Nuchal Crest
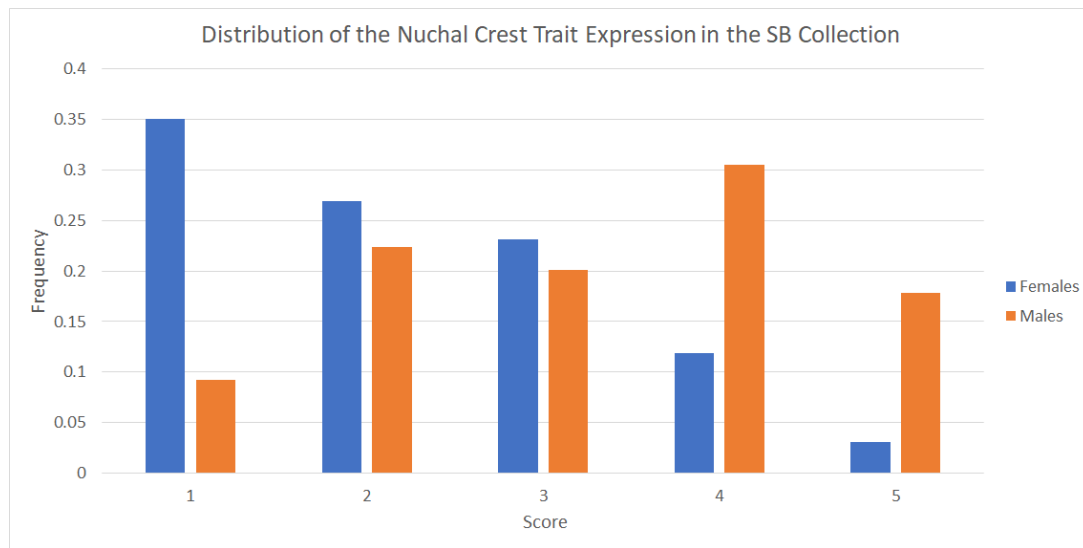
Figure C.1: The distribution of the nuchal crest trait expression in the SB Collection represented using a bar chart. Females are in blue while males are in orange.
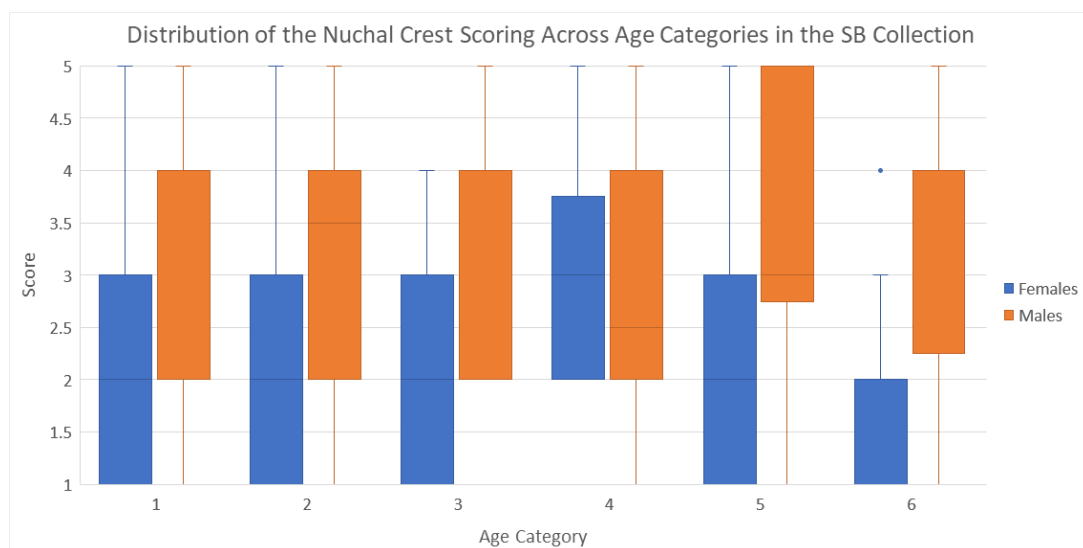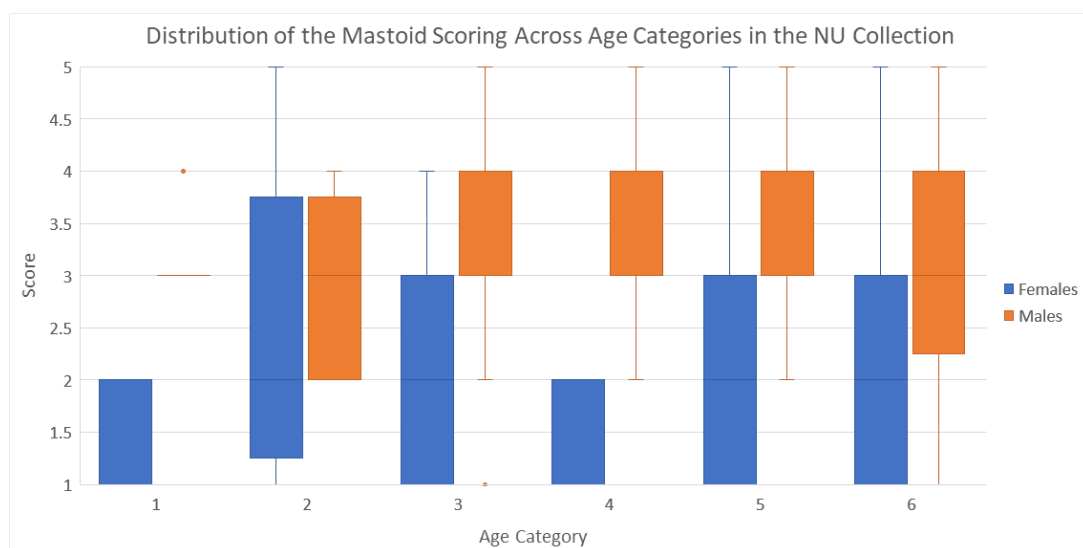

Figure C.2: A boxplot distribution of nuchal crest scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table C.1: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the SB collection when comparing nuchal crest trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 18 M = 26	F = 2.0 M = 4.0	$U = 330.0$ $p = 0.019$ $z = -1.79$ $r = -0.27$	0.829
2	F = 22 M = 18	F = 2.0 M = 3.5	$U = 272.5$ $p = 0.039$ $z = -4.85$ $r = -0.77$	0.806
3	F = 20 M = 14	F = 2.0 M = 4.0	$U = 233.0$ $p < 0.001$ $z = -4.09$ $r = -0.70$	0.879
4	F = 20 M = 32	F = 3.0 M = 2.5	$U = 277.5$ $p = 0.411$ $z = -4.75$ $r = -0.66$	0.745
5	F = 48 M = 50	F = 2.0 M = 3.0	$U = 1739.5$ $p \ll 0.001$ $z = -4.52$ $r = -0.46$	0.820
6	F = 32 M = 32	F = 1.0 M = 4.0	$U = 904.5$ $p \ll 0.001$ $z = -1.82$ $r = -0.23$	0.878
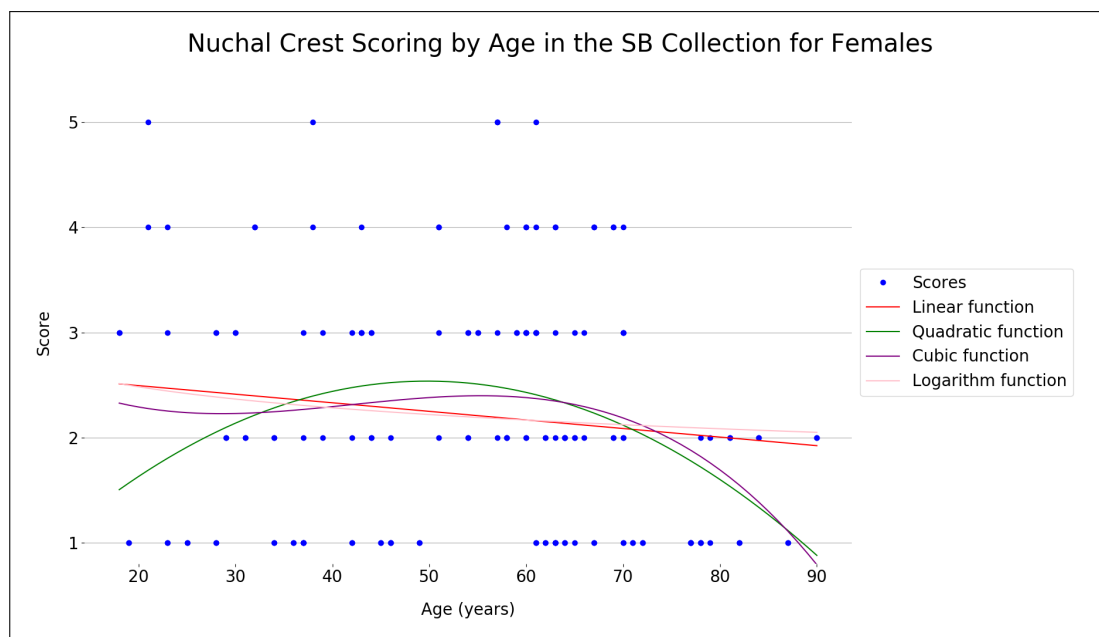
Figure C.3: A scatterplot of age vs. nuchal crest trait scoring for females in the SB collection, with four fitting functions.
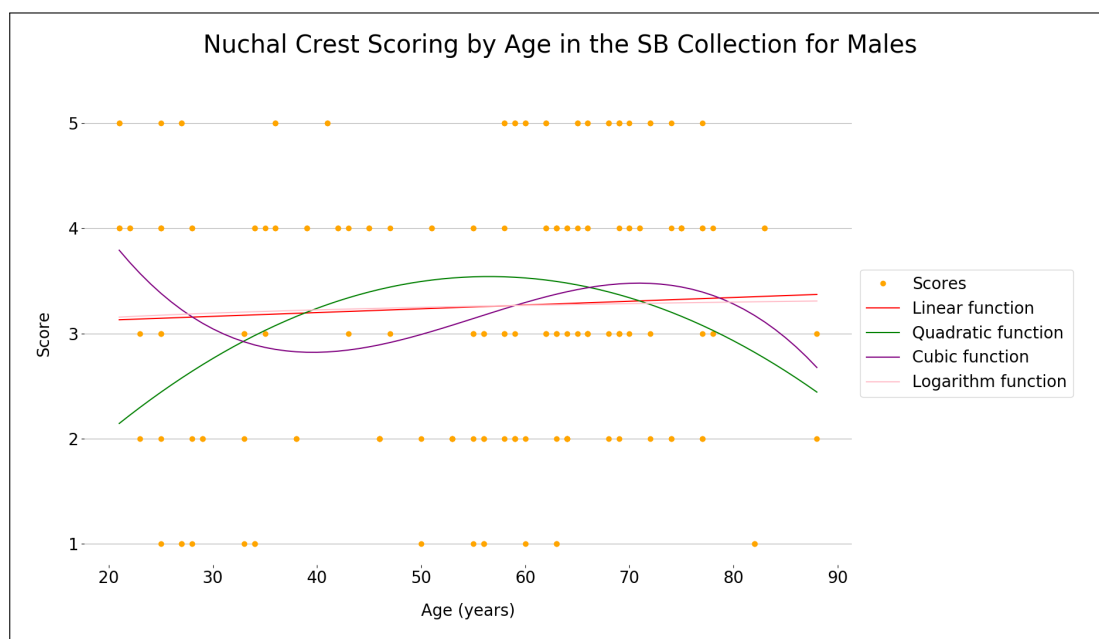
Figure C.4: A scatterplot of age vs. nuchal crest trait scoring for males in the SB collection, with four fitting functions.
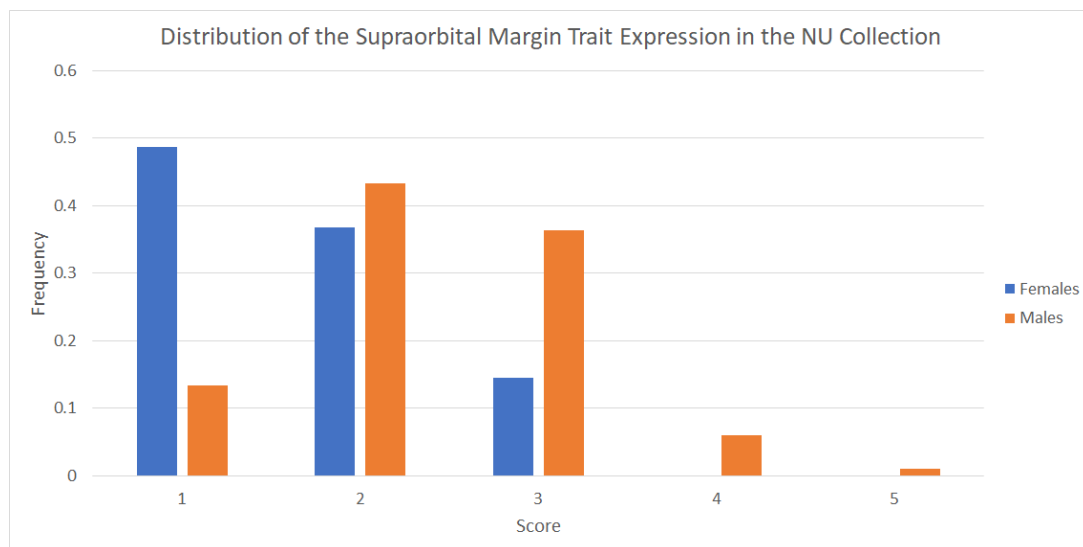
C.2   Mastoid Process

Figure C.5: The distribution of the mastoid process trait expression in the SB Collection represented using a bar chart. Females are in blue while males are in orange.
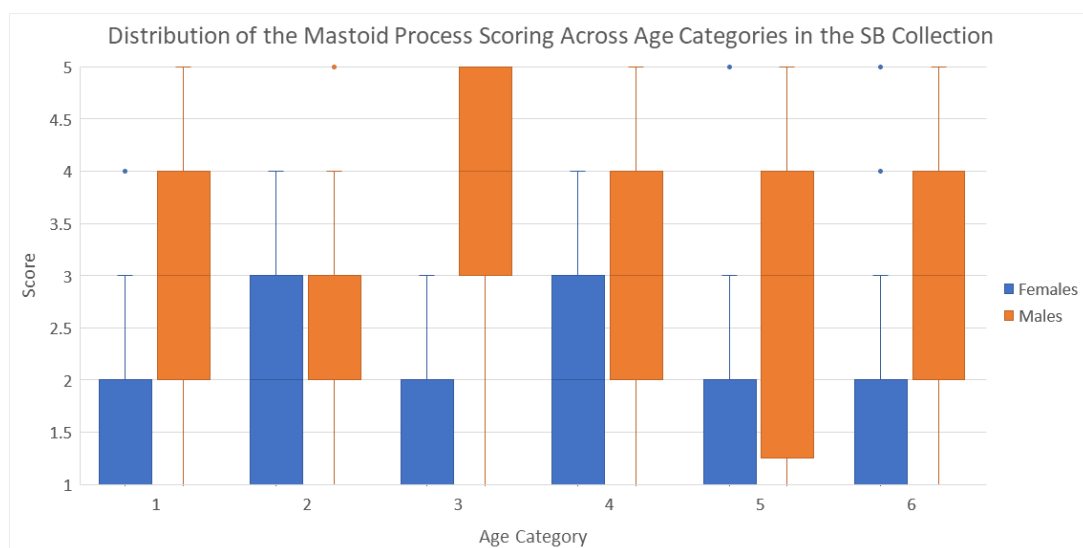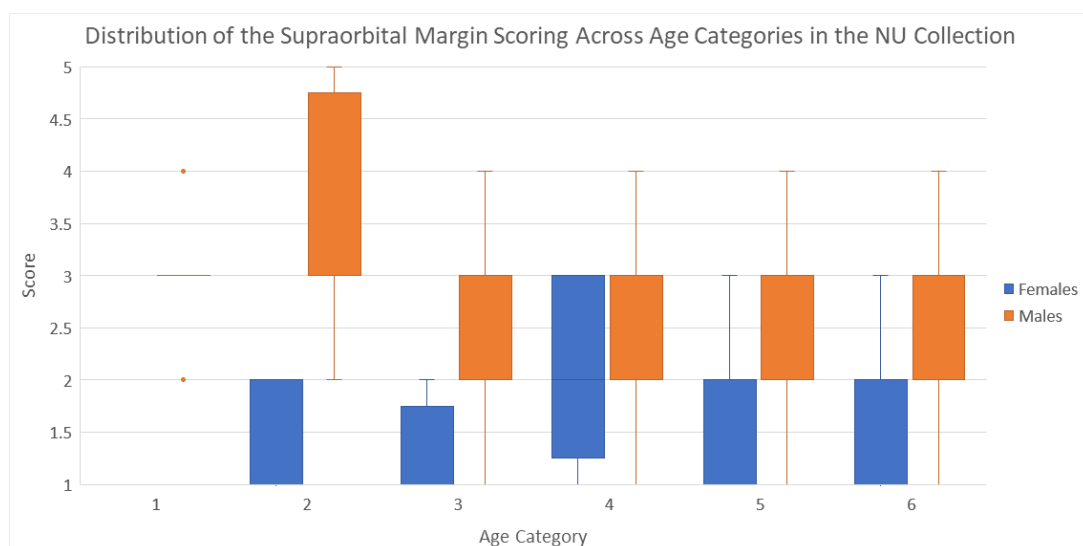

Figure C.6: A boxplot distribution of mastoid process scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table C.2: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the SB collection when comparing mastoid process scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 32 M = 52	F = 1.0 M = 3.0	$U = 1343.5$ $p << 0.001$ $z = $ -0.15 $r = $ -0.02	0.830
2	F = 50 M = 38	F = 2.0 M = 3.0	$U = 1195.0$ $p = 0.032$ $z = $ -8.68 $r = $ -0.92	0.761
3	F = 34 M = 36	F = 1.5 M = 4.0	$U = 1137.5$ $p << 0.001$ $z = $ -0.82 $r = $ -0.10	0.924
4	F = 48 M = 58	F = 2.0 M = 3.0	$U = 2109.0$ $p << 0.001$ $z = $ -2.91 $r = $ -0.28	0.762
5	F = 94 M = 83	F = 2.0 M = 3.0	$U = 6025.5$ $p << 0.001$ $z = $ -6.88 $r = $ -0.52	0.810
6	F = 82 M = 68	F = 2.0 M = 3.0	$U = 4374.5$ $p << 0.001$ $z = $ -6.86 $r = $ -0.56	0.810
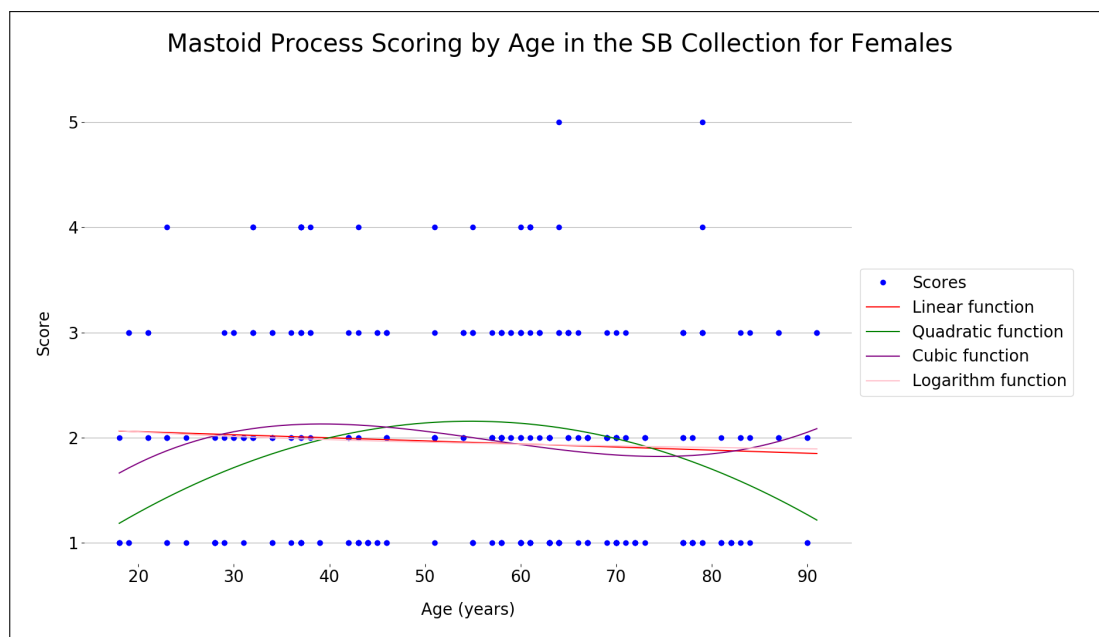
Figure C.7: A scatterplot of age vs. mastoid process trait scoring for females in the SB collection, with four fitting functions.
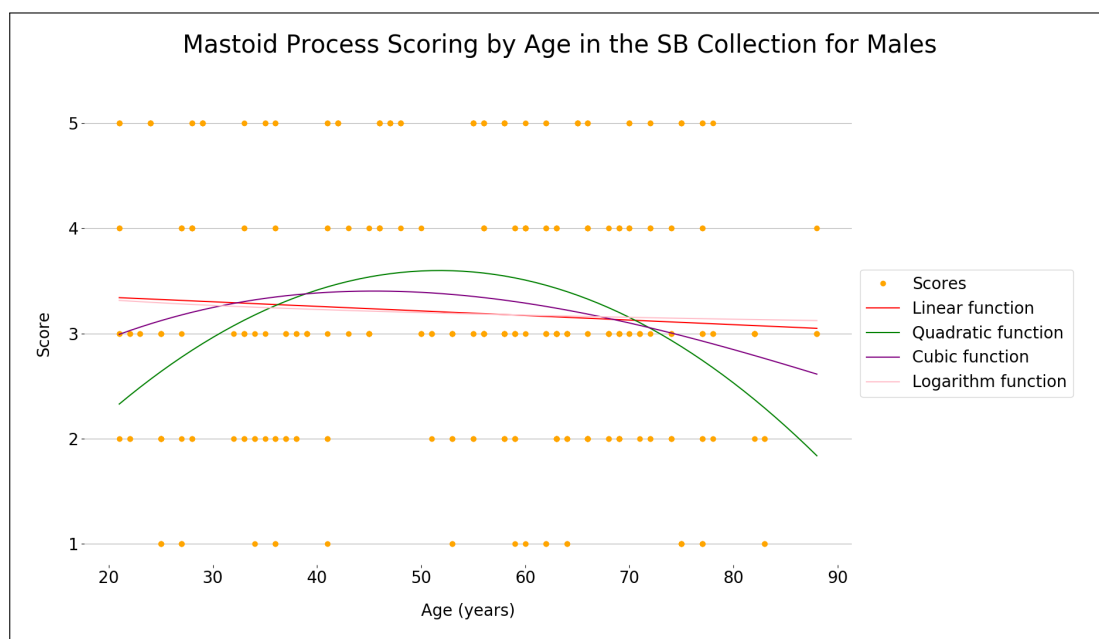
Figure C.8: A scatterplot of age vs. mastoid process trait scoring for males in the SB collection, with four fitting functions.
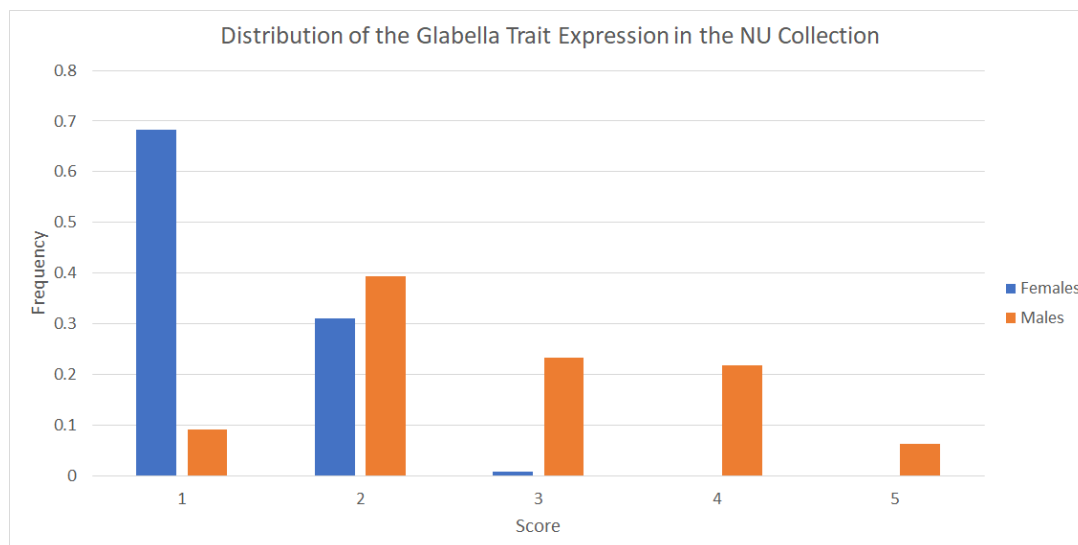
C.3 Supraorbital Margin
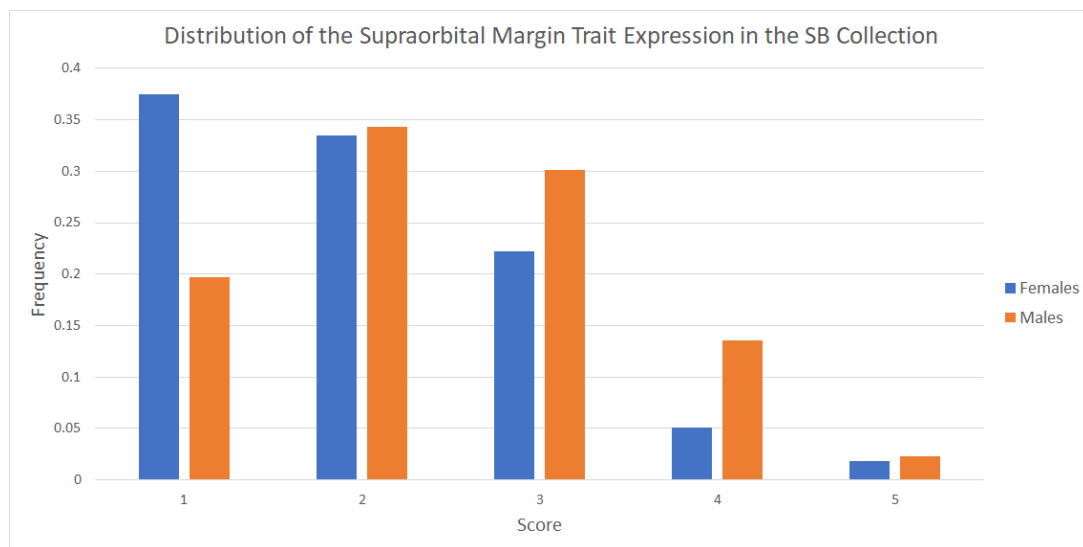
Figure C.9: The distribution of the supraorbital margin trait expression in the SB Collection represented using a bar chart. Females are in blue while males are in orange.
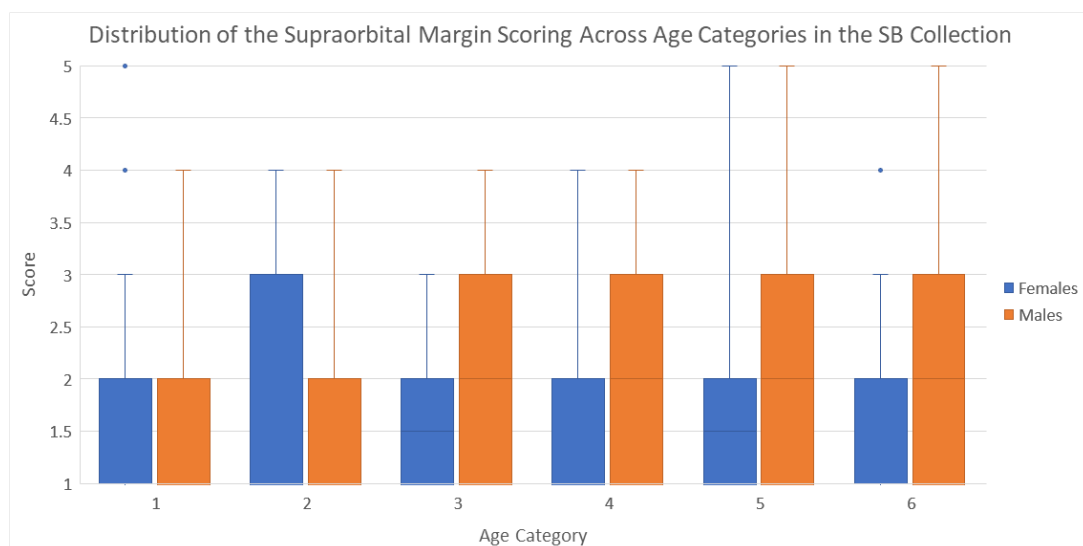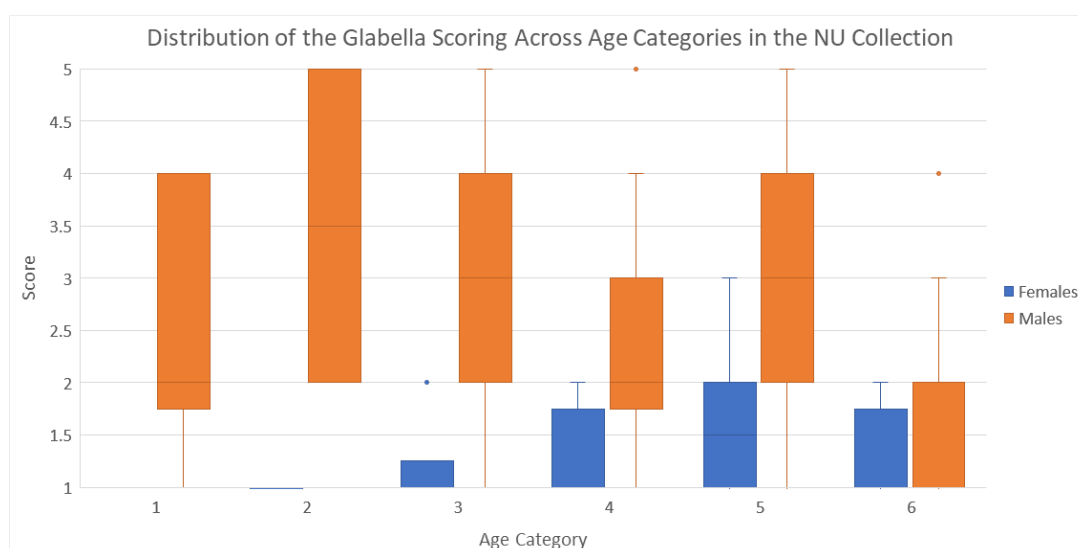

Figure C.10: A boxplot distribution of supraorbital margin scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table C.3: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the SB collection when comparing supraorbital margin scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 30 M = 38	F = 2.0 M = 2.0	$U = 620.0$ $p = 0.518$ $z = $ -5.13 $r = $ -0.62	0.702
2	F = 36 M = 29	F = 2.0 M = 2.0	$U = 536.5$ $p = 0.846$ $z = $ -8.60 $r = $ -1.07	0.742
3	F = 30 M = 28	F = 2.0 M = 3.0	$U = 637.0$ $p < 0.001$ $z = $ -3.86 $r = $ -0.51	0.762
4	F = 38 M = 43	F = 1.0 M = 3.0	$U = 1189.0$ $p < 0.001$ $z = $ -3.49 $r = $ -0.39	0.781
5	F = 74 M = 67	F = 2.0 M = 3.0	$U = 3024.5$ $p = 0.020$ $z = $ -9.20 $r = $ -0.78	0.763
6	F = 67 M = 50	F = 2.0 M = 2.0	$U = 2078.0$ $p = 0.020$ $z = $ -10.33 $r = $ -0.96	0.711
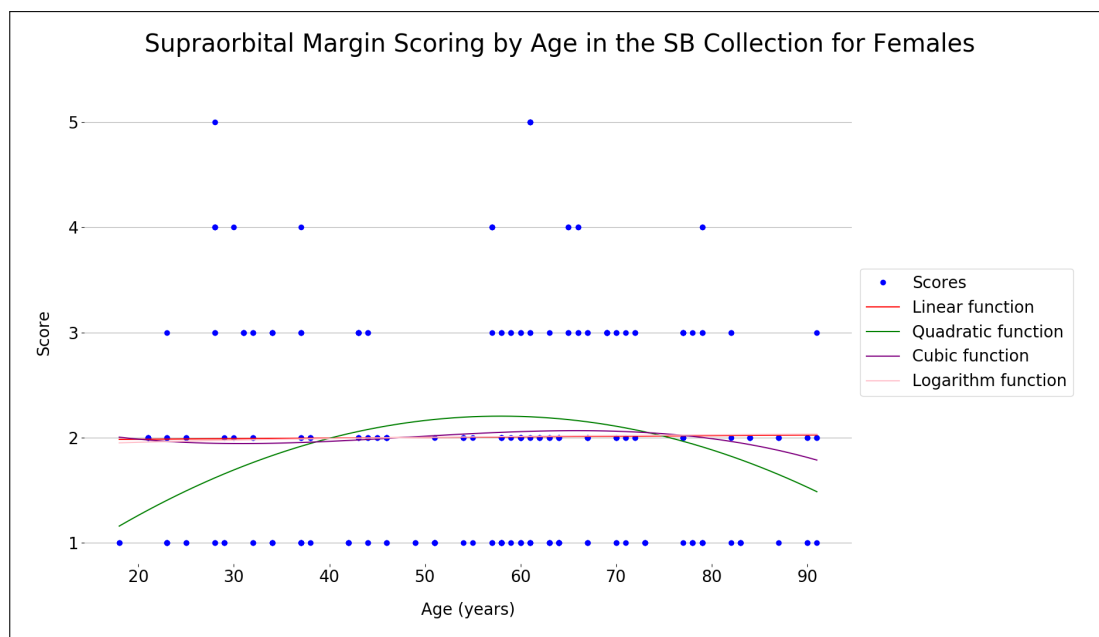
Figure C.11: A scatterplot of age vs. supraorbital margin trait scoring for females in the SB collection, with four fitting functions.
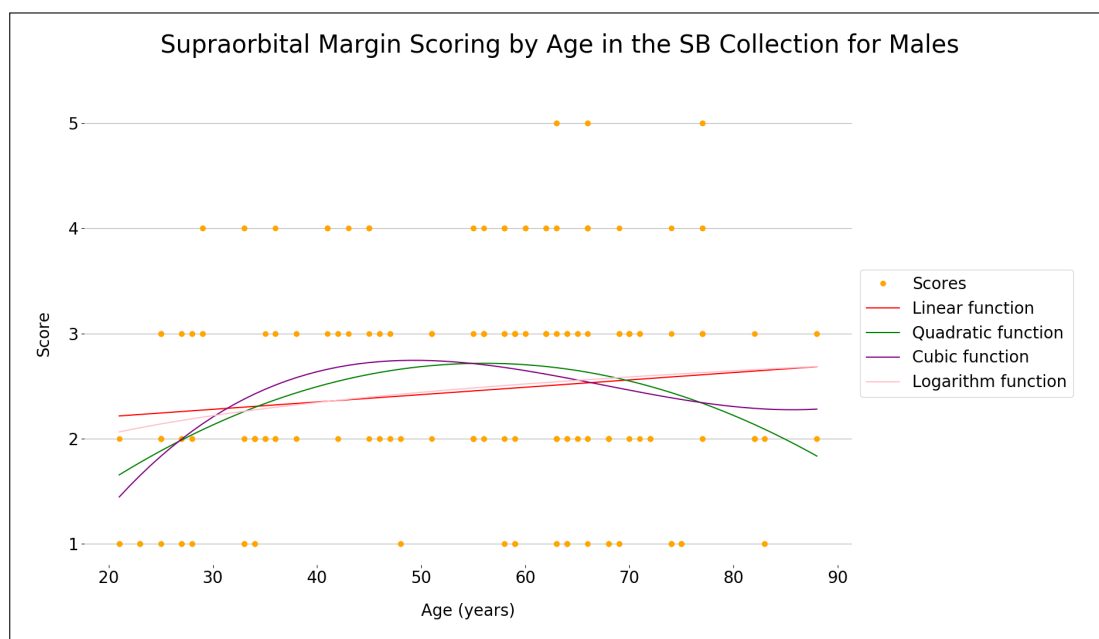
Figure C.12: A scatterplot of age vs. supraorbital margin trait scoring for males in the SB collection, with four fitting functions.

C.4   Glabella
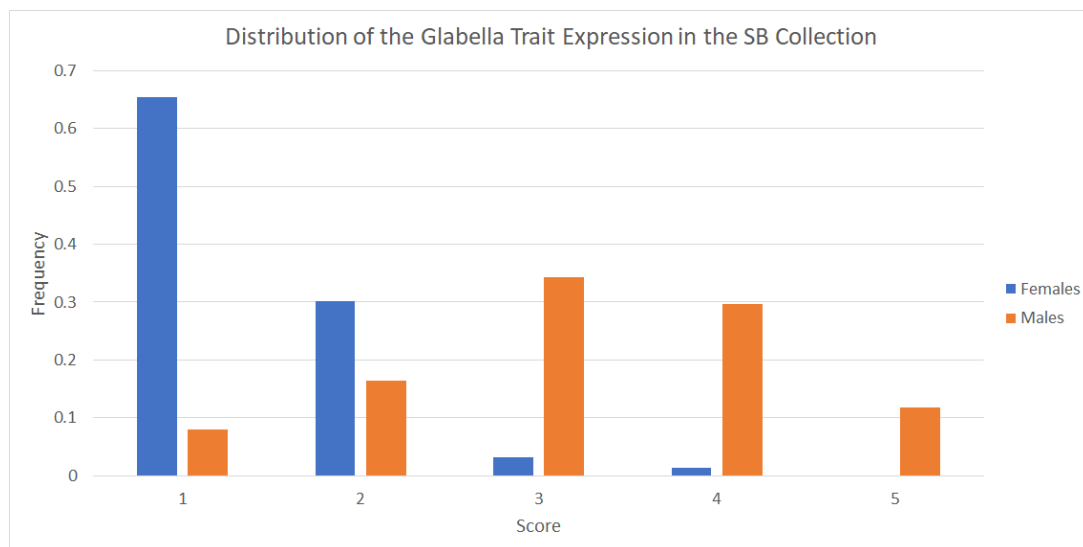
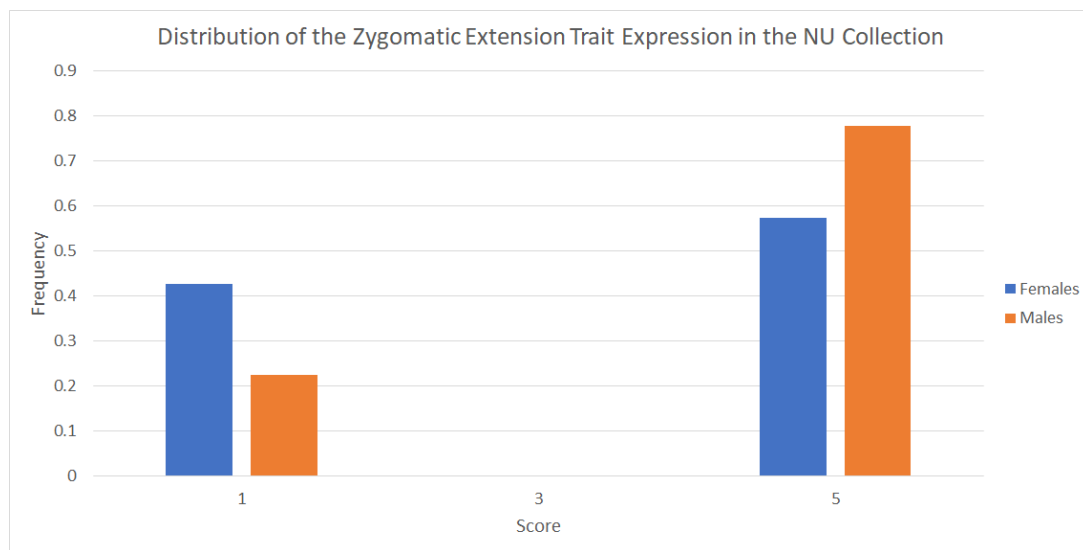**Figure C.13:** The distribution of the glabella trait expression in the SB Collection represented using a bar chart. Females are in blue while males are in orange.
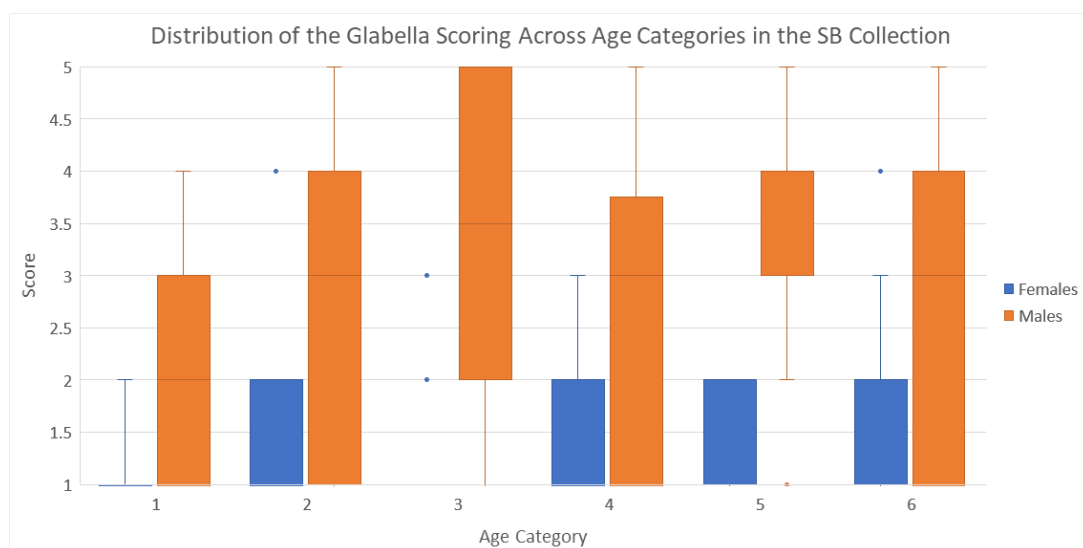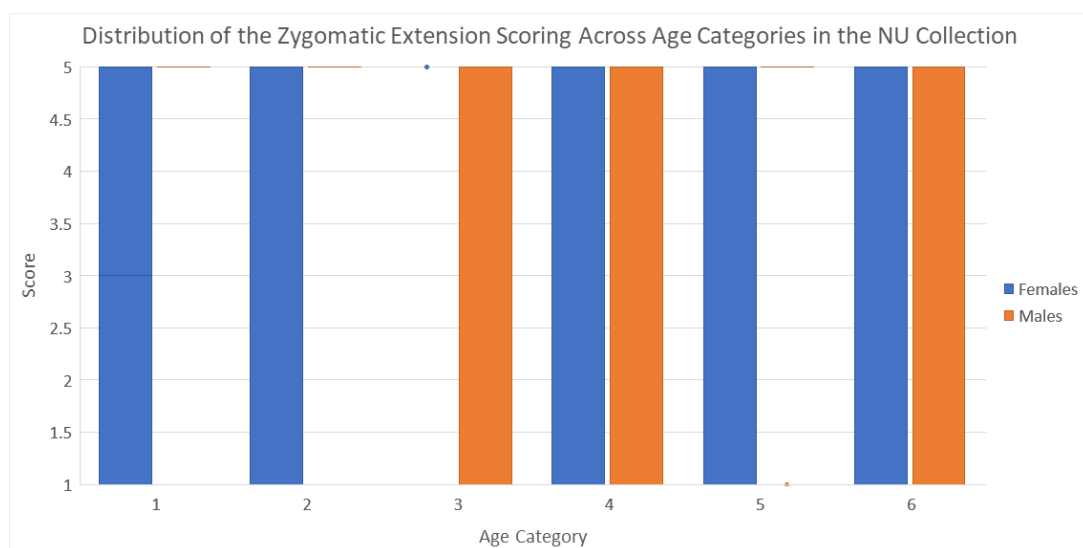

Figure C.14: A boxplot distribution of glabella scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table C.4: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the SB collection when comparing glabella trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 12 M = 20	F = 1.0 M = 3.0	$U = 213.0$ $p < 0.001$ $z = 0.58$ $r = 0.10$	0.825
2	F = 20 M = 18	F = 1.0 M = 3.0	$U = 307.0$ $p < 0.001$ $z = -2.43$ $r = -0.39$	0.856
3	F = 18 M = 20	F = 1.0 M = 3.0	$U = 330.0$ $p << 0.001$ $z = -0.61$ $r = -0.10$	0.883
4	F = 20 M = 22	F = 1.0 M = 4.0	$U = 415.5$ $p << 0.001$ $z = -0.37$ $r = -0.06$	0.911
5	F = 46 M = 39	F = 1.0 M = 3.0	$U = 1657.0$ $p << 0.001$ $z = -2.83$ $r = -0.31$	0.904
6	F = 40 M = 31	F = 1.0 M = 4.0	$U = 1119.5$ $p << 0.001$ $z = -3.72$ $r = -0.44$	0.894
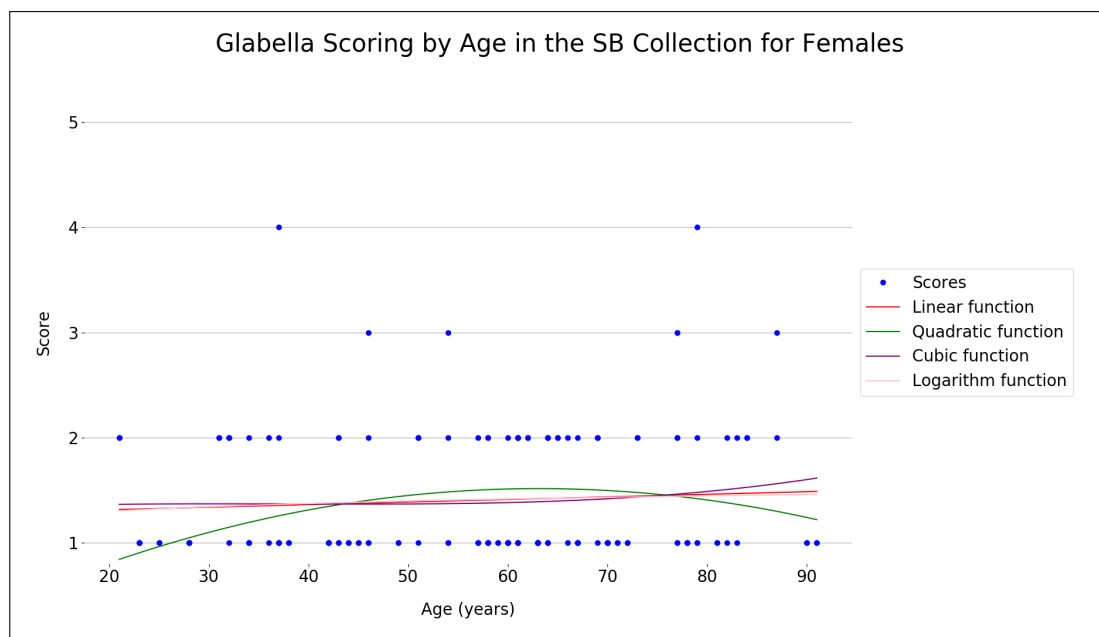
Figure C.15: A scatterplot of age vs. glabella trait scoring for females in the SB collection, with four fitting functions.
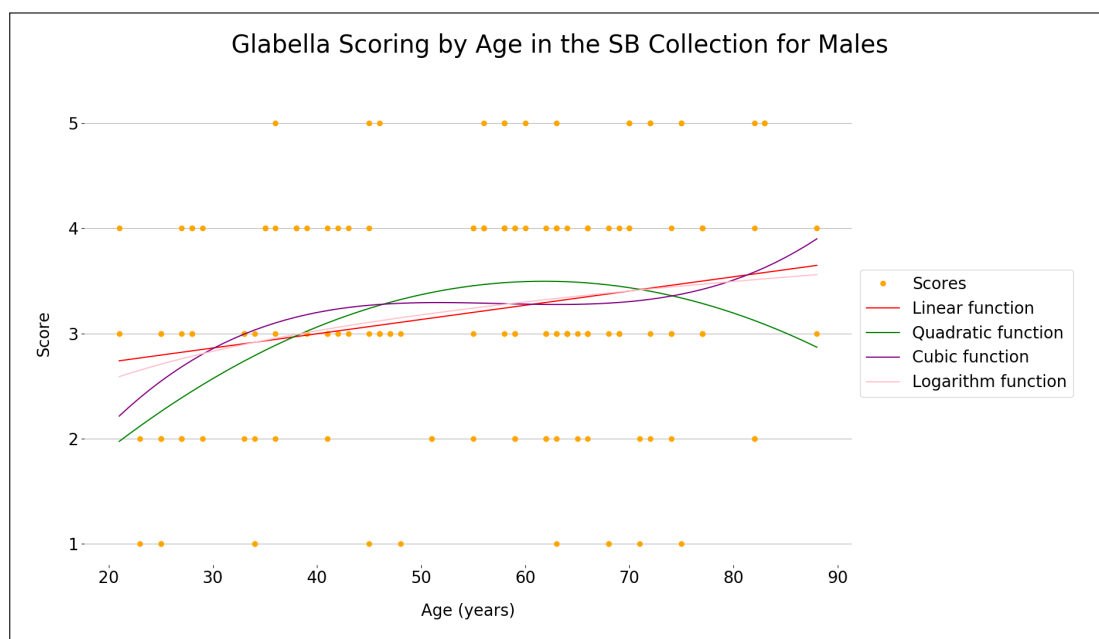
Figure C.16: A scatterplot of age vs. glabella trait scoring for males in the SB collection, with four fitting functions.

C.5   Zygomatic Extension
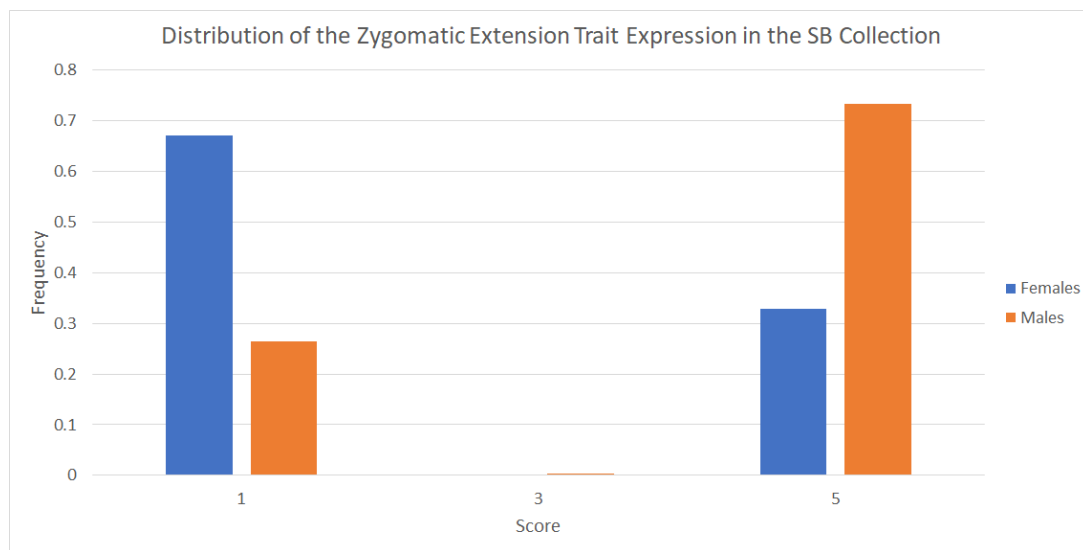
Figure C.17: The distribution of the zygomatic extension trait expression in the SB Collection represented using a bar chart. Females are in blue while males are in orange.
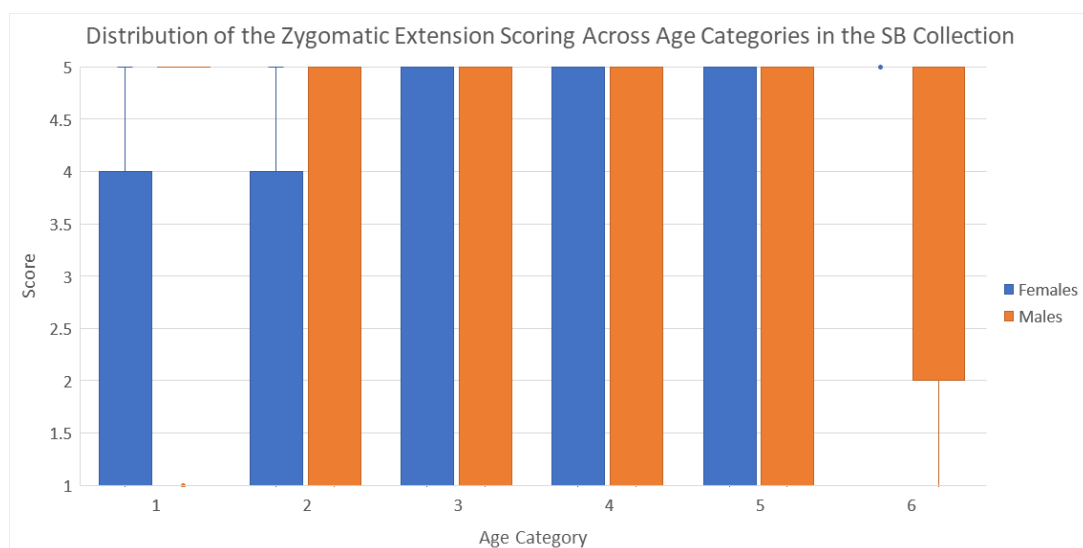
Figure C.18: A boxplot distribution of zygomatic extension scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table C.5: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the SB collection when comparing zygomatic extension trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 33 M = 54	F = 1.0 M = 5.0	$U = 1390.5$ $p << 0.001$ $z = $ -0.54 $r = $ -0.06	0.652
2	F = 44 M = 34	F = 1.0 M = 5.0	$U = 1143.0$ $p << 0.001$ $z = $ -6.00 $r = $ -0.68	0.632
3	F = 38 M = 38	F = 1.0 M = 3.0	$U = 779.0$ $p = 0.497$ $z = $ -7.11 $r = $ -0.82	0.500
4	F = 46 M = 58	F = 1.0 M = 5.0	$U = 1622.0$ $p = 0.028$ $z = $ -5.19 $r = $ -0.51	0.515
5	F = 102 M = 94	F = 1.0 M = 5.0	$U = 6672.0$ $p << 0.001$ $z = $ -8.51 $r = $ -0.61	0.572
6	F = 78 M = 65	F = 1.0 M = 5.0	$U = 3971.5$ $p << 0.001$ $z = $ -6.67 $r = $ -0.56	0.661
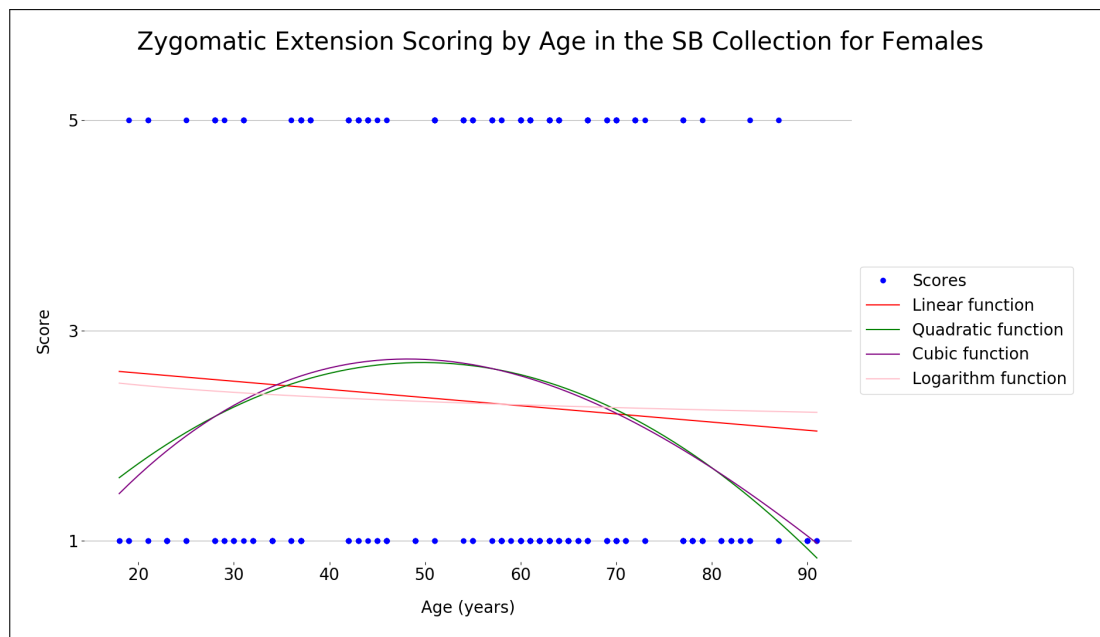
Figure C.19: A scatterplot of age vs. zygomatic extension trait scoring for females in the SB collection, with four fitting functions.
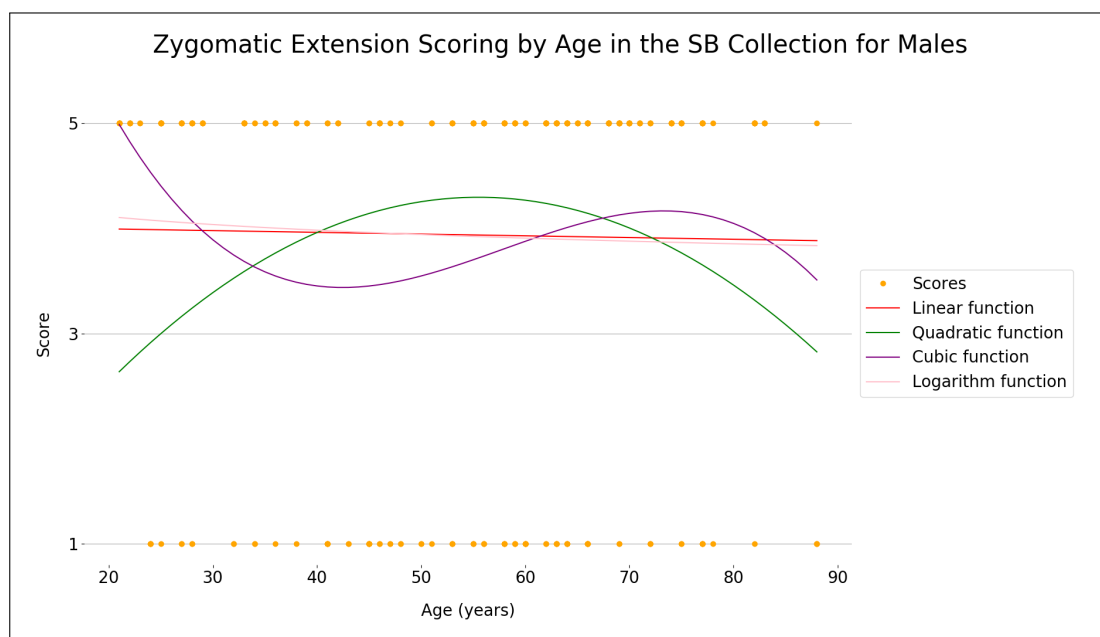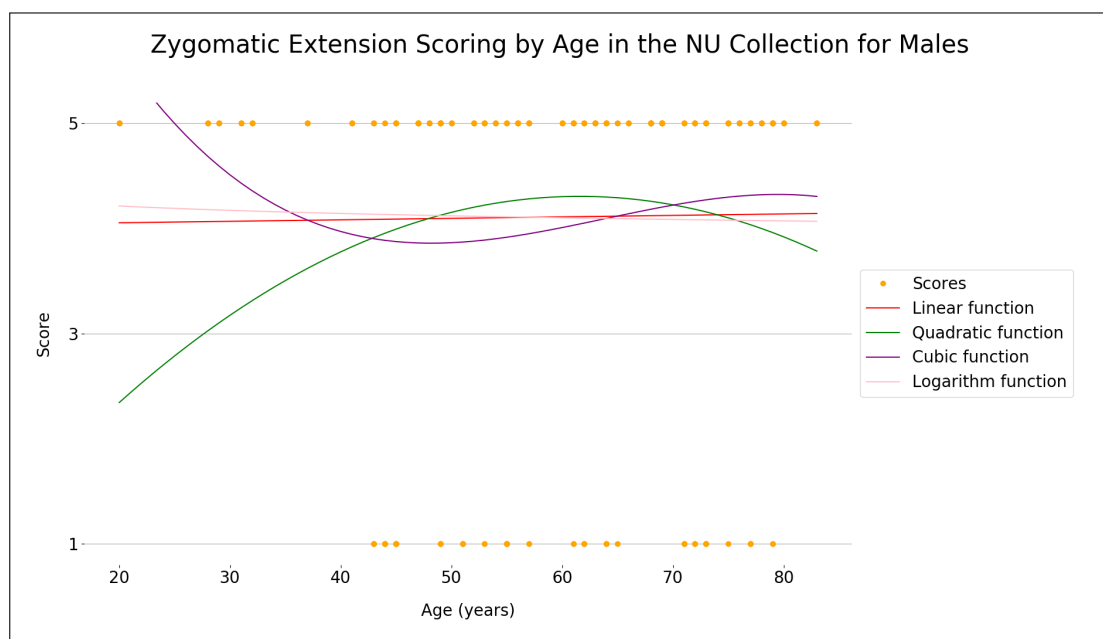

Figure C.20: A scatterplot of age vs. zygomatic extension trait scoring for males in the SB collection, with four fitting functions.
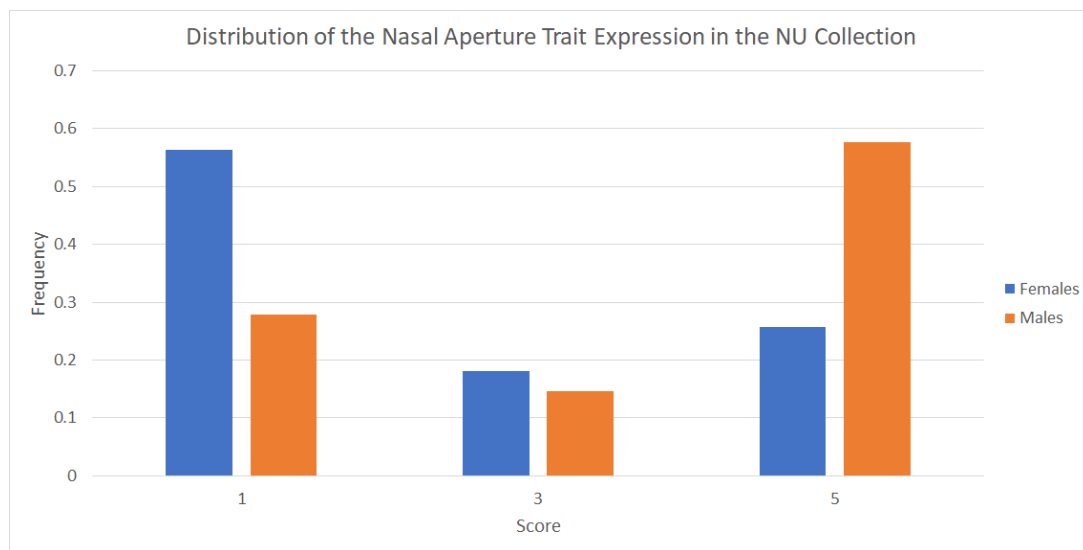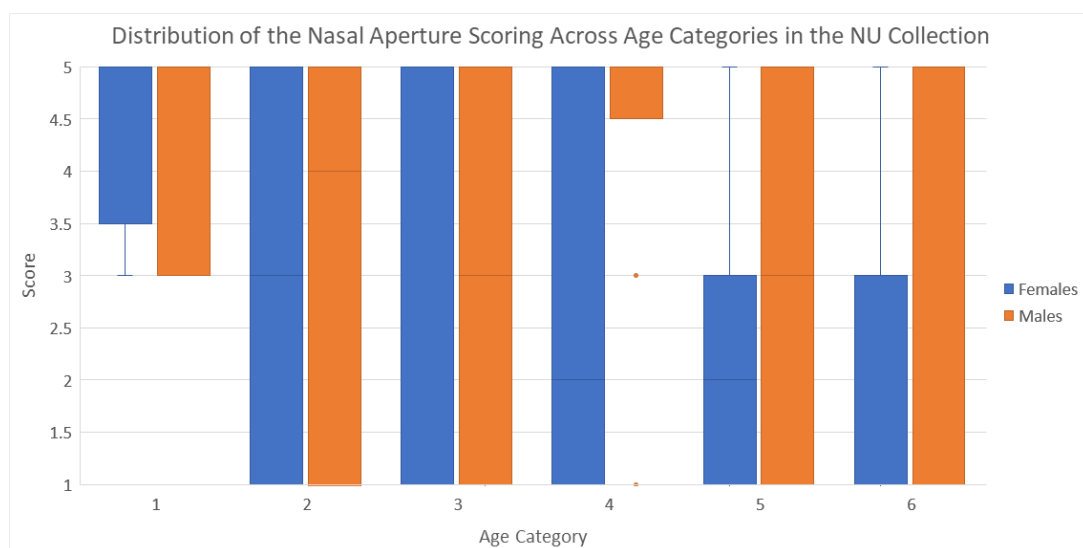
C.6   Nasal Aperture
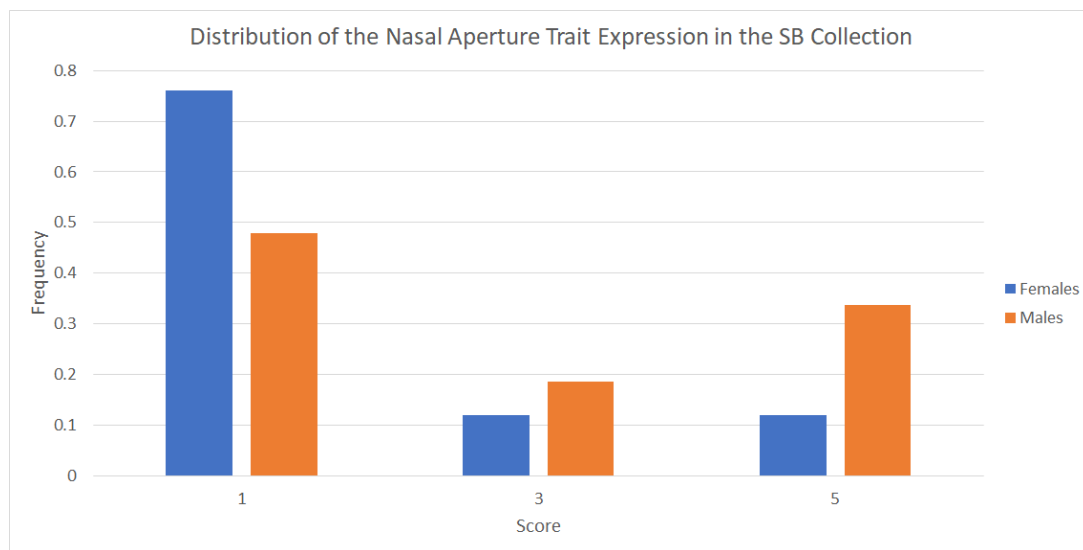
Figure C.21: The distribution of the nasal aperture trait expression in the SB Collection represented using a bar chart. Females are in blue while males are in orange.
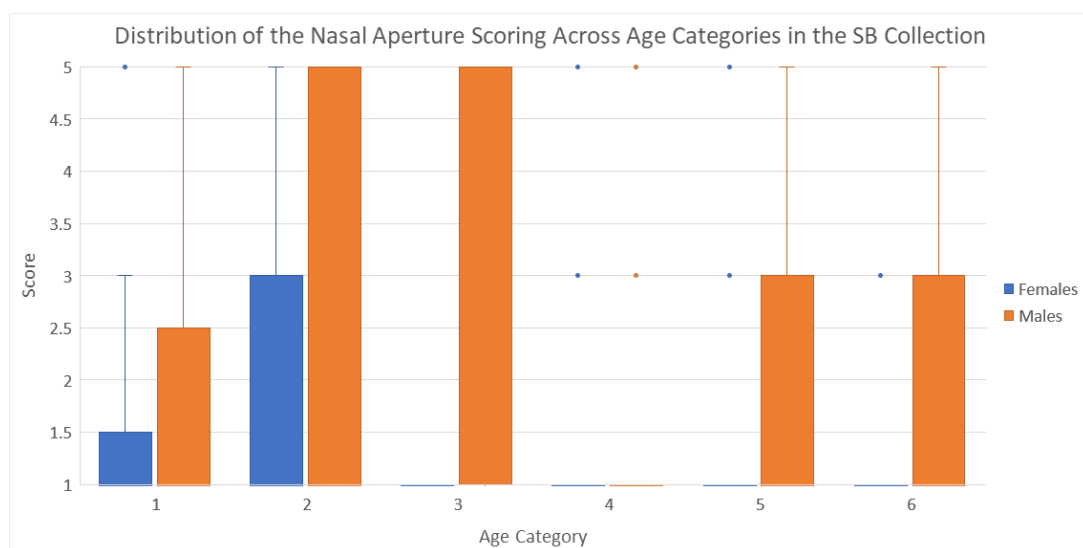
Figure C.22: A boxplot distribution of nasal aperture scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table C.6: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the SB collection when comparing nasal aperture trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 10 M = 14	F = 1.0 M = 2.0	$U = 80.0$ $p = 0.537$ $z = $ -2.63 $r = $ -0.54	0.600
2	F = 18 M = 14	F = 1.0 M = 3.0	$U = 152.5$ $p = 0.284$ $z = $ -5.49 $r = $ -0.97	0.647
3	F = 13 M = 16	F = 1.0 M = 3.0	$U = 148.0$ $p = 0.017$ $z = $ -2.06 $r = $ -0.38	0.500
4	F = 18 M = 18	F = 1.0 M = 1.0	$U = 179.0$ $p = 0.506$ $z = $ -4.87 $r = $ -0.81	0.438
5	F = 34 M = 28	F = 1.0 M = 3.0	$U = 621.0$ $p = 0.021$ $z = $ -6.37 $r = $ -0.81	0.616
6	F = 32 M = 21	F = 1.0 M = 3.0	$U = 489.5$ $p < 0.001$ $z = $ -6.81 $r = $ -0.94	0.546
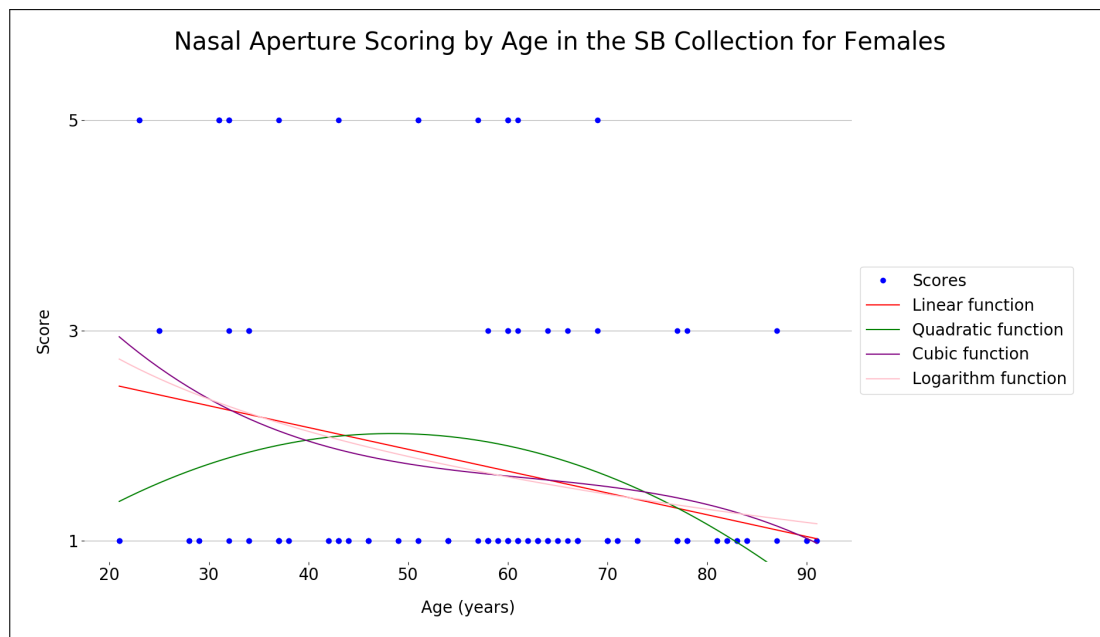
Figure C.23: A scatterplot of age vs. nasal aperture trait scoring for females in the SB collection, with four fitting functions.
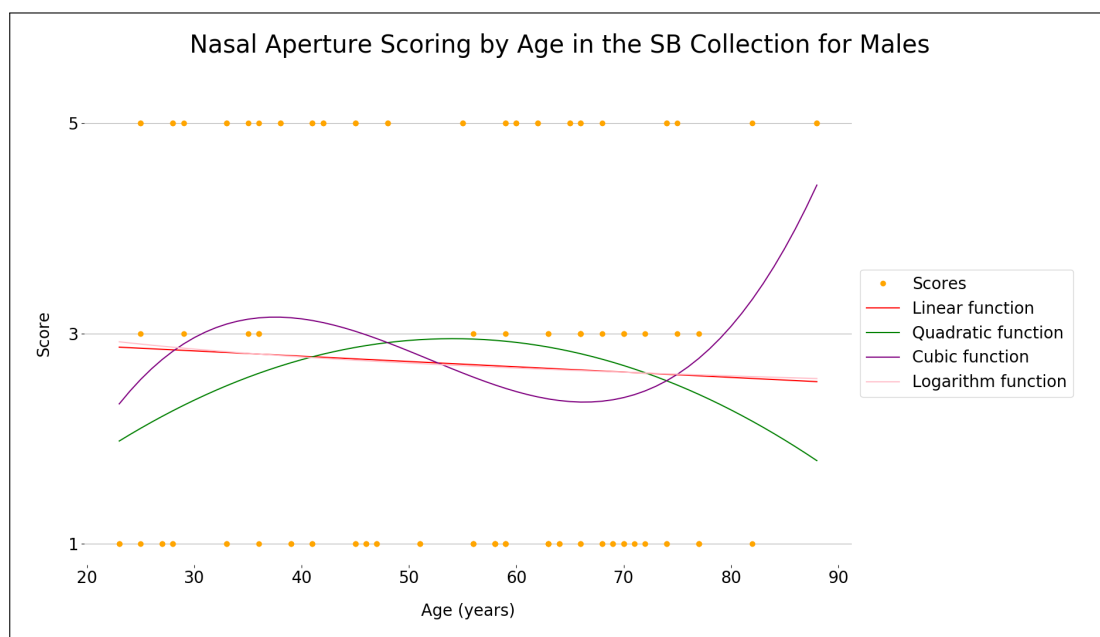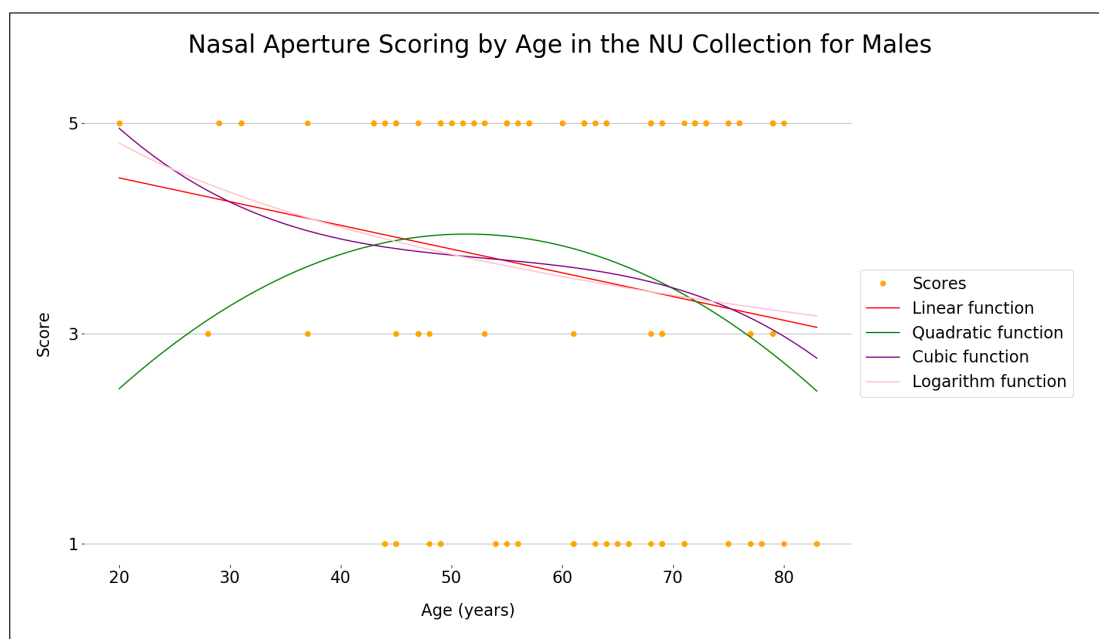

Figure C.24: A scatterplot of age vs. nasal aperture trait scoring for males in the SB collection, with four fitting functions.
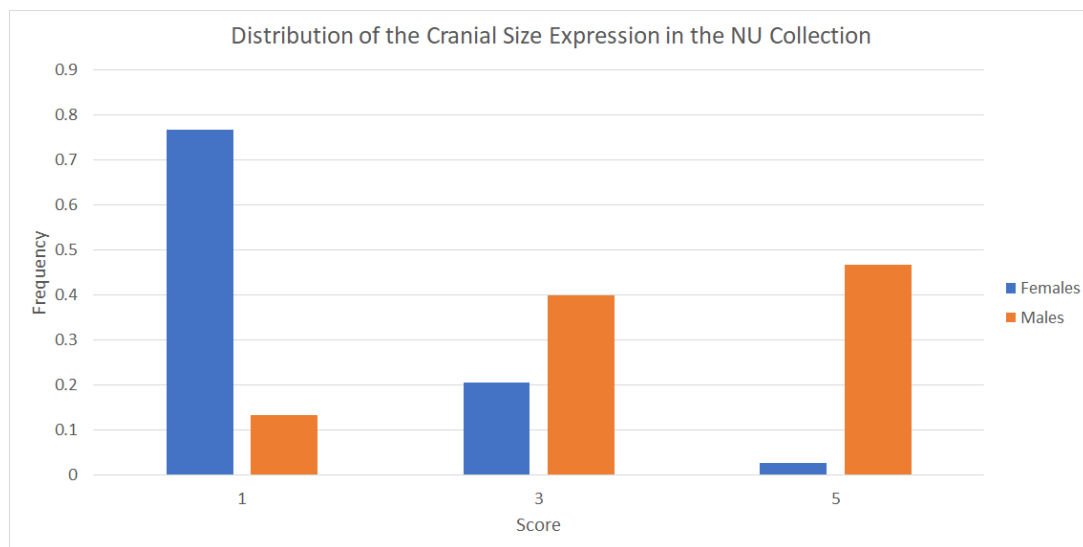
C.7 Cranial Size
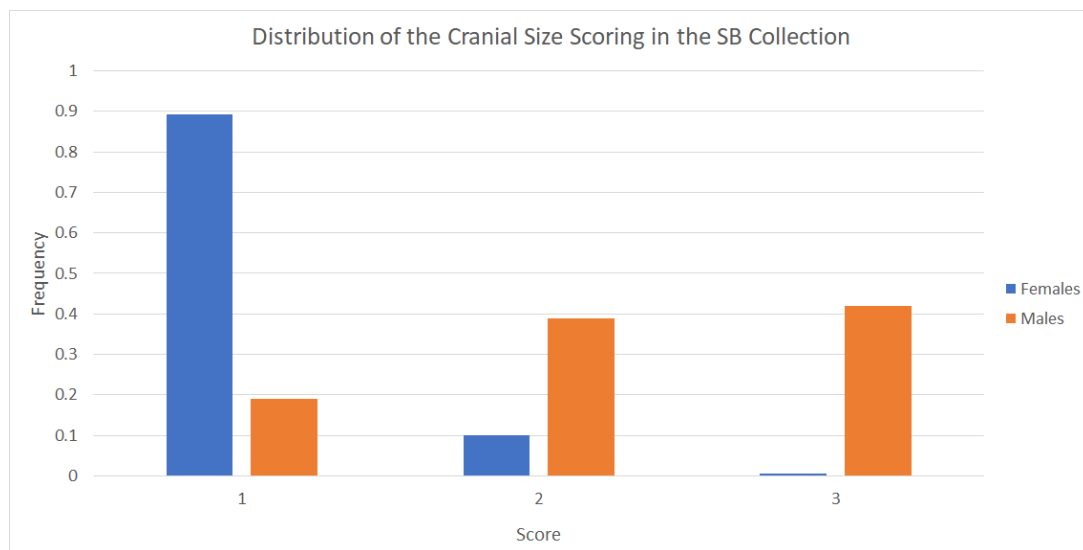
Figure C.25: The distribution of cranial size in the SB Collection represented using a bar chart. Females are in blue while males are in orange.
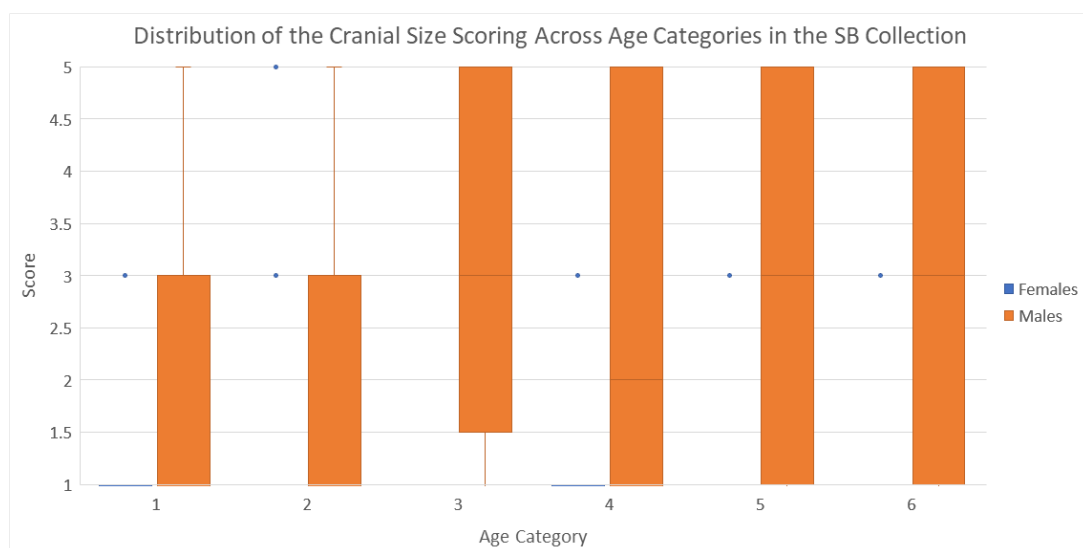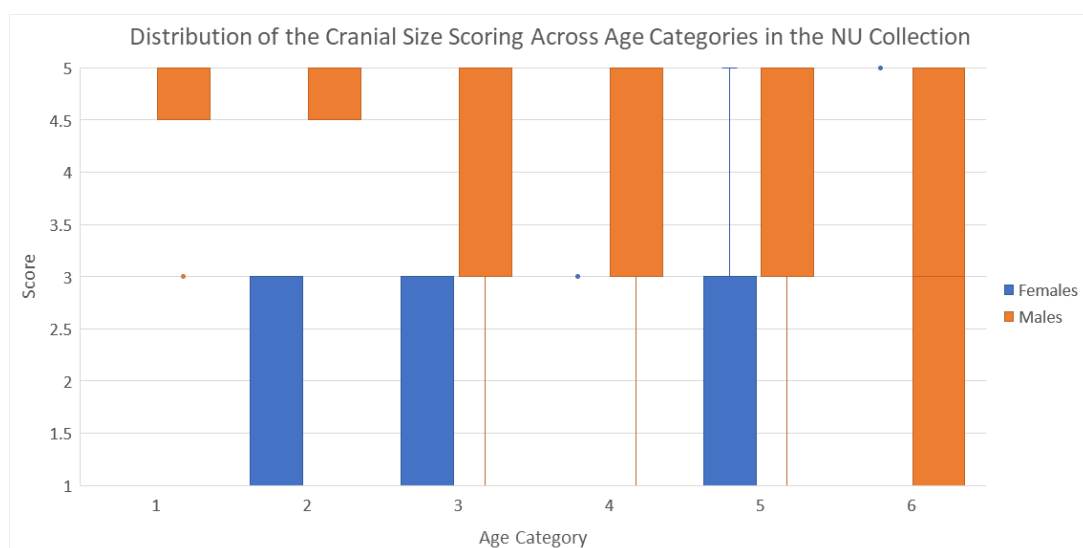

Figure C.26: A boxplot distribution of cranial size scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table C.7: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the SB collection when comparing cranial size scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 12 M = 14	F = 1.0 M = 3.0	$U = 115.5$ $p = 0.071$ $z = $ -2.39 $r = $ -0.47	0.589
2	F = 22 M = 16	F = 1.0 M = 3.0	$U = 272.0$ $p = 0.002$ $z = $ -4.64 $r = $ -0.75	0.705
3	F = 18 M = 16	F = 1.0 M = 3.0	$U = 279.0$ $p << 0.001$ $z = $ -1.24 $r = $ -0.21	0.938
4	F = 18 M = 20	F = 1.0 M = 5.0	$U = 316.0$ $p << 0.001$ $z = $ -1.02 $r = $ -0.17	0.800
5	F = 43 M = 40	F = 1.0 M = 3.0	$U = 1502.0$ $p << 0.001$ $z = $ -2.77 $r = $ -0.30	0.784
6	F = 36 M = 28	F = 1.0 M = 3.0	$U = 938.0$ $p << 0.001$ $z = $ -3.14 $r = $ -0.39	0.873

Figure C.27: A scatterplot of age vs. cranial size scoring for females in the SB collection, with four fitting functions.
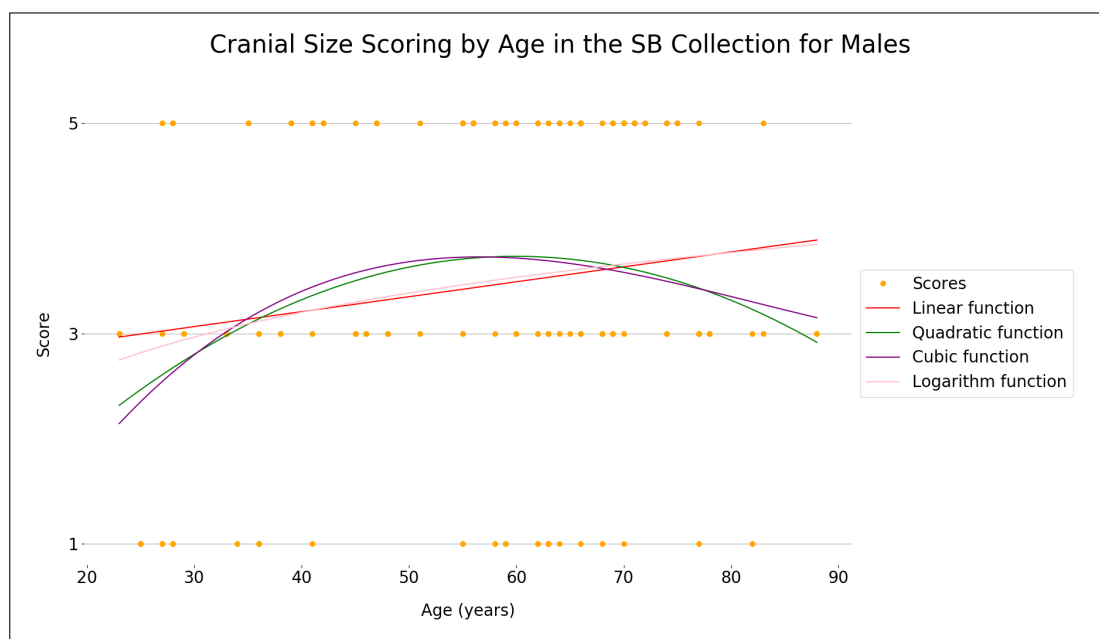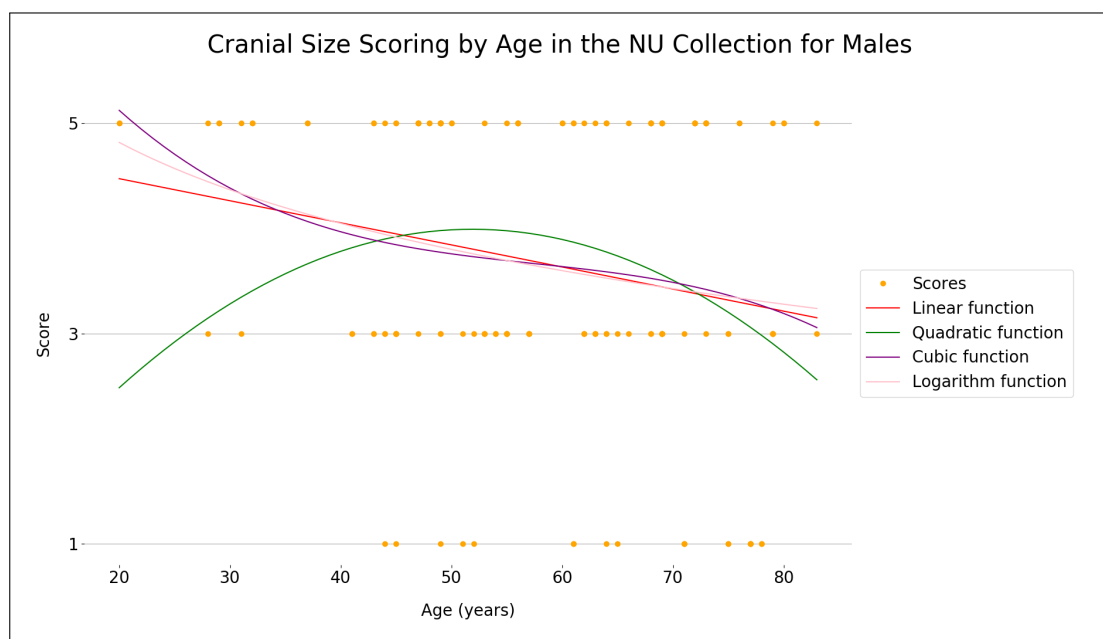

Figure C.28: A scatterplot of age vs. cranial size scoring for males in the SB collection, with four fitting functions.

APPENDIX D: Trait Distribution Graphs for the NU Collection
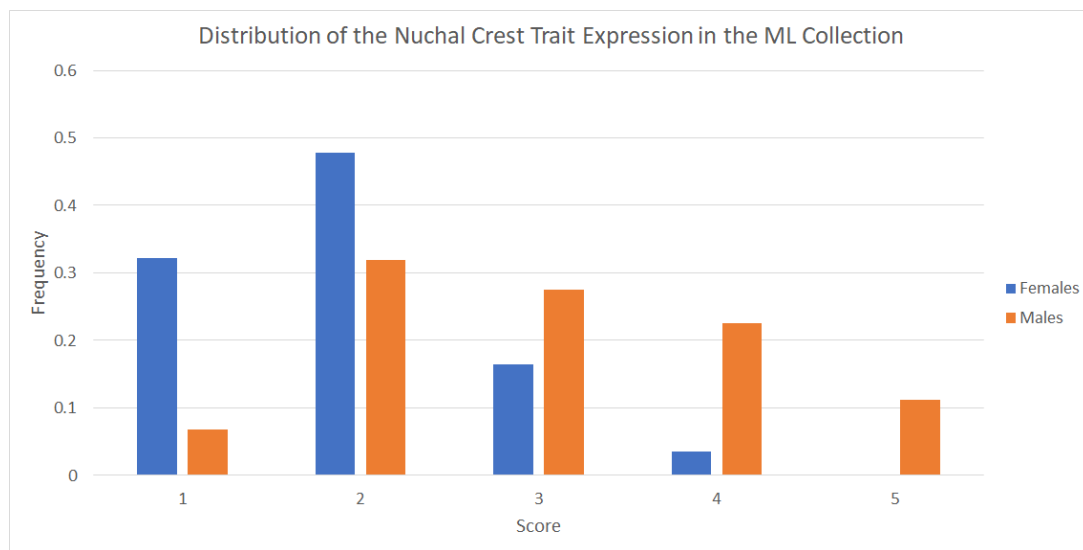
D.1   Nuchal Crest
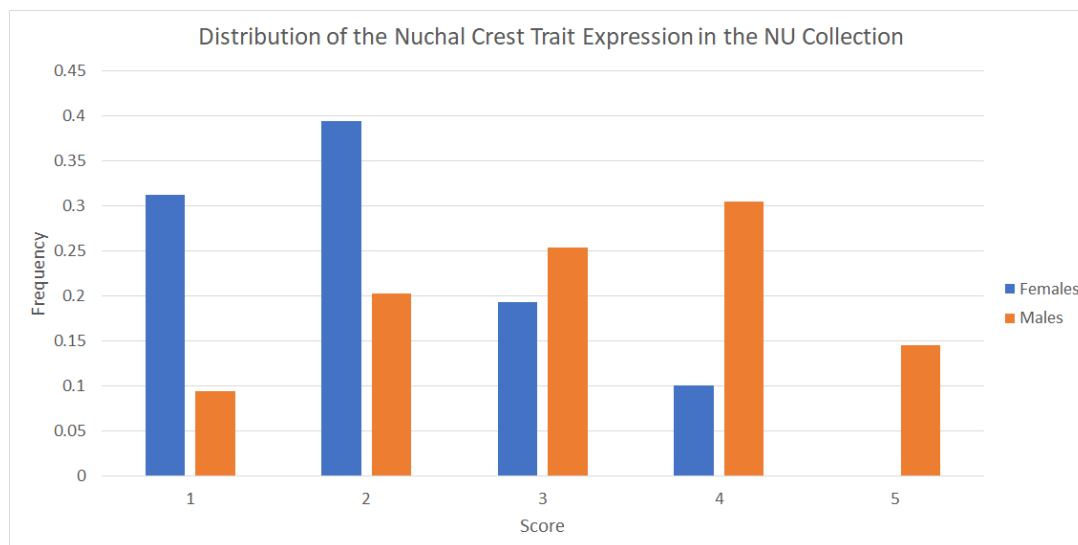
Figure D.1: The distribution of the nuchal crest trait expression in the NU Collection represented using a bar chart. Females are in blue while males are in orange.
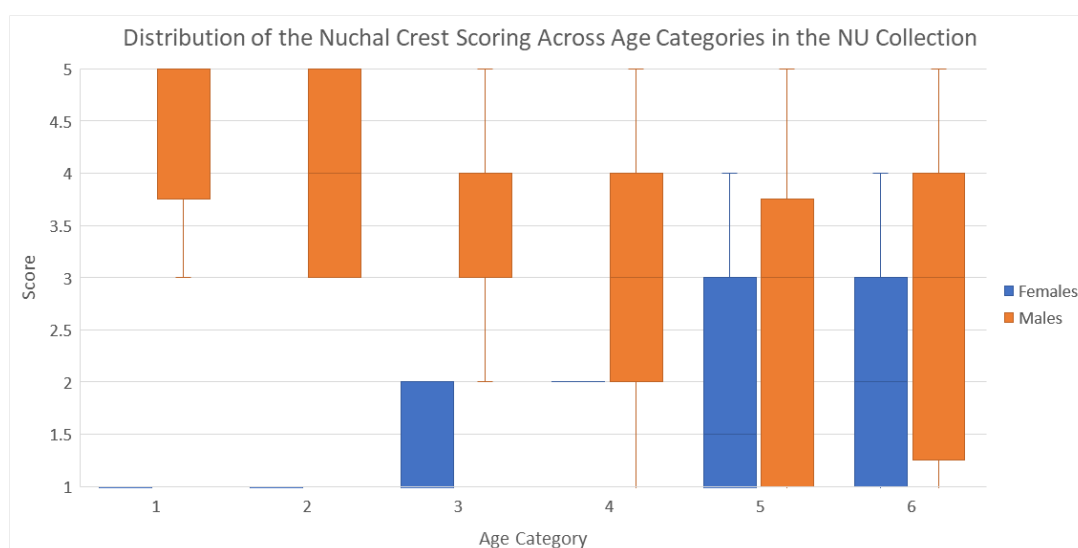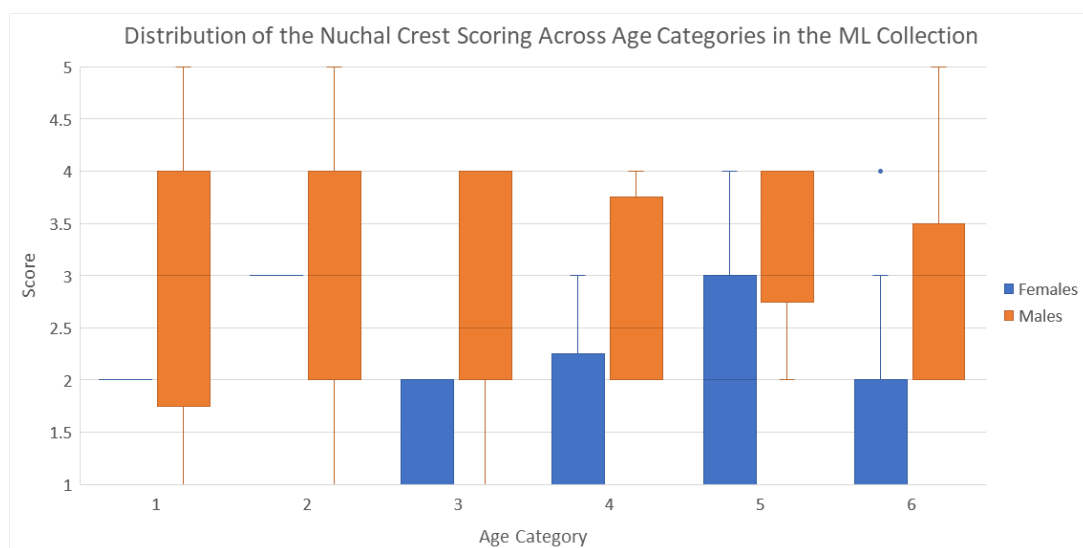

Figure D.2: A boxplot distribution of nuchal crest scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table D.1: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the NU collection when comparing nuchal crest trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F $= 2$ M $= 6$	F $= 1.0$ M $= 5.0$	$U = 12.0$ $p = 0.049$ $z = 1.00$ $r = 0.35$	1.000
2	F $= 2$ M $= 6$	F $= 1.0$ M $= 4.0$	$U = 12.0$ $p = 0.060$ $z = 1.00$ $r = 0.35$	1.000
3	F $= 6$ M $= 35$	F $= 1.5$ M $= 4.0$	$U = 201.0$ $p < 0.001$ $z = 2.77$ $r = 0.43$	0.914
4	F $= 8$ M $= 24$	F $= 2.0$ M $= 3.0$	$U = 152.0$ $p = 0.010$ $z = 0.87$ $r = 0.15$	0.750
5	F $= 27$ M $= 33$	F $= 2.0$ M $= 3.0$	$U = 594.0$ $p = 0.023$ $z = -3.41$ $r = -0.44$	0.782
6	F $= 64$ M $= 34$	F $= 2.0$ M $= 3.0$	$U = 1379.0$ $p = 0.025$ $z = -13.35$ $r = -1.35$	0.770
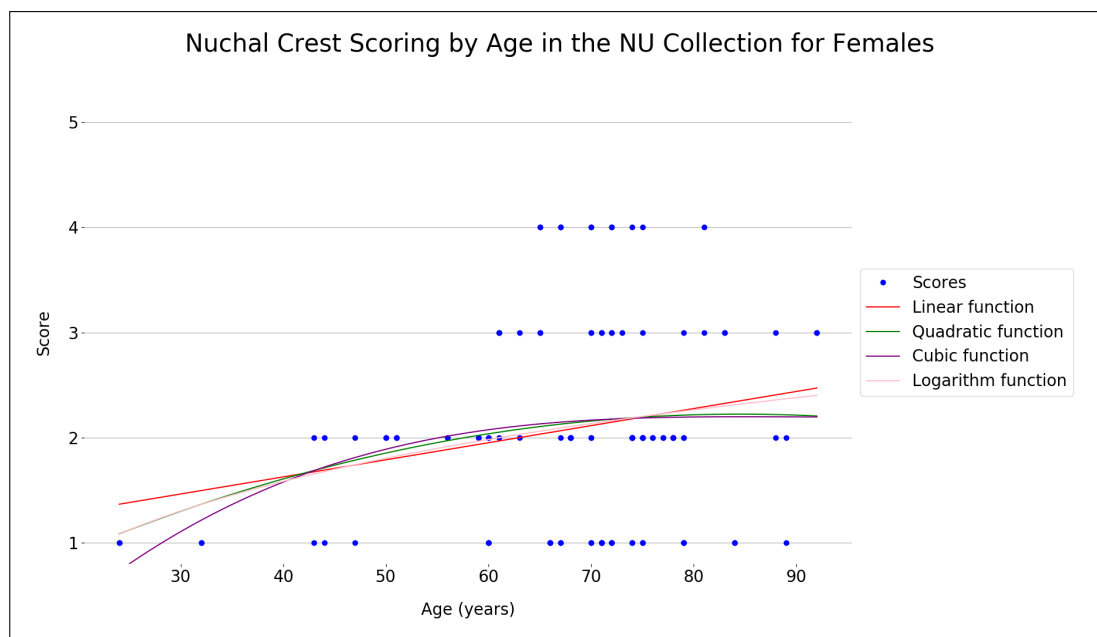
Figure D.3: A scatterplot of age vs. nuchal crest trait scoring for females in the NU collection, with four fitting functions.
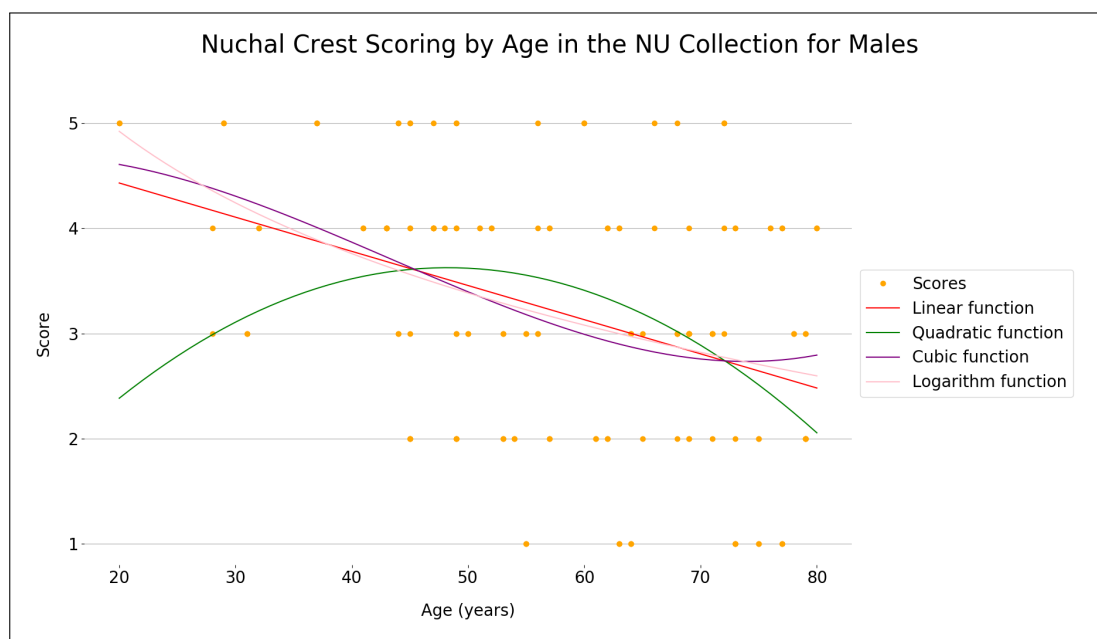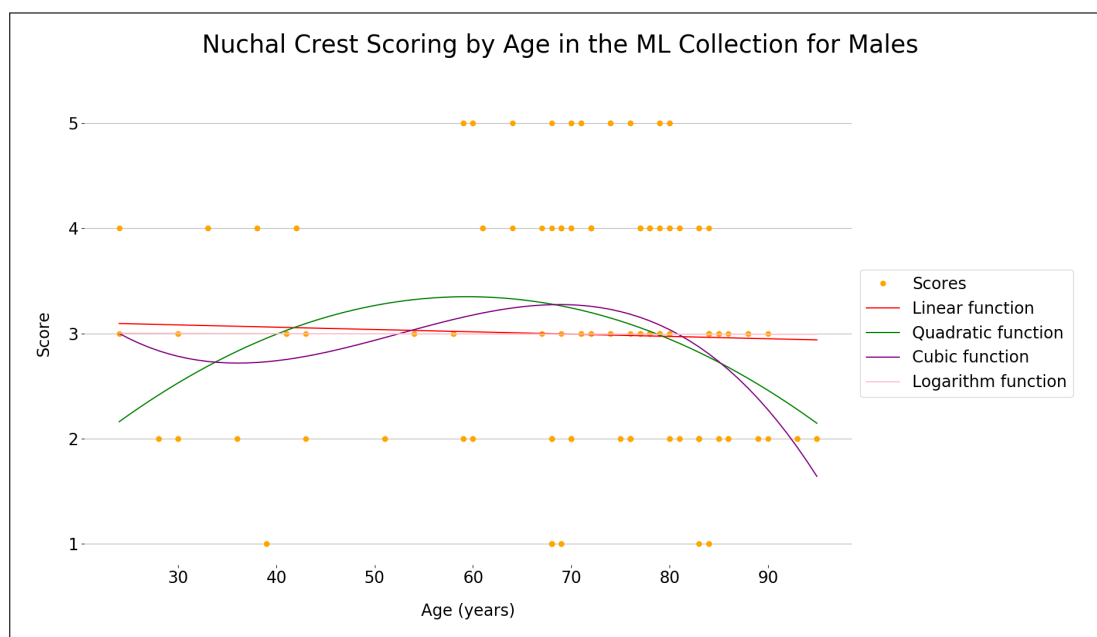

Figure D.4: A scatterplot of age vs. nuchal crest trait scoring for males in the NU collection, with four fitting functions.
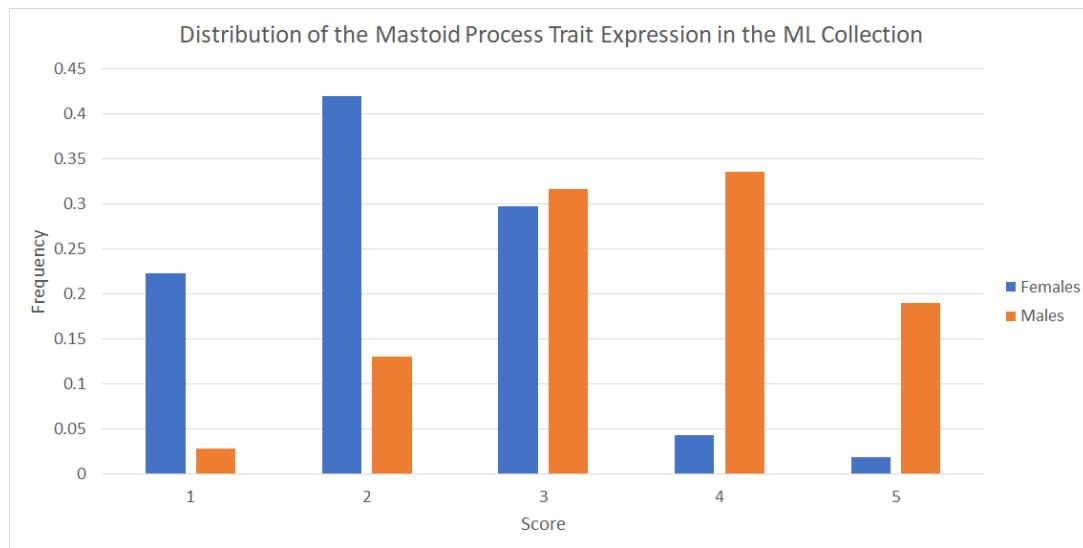
D.2 Mastoid Process
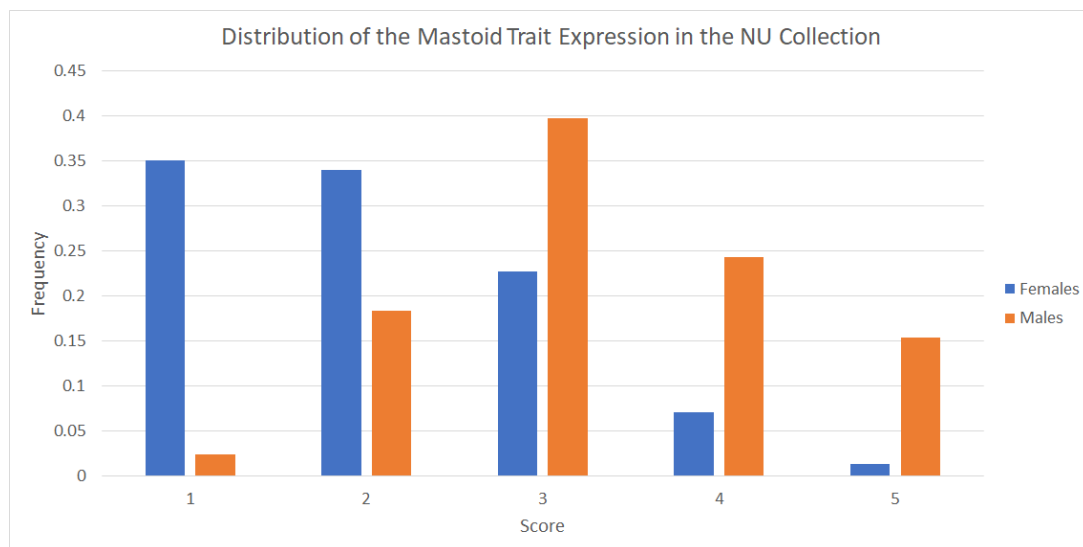
Figure D.5: The distribution of the mastoid process trait expression in the NU Collection represented using a bar chart. Females are in blue while males are in orange.
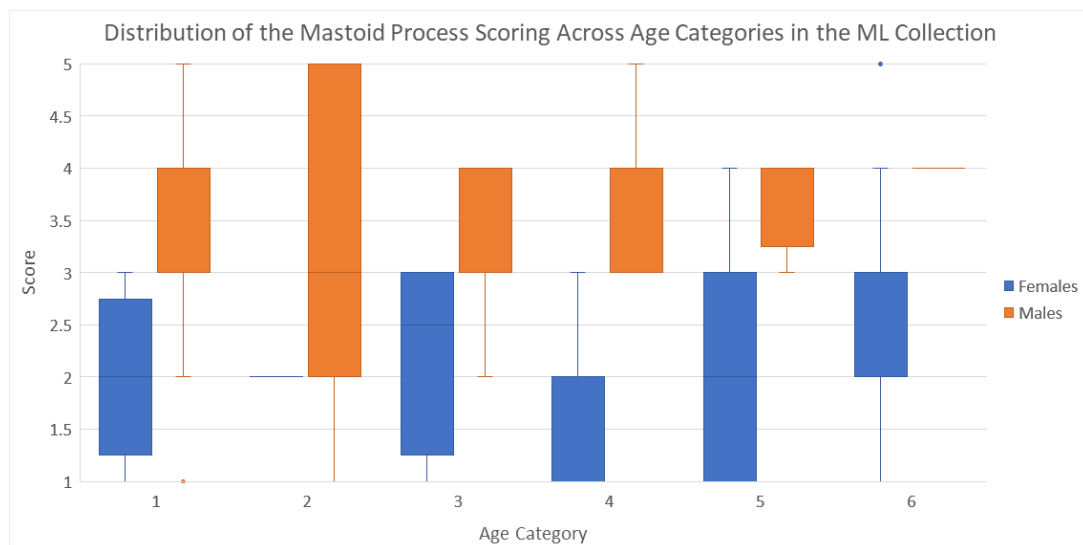

Figure D.6: A boxplot distribution of mastoid process scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table D.2: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the NU collection when comparing mastoid process trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 12	F = 1.0 M = 3.0	$U = 96.0$ $p << 0.001$ $z = 0.93$ $r = 0.21$	1.000
2	F = 12 M = 12	F = 2.0 M = 3.0	$U = 94.5$ $p = 0.186$ $z = -3.20$ $r = -0.65$	0.785
3	F = 20 M = 72	F = 2.0 M = 3.0	$U = 1091.5$ $p < 0.001$ $z = 1.53$ $r = 0.16$	0.794
4	F = 16 M = 52	F = 1.0 M = 3.0	$U = 807.5$ $p << 0.001$ $z = 3.69$ $r = 0.45$	0.941
5	F = 76 M = 80	F = 2.0 M = 3.0	$U = 4910.5$ $p << 0.001$ $z = -3.74$ $r = -0.30$	0.831
6	F = 168 M = 72	F = 2.0 M = 3.0	$U = 9232.0$ $p << 0.001$ $z = -22.34$ $r = -1.44$	0.795
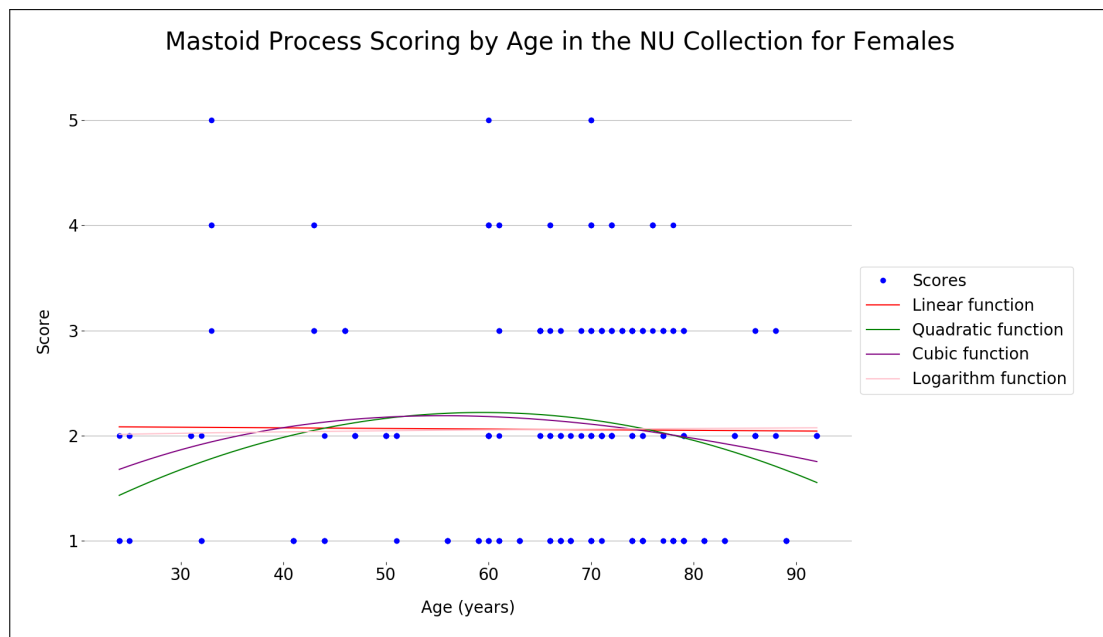
Figure D.7: A scatterplot of age vs. mastoid process trait scoring for females in the NU collection, with four fitting functions.
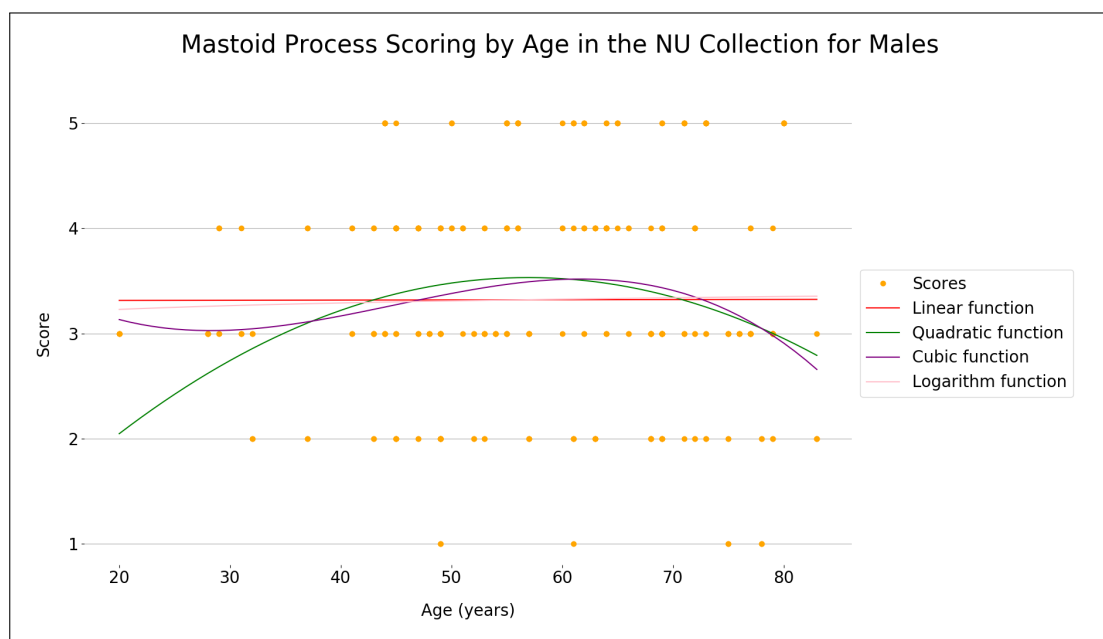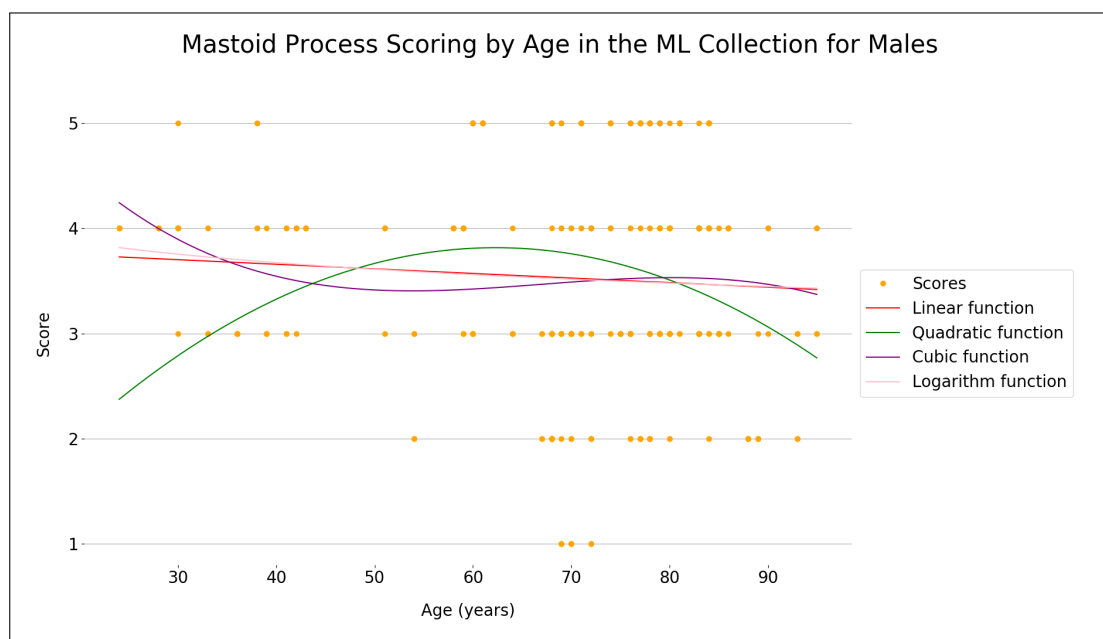

Figure D.8: A scatterplot of age vs. mastoid process trait scoring for males in the NU collection, with four fitting functions.
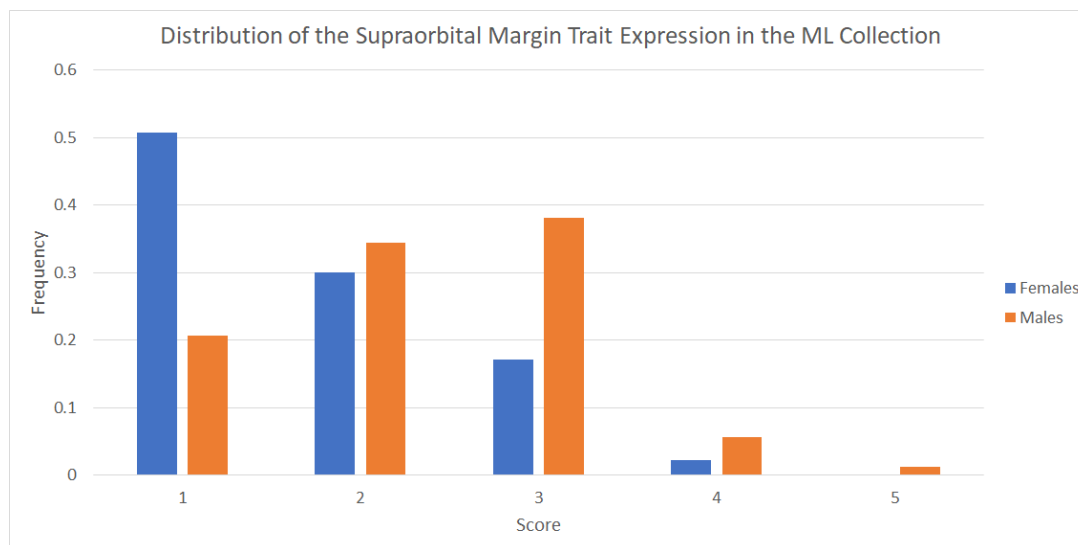
D.3 Supraorbital Margin

Figure D.9: The distribution of the supraorbital margin trait expression in the NU Collection represented using a bar chart. Females are in blue while males are in orange.
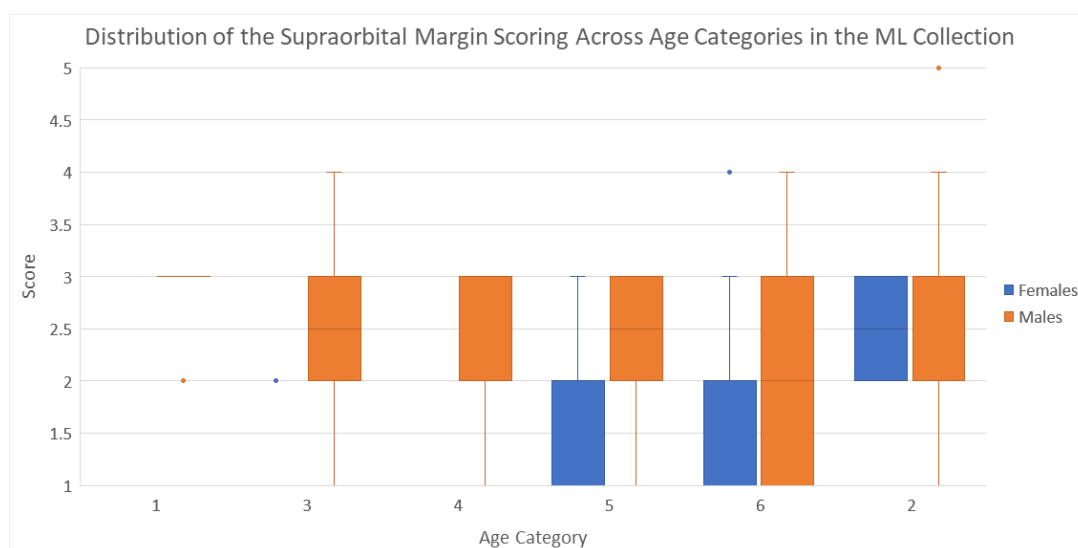

Figure D.10: A boxplot distribution of supraorbital margin scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table D.3: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the NU collection when comparing supraorbital margin trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 12	F = 1.0 M = 3.0	$U = 96.0$ $p << 0.001$ $z = 0.93$ $r = 0.21$	1.000
2	F = 10 M = 12	F = 2.0 M = 3.0	$U = 113.0$ $p < 0.001$ $z = -0.13$ $r = -0.03$	0.883
3	F = 20 M = 72	F = 1.0 M = 2.0	$U = 1222.5$ $p << 0.001$ $z = 2.77$ $r = 0.29$	0.774
4	F = 16 M = 52	F = 2.0 M = 2.0	$U = 466.0$ $p = 0.446$ $z = -1.24$ $r = -0.15$	0.668
5	F = 76 M = 80	F = 2.0 M = 2.0	$U = 4531.5$ $p << 0.001$ $z = -5.09$ $r = -0.41$	0.727
6	F = 166 M = 72	F = 2.0 M = 2.0	$U = 8025.0$ $p << 0.001$ $z = -24.21$ $r = -1.57$	0.682
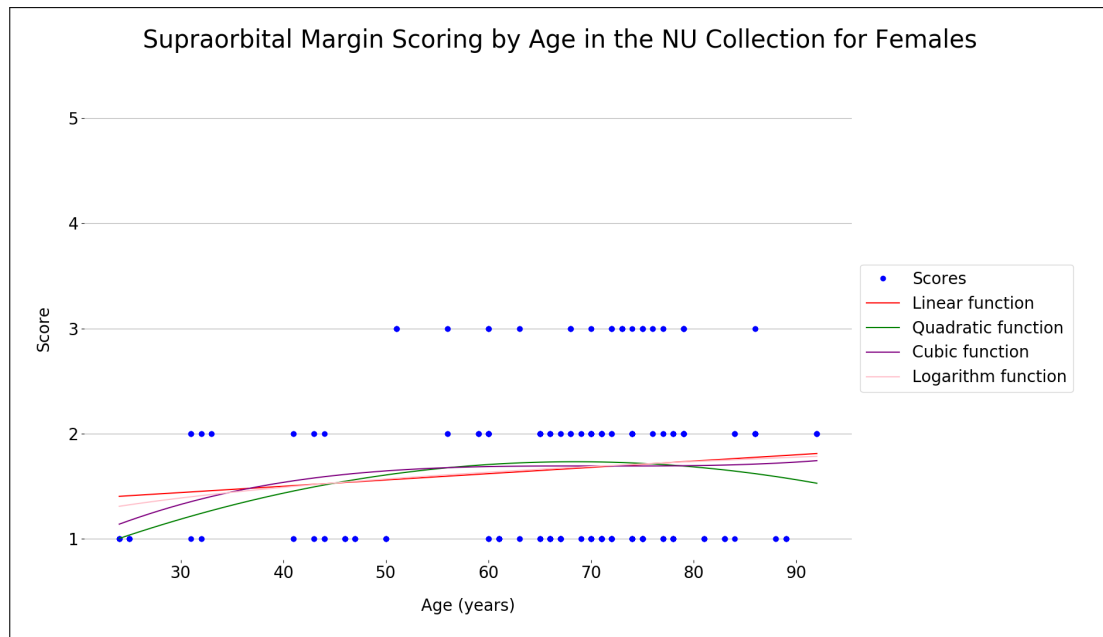
Figure D.11: A scatterplot of age vs. supraorbital margin trait scoring for females in the NU collection, with four fitting functions.
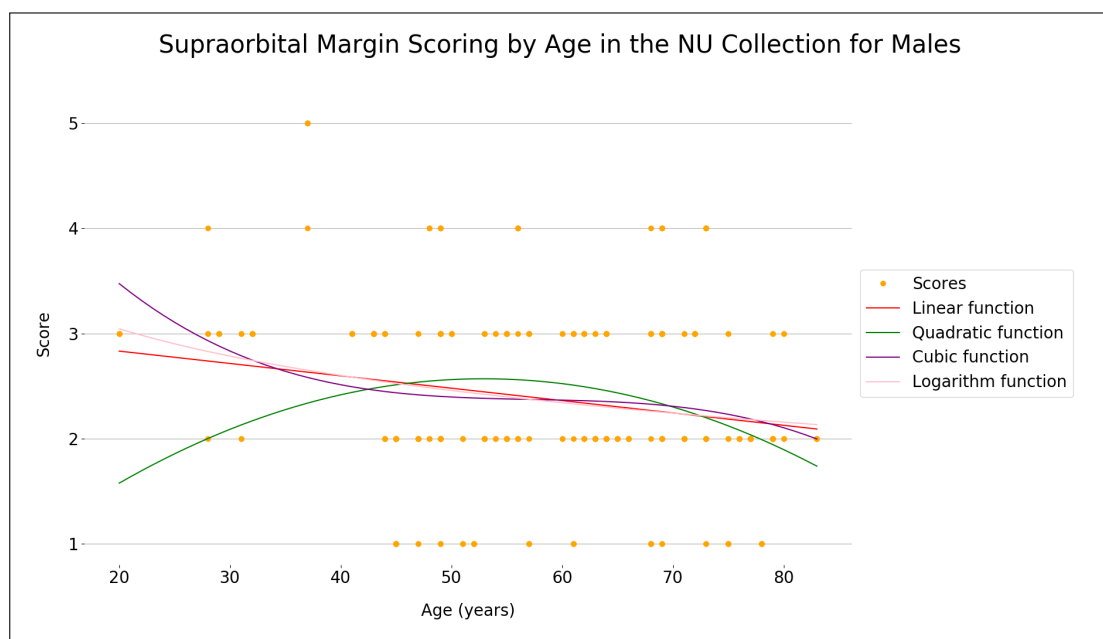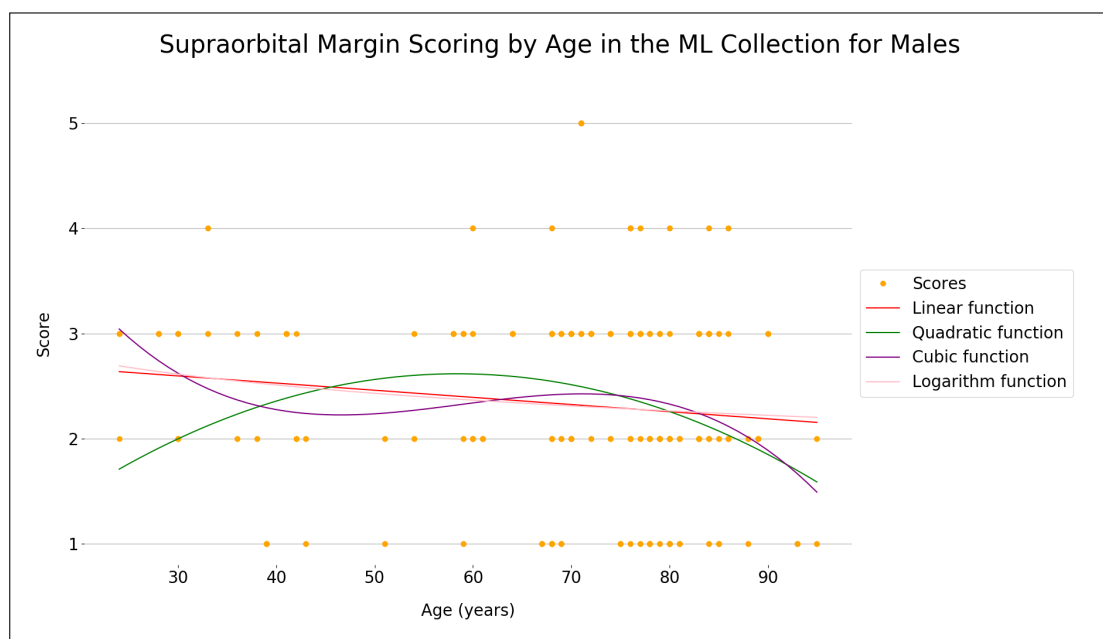

Figure D.12: A scatterplot of age vs. supraorbital margin trait scoring for males in the NU collection, with four fitting functions.
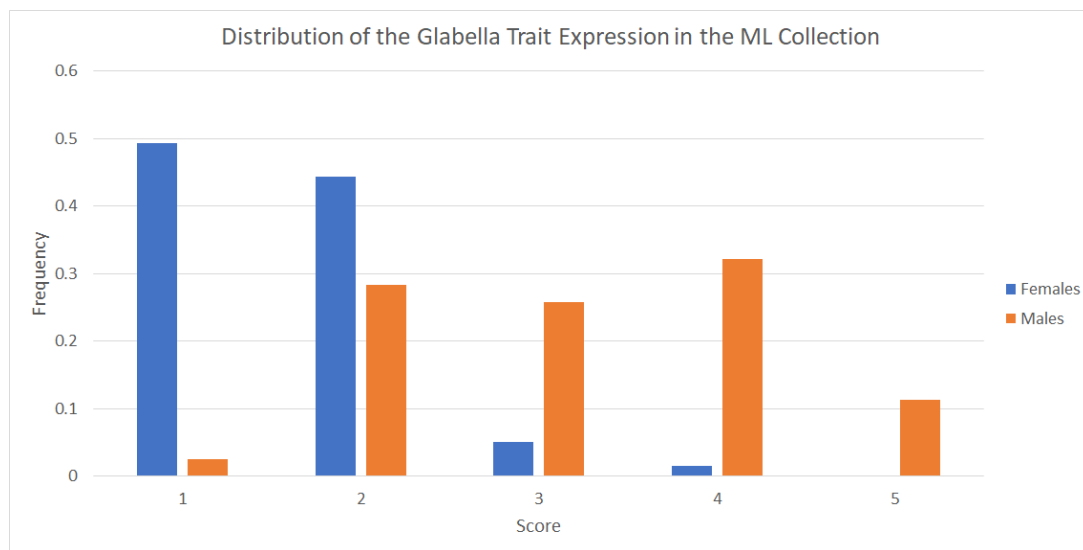
D.4 Glabella

Figure D.13: The distribution of the glabella trait expression in the NU Collection represented using a bar chart. Females are in blue while males are in orange.
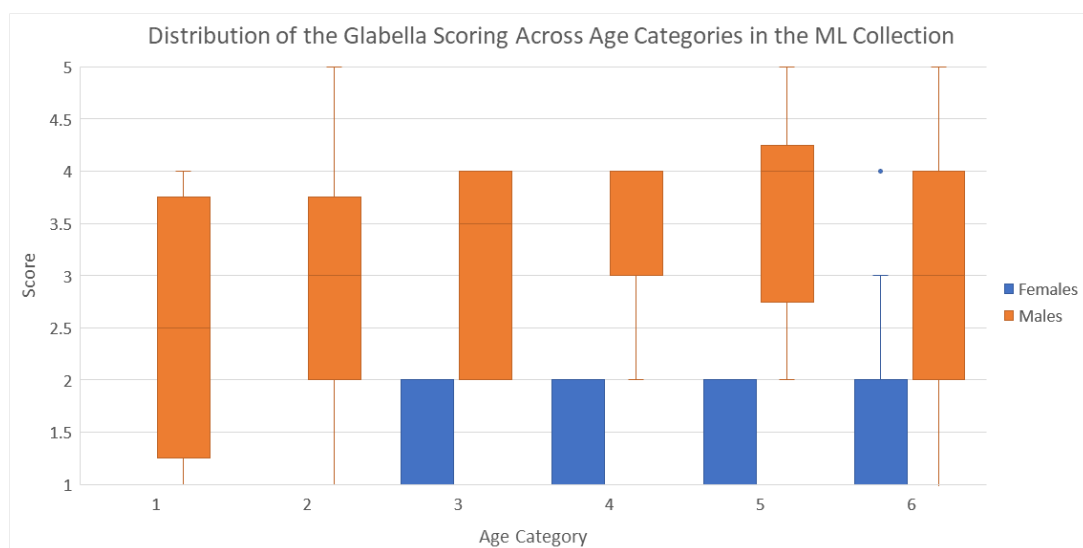

Figure D.14: A boxplot distribution of glabella scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table D.4: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the NU collection when comparing glabella trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 6	F = 1.0 M = 2.0	$U = 22.0$ $p = 0.028$ $z = 0.00$ $r = 0.00$	0.833
2	F = 5 M = 6	F = 1.0 M = 3.5	$U = 30.0$ $p = 0.005$ $z = 0.00$ $r = 0.00$	1.000
3	F = 10 M = 36	F = 1.0 M = 3.0	$U = 344.0$ $p << 0.001$ $z = 2.90$ $r = 0.43$	0.922
4	F = 8 M = 26	F = 1.0 M = 2.0	$U = 170.0$ $p = 0.006$ $z = 1.22$ $r = 0.21$	0.750
5	F = 36 M = 38	F = 2.0 M = 3.0	$U = 1185.5$ $p << 0.001$ $z = -1.78$ $r = -0.21$	0.783
6	F = 76 M = 30	F = 1.0 M = 2.0	$U = 1897.0$ $p << 0.001$ $z = -15.21$ $r = -1.48$	0.738
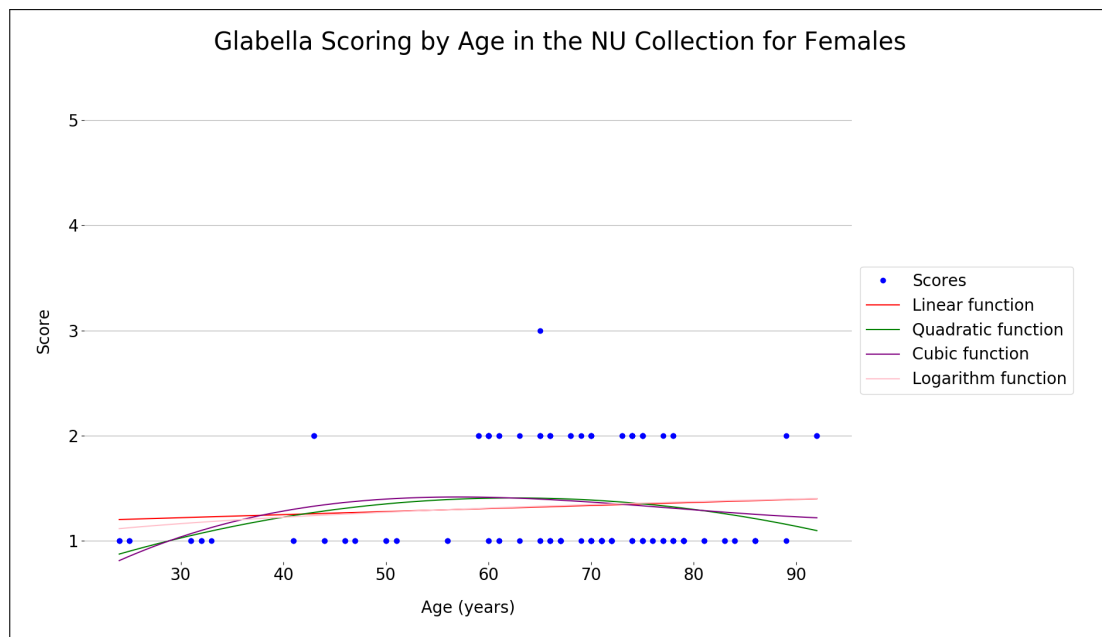
Figure D.15: A scatterplot of age vs. glabella trait scoring for females in the NU collection, with four fitting functions.
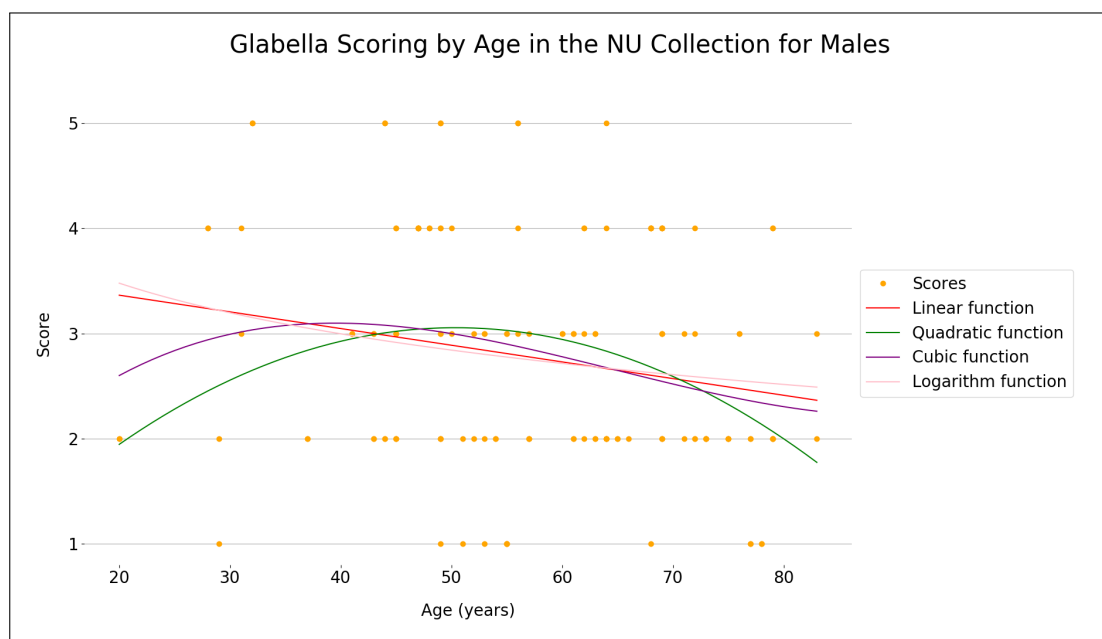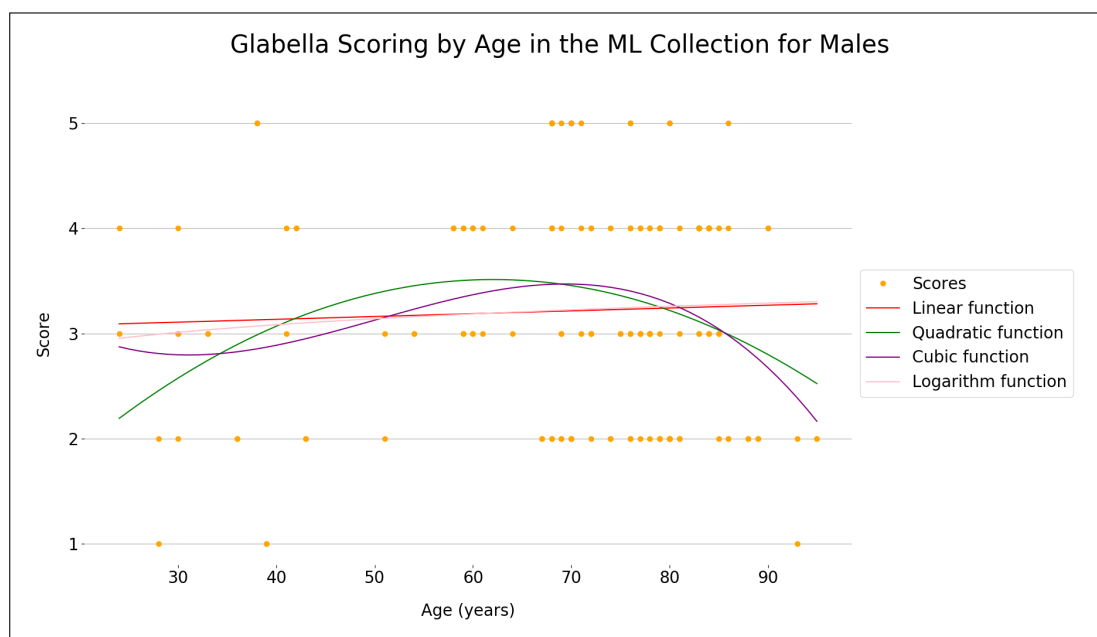

Figure D.16: A scatterplot of age vs. glabella trait scoring for males in the NU collection, with four fitting functions.

D.5   Zygomatic Extension

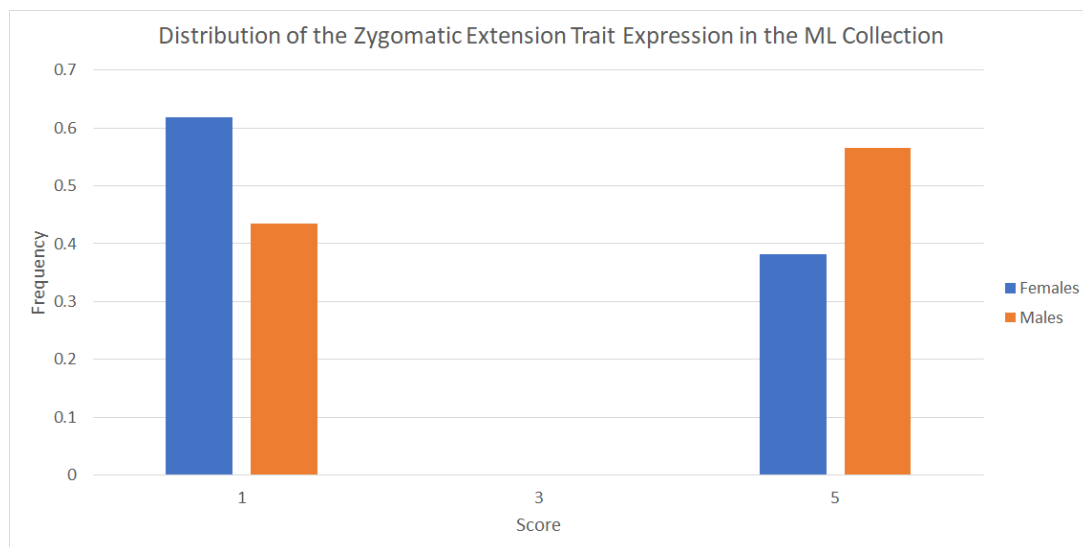**Figure D.17:** The distribution of the zygomatic extension trait expression in the NU Collection represented using a bar chart. Females are in blue while males are in orange.
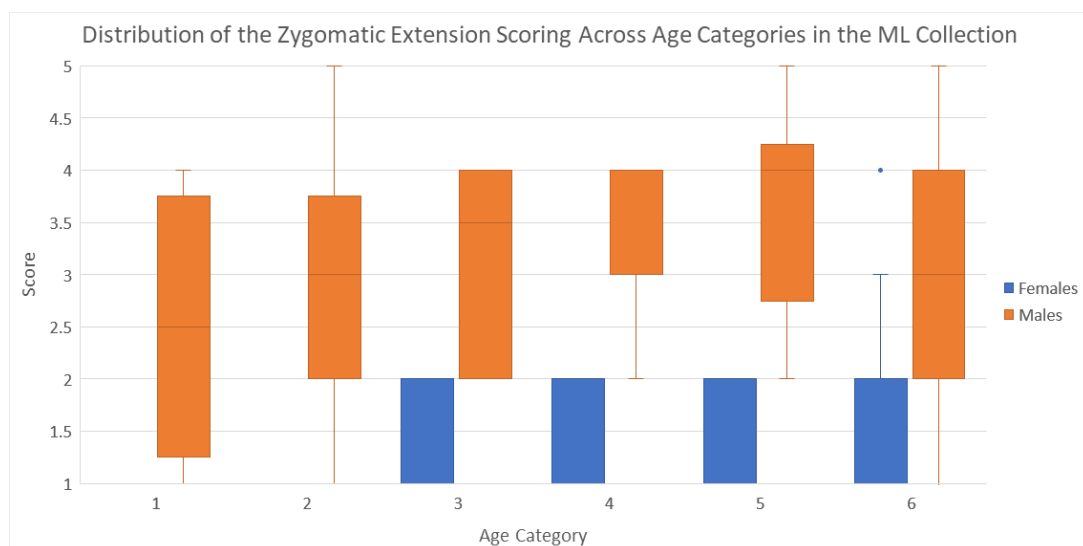

Figure D.18: A boxplot distribution of zygomatic extension scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table D.5: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the NU collection when comparing zygomatic extension trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F $= 8$ M $= 12$	F $= 3.0$ M $= 5.0$	$U = 72.0$ $p = 0.009$ $z = $ -0.93 $r = $ -0.21	0.500
2	F $= 12$ M $= 12$	F $= 1.0$ M $= 5.0$	$U = 120.0$ $p < 0.001$ $z = $ -1.73 $r = $ -0.35	0.667
3	F $= 20$ M $= 72$	F $= 1.0$ M $= 5.0$	$U = 1076.0$ $p << 0.001$ $z = 1.38$ $r = 0.14$	0.617
4	F $= 16$ M $= 52$	F $= 5.0$ M $= 5.0$	$U = 470.0$ $p = 0.345$ $z = $ -1.19 $r = $ -0.14	0.476
5	F $= 76$ M $= 80$	F $= 5.0$ M $= 5.0$	$U = 3820.0$ $p < 0.001$ $z = $ -7.61 $r = $ -0.61	0.411
6	F $= 168$ M $= 72$	F $= 5.0$ M $= 5.0$	$U = 6756.0$ $p = 0.081$ $z = $ -27.37 $r = $ -1.77	0.444
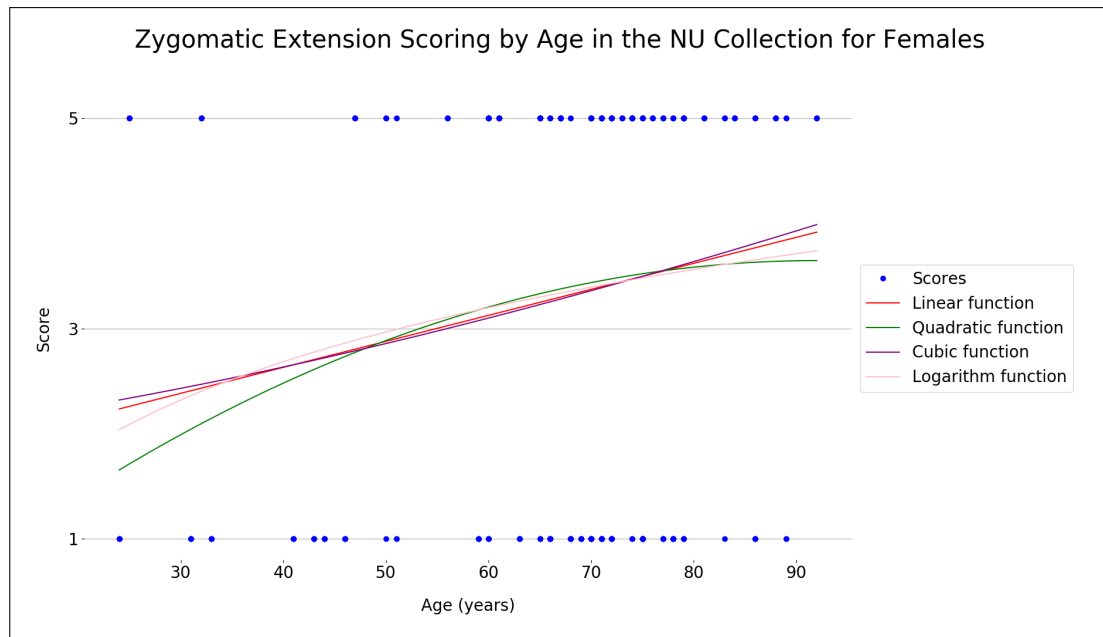
Figure D.19: A scatterplot of age vs. zygomatic extension trait scoring for females in the NU collection, with four fitting functions.
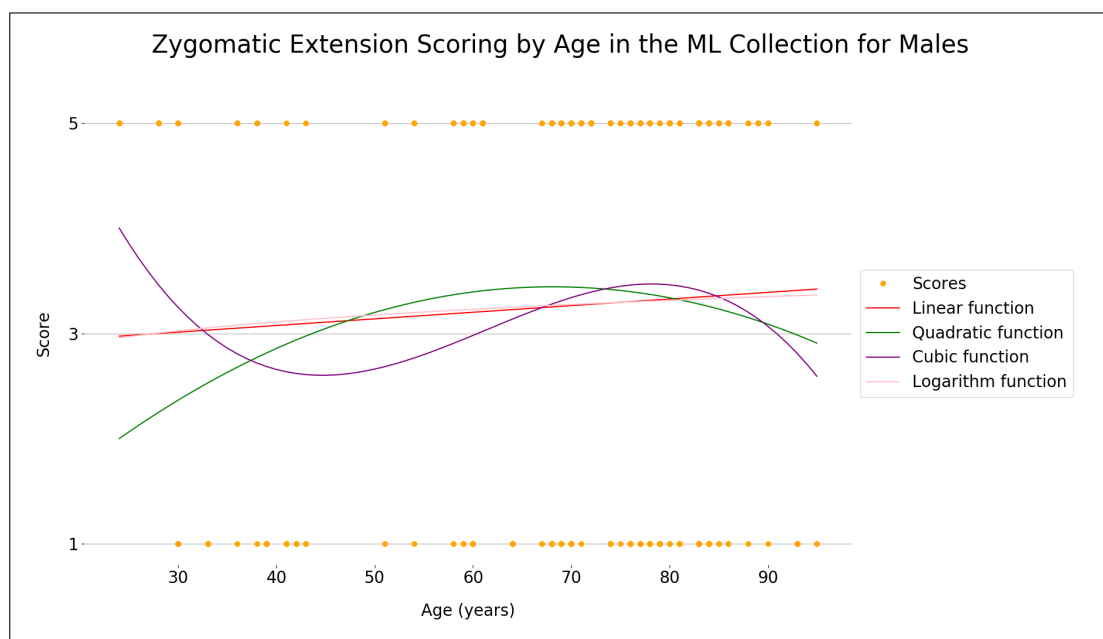

Figure D.20: A scatterplot of age vs. zygomatic extension trait scoring for males in the NU collection, with four fitting functions.

D.6   Nasal Aperture
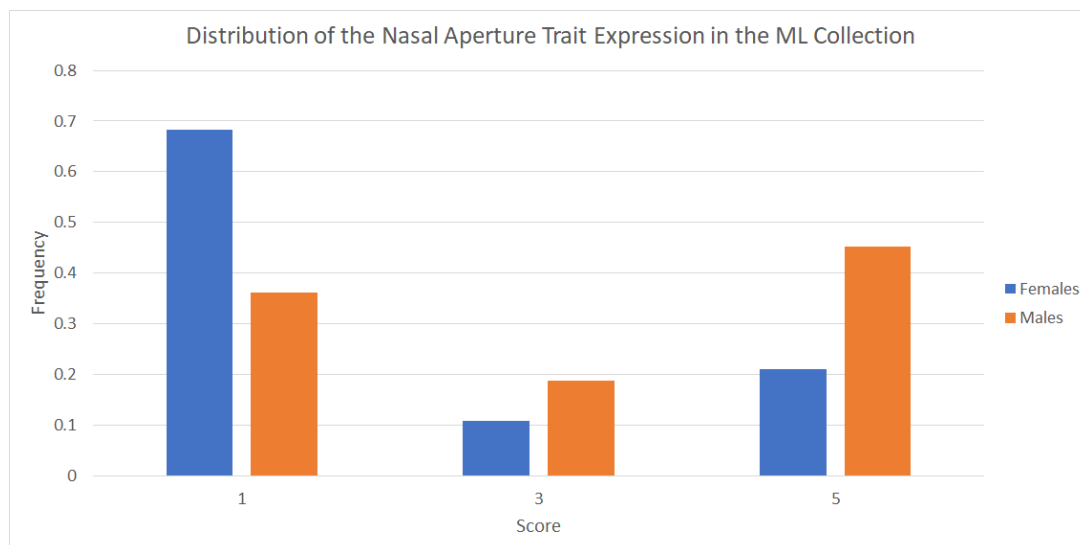

Figure D.21: The distribution of the nasal aperture trait expression in the NU Collection represented using a bar chart. Females are in blue while males are in orange.
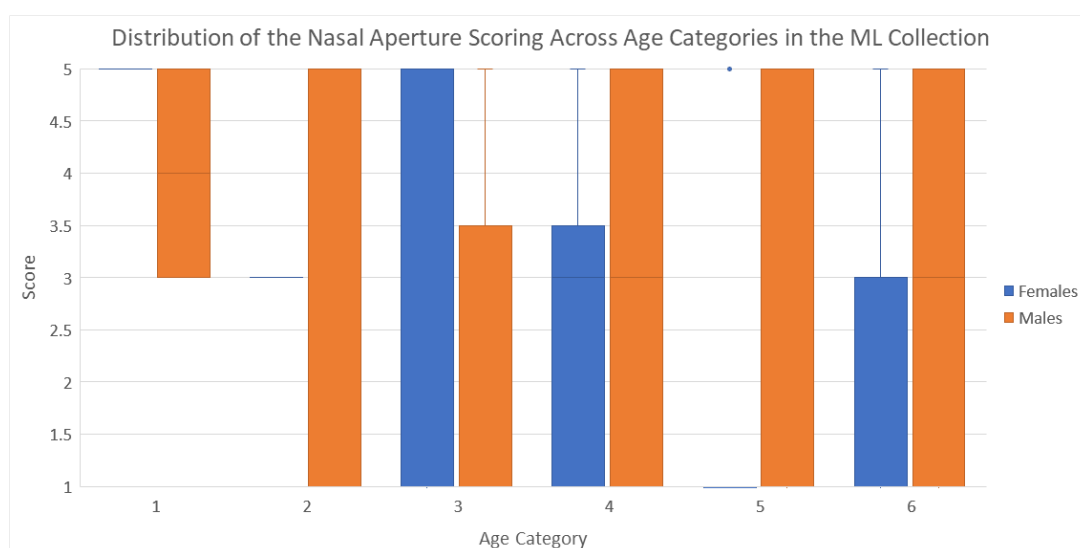

Figure D.22: A boxplot distribution of nasal aperture scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table D.6: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the NU collection when comparing nasal aperture trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 6	F = 5.0 M = 5.0	$U = 11.0$ $p = 0.894$ $z = -2.35$ $r = -0.74$	0.417
2	F = 6 M = 4	F = 3.0 M = 5.0	$U = 18.0$ $p = 0.203$ $z = -3.20$ $r = -1.01$	0.667
3	F = 10 M = 32	F = 1.0 M = 5.0	$U = 214.0$ $p = 0.086$ $z = -0.03$ $r = -0.00$	0.669
4	F = 8 M = 26	F = 2.0 M = 5.0	$U = 144.5$ $p = 0.048$ $z = 0.18$ $r = 0.03$	0.611
5	F = 36 M = 40	F = 3.0 M = 3.0	$U = 870.5$ $p = 0.094$ $z = -5.36$ $r = -0.62$	0.672
6	F = 80 M = 36	F = 1.0 M = 5.0	$U = 1960.0$ $p < 0.001$ $z = -16.23$ $r = -1.51$	0.644
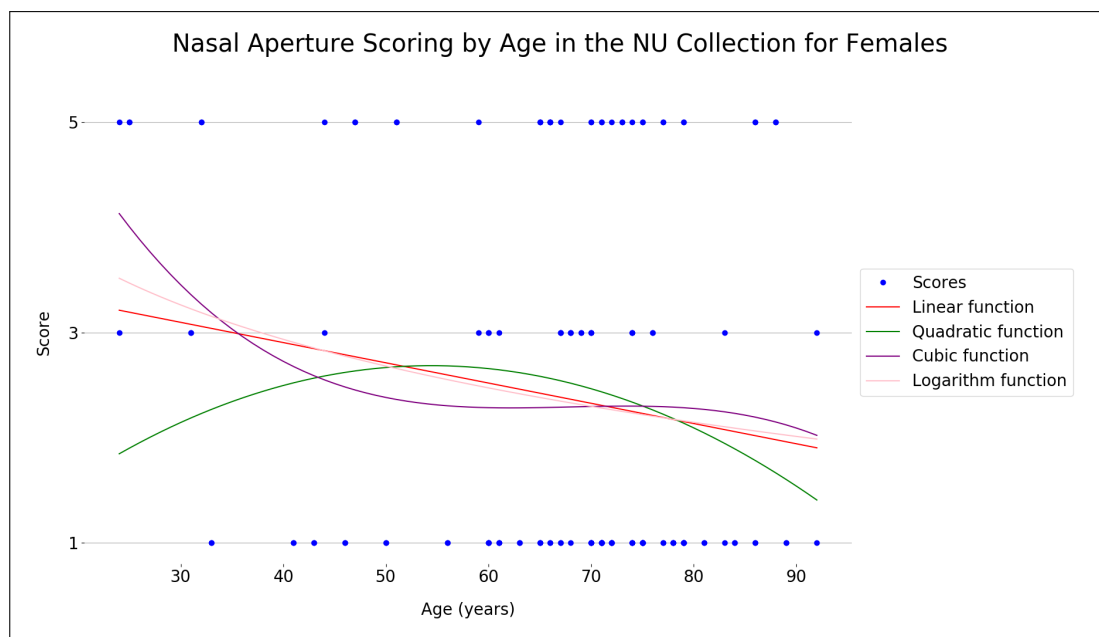
Figure D.23: A scatterplot of age vs. nasal aperture trait scoring for females in the NU collection, with four fitting functions.

Figure D.24: A scatterplot of age vs. nasal aperture trait scoring for males in the NU collection, with four fitting functions.

D.7 Cranial Size

Figure D.25: The distribution of cranial size in the NU Collection represented using a bar chart. Females are in blue while males are in orange.

Figure D.26: A boxplot distribution of cranial size scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table D.7: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the NU collection when comparing cranial size scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

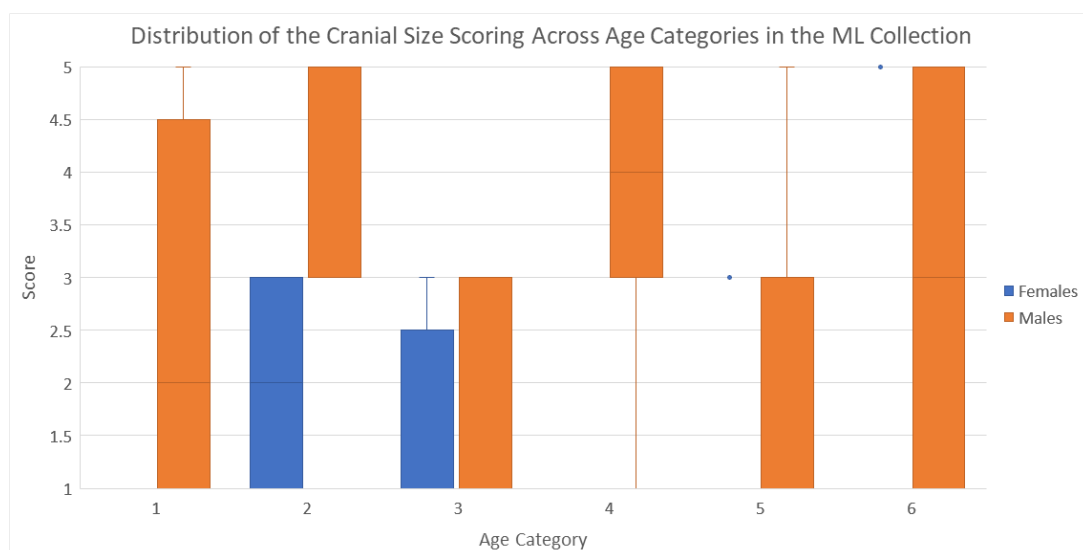| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 6	F = 1.0 M = 5.0	$U = 24.0$ $p = 0.007$ $z = 0.43$ $r = 0.13$	1.000
2	F = 6 M = 6	F = 1.0 M = 5.0	$U = 35.0$ $p = 0.005$ $z = $ -0.64 $r = $ -0.18	0.944
3	F = 10 M = 36	F = 1.0 M = 5.0	$U = 311.0$ $p < 0.001$ $z = 2.02$ $r = 0.30$	0.794
4	F = 8 M = 26	F = 1.0 M = 3.0	$U = 191.0$ $p < 0.001$ $z = 2.07$ $r = 0.36$	0.856
5	F = 38 M = 40	F = 1.0 M = 3.0	$U = 1298.0$ $p << 0.001$ $z = $ -2.03 $r = $ -0.23	0.808
6	F = 84 M = 36	F = 1.0 M = 3.0	$U = 2348.0$ $p << 0.001$ $z = $ -15.66 $r = $ -1.43	0.697
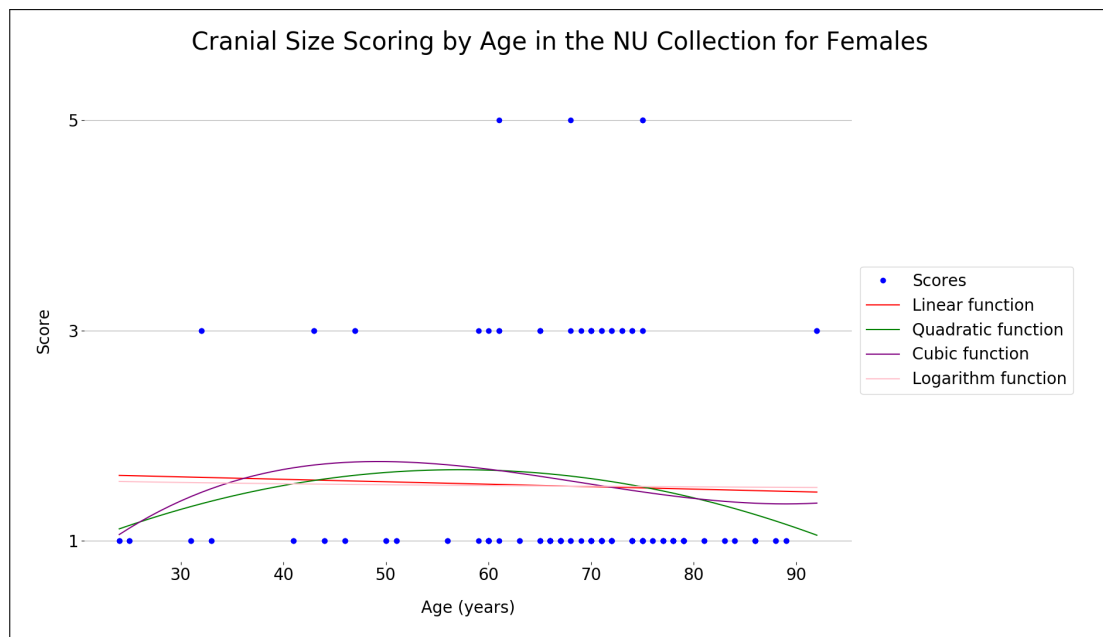
Figure D.27: A scatterplot of age vs. cranial size scoring for females in the NU collection, with four fitting functions.

Figure D.28: A scatterplot of age vs. cranial size scoring for males in the NU collection, with four fitting functions.

APPENDIX E: Trait Distribution Graphs for the ML Collection

E.1 Nuchal Crest

Figure E.1: The distribution of the nuchal crest trait expression in the ML Collection represented using a bar chart. Females are in blue while males are in orange.

Figure E.2: A boxplot distribution of nuchal crest scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table E.1: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the ML collection when comparing nuchal crest trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 2 M = 4	F = 2.0 M = 2.5	$U = 6.0$ $p = 0.411$ $z = -0.46$ $r = -0.19$	0.500
2	F = 2 M = 12	F = 3.0 M = 2.5	$U = 10.0$ $p = 0.777$ $z = -0.91$ $r = -0.24$	0.833
3	F = 8 M = 6	F = 2.0 M = 3.0	$U = 45.5$ $p = 0.004$ $z = -1.87$ $r = -0.50$	0.896
4	F = 6 M = 10	F = 1.0 M = 3.0	$U = 52.0$ $p = 0.015$ $z = 0.11$ $r = 0.03$	0.867
5	F = 16 M = 26	F = 2.0 M = 3.0	$U = 269.0$ $p = 0.107$ $z = -1.94$ $r = -0.30$	0.784
6	F = 106 M = 102	F = 2.0 M = 3.0	$U = 8604.0$ $p \ll 0.001$ $z = -5.70$ $r = -0.40$	0.764
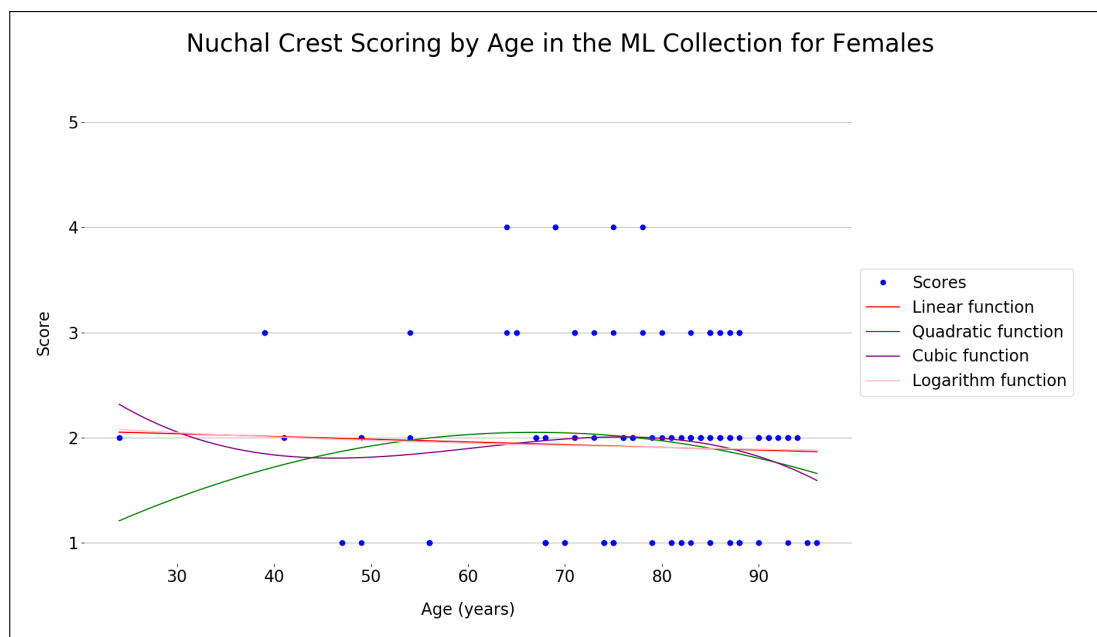
Figure E.3: A scatterplot of age vs. nuchal crest trait scoring for females in the ML collection, with four fitting functions.


Figure E.4: A scatterplot of age vs. nuchal crest trait scoring for males in the ML collection, with four fitting functions.

E.2   Mastoid Process

Figure E.5: The distribution of the mastoid process trait expression in the ML Collection represented using a bar chart. Females are in blue while males are in orange.
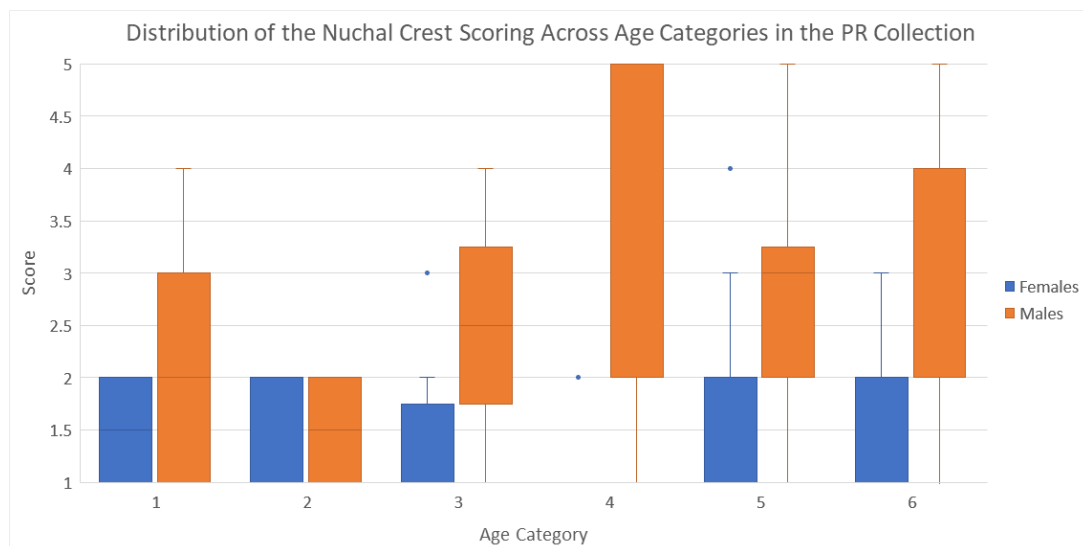

Figure E.6: A boxplot distribution of mastoid process scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.
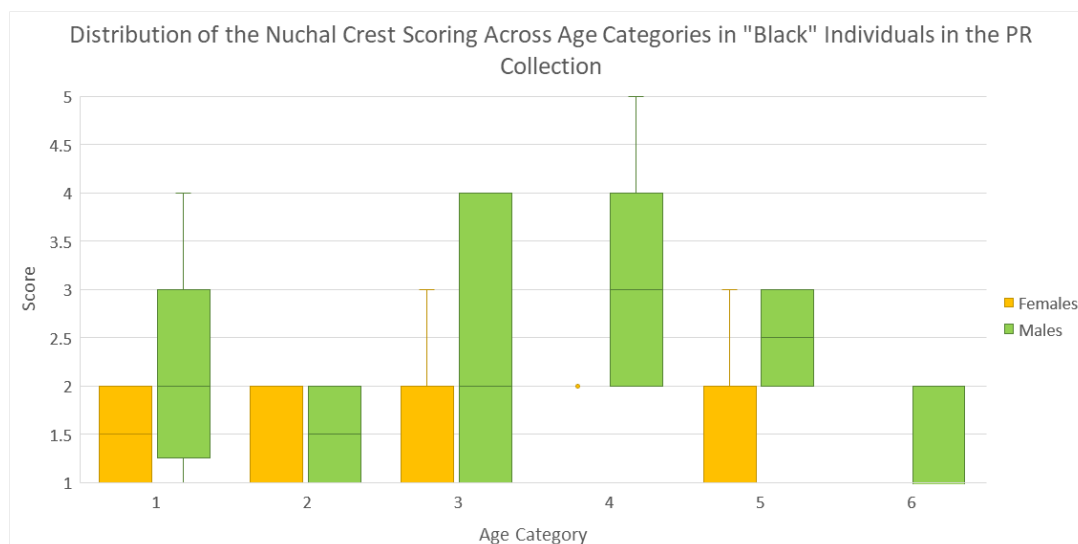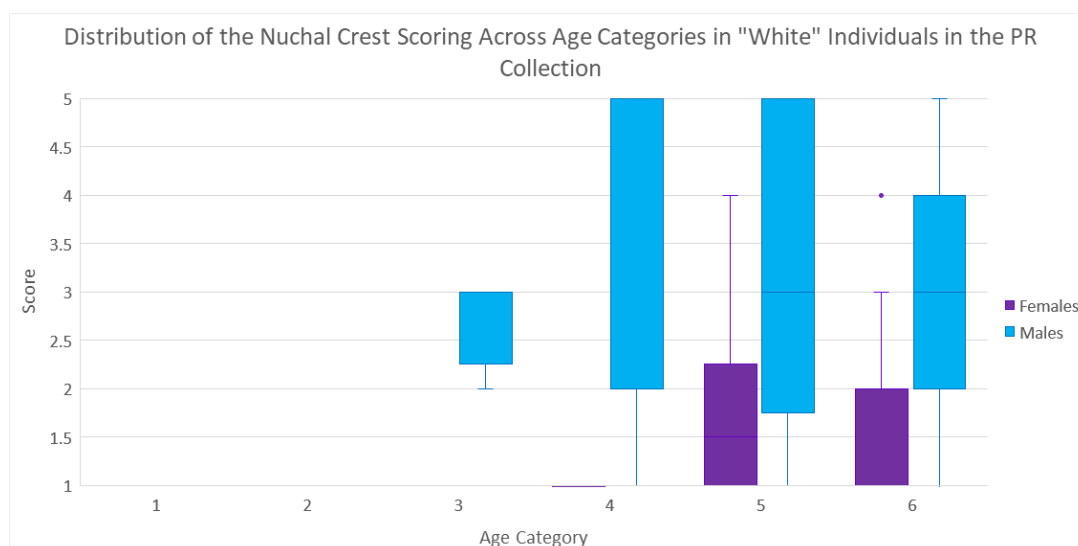
Table E.2: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the ML collection when comparing mastoid process trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 8	F = 2.0 M = 4.0	$U = 32.0$ $p = 0.002$ $z = 1.02$ $r = 0.29$	1.000
2	F = 4 M = 24	F = 2.0 M = 4.0	$U = 96.0$ $p < 0.001$ $z = 2.49$ $r = 0.47$	1.000
3	F = 16 M = 12	F = 2.5 M = 4.0	$U = 180.0$ $p << 0.001$ $z = -2.41$ $r = -0.46$	0.875
4	F = 12 M = 20	F = 2.0 M = 4.0	$U = 228.5$ $p << 0.001$ $z = 1.19$ $r = 0.21$	0.921
5	F = 32 M = 52	F = 2.0 M = 3.0	$U = 1295.0$ $p << 0.001$ $z = -0.60$ $r = -0.07$	0.822
6	F = 211 M = 200	F = 2.0 M = 4.0	$U = 33967.5$ $p << 0.001$ $z = -7.89$ $r = -0.39$	0.819
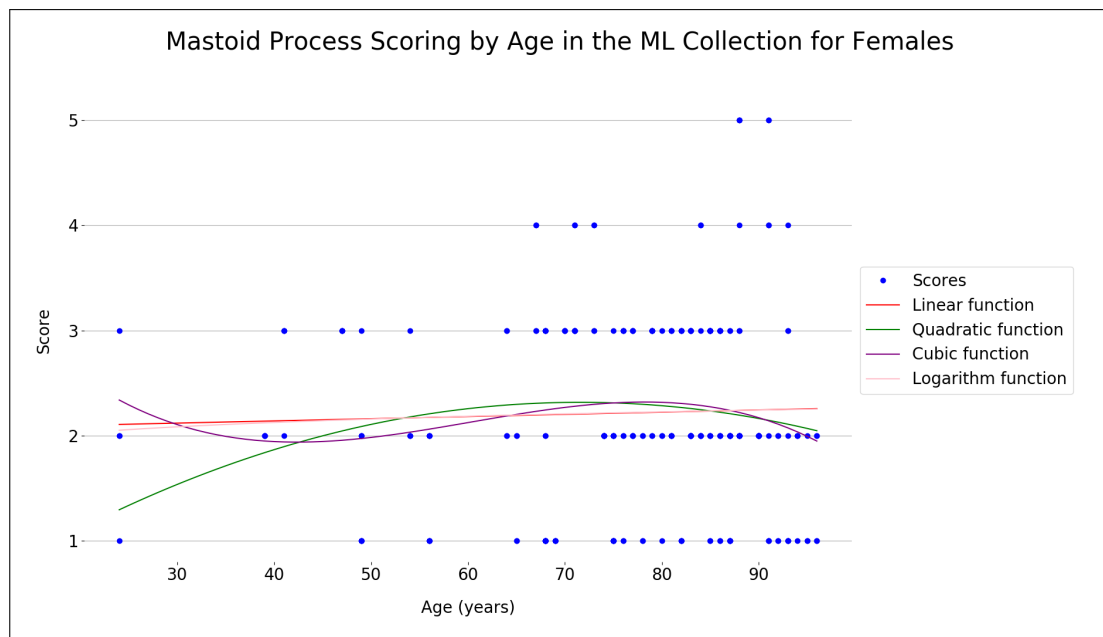
Figure E.7: A scatterplot of age vs. mastoid process trait scoring for females in the ML collection, with four fitting functions.

Figure E.8: A scatterplot of age vs. mastoid process trait scoring for males in the ML collection, with four fitting functions.

E.3 Supraorbital Margin

Figure E.9: The distribution of the supraorbital margin trait expression in the ML Collection represented using a bar chart. Females are in blue while males are in orange.

Figure E.10: A boxplot distribution of supraorbital margin scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table E.3: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the ML collection when comparing supraorbital margin trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F $= 4$ M $= 8$	F $= 1.0$ M $= 3.0$	$U = 32.0$ $p = 0.003$ $z = 1.02$ $r = 0.29$	1.000
2	F $= 4$ M $= 24$	F $= 2.5$ M $= 2.5$	$U = 46.0$ $p = 0.916$ $z = -0.79$ $r = -0.15$	0.625
3	F $= 16$ M $= 12$	F $= 1.0$ M $= 2.0$	$U = 169.0$ $p < 0.001$ $z = -2.92$ $r = -0.55$	0.802
4	F $= 12$ M $= 20$	F $= 1.0$ M $= 2.5$	$U = 206.0$ $p < 0.001$ $z = 0.31$ $r = 0.06$	0.783
5	F $= 32$ M $= 52$	F $= 1.0$ M $= 2.0$	$U = 1126.0$ $p = 0.004$ $z = -2.16$ $r = -0.24$	0.692
6	F $= 212$ M $= 204$	F $= 2.0$ M $= 2.0$	$U = 28564.0$ $p \ll 0.001$ $z = -12.76$ $r = -0.63$	0.719
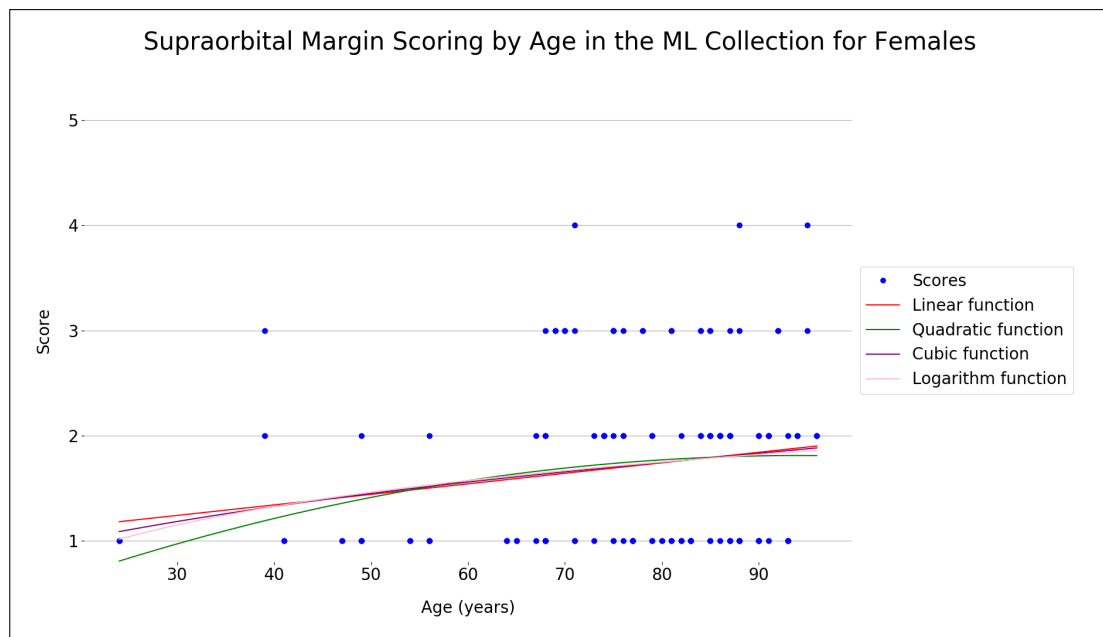
Figure E.11: A scatterplot of age vs. supraorbital margin trait scoring for females in the ML collection, with four fitting functions.

Figure E.12: A scatterplot of age vs. supraorbital margin trait scoring for males in the ML collection, with four fitting functions.

E.4   Glabella

Figure E.13: The distribution of the glabella trait expression in the ML Collection represented using a bar chart. Females are in blue while males are in orange.

Figure E.14: A boxplot distribution of glabella scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table E.4: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the ML collection when comparing glabella trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 2 M = 4	F = 1.0 M = 2.5	$U = 7.0$ $p = 0.219$ $z = 0.00$ $r = 0.00$	0.750
2	F = 2 M = 12	F = 1.0 M = 3.0	$U = 22.0$ $p = 0.074$ $z = 1.28$ $r = 0.34$	0.833
3	F = 8 M = 6	F = 1.0 M = 3.5	$U = 45.0$ $p = 0.005$ $z = -1.94$ $r = -0.52$	0.875
4	F = 6 M = 10	F = 1.0 M = 3.0	$U = 59.0$ $p = 0.001$ $z = 0.87$ $r = 0.22$	0.967
5	F = 16 M = 26	F = 1.0 M = 4.0	$U = 401.0$ $p \ll 0.001$ $z = 1.48$ $r = 0.23$	0.928
6	F = 106 M = 101	F = 2.0 M = 3.0	$U = 9336.5$ $p \ll 0.001$ $z = -3.92$ $r = -0.27$	0.818
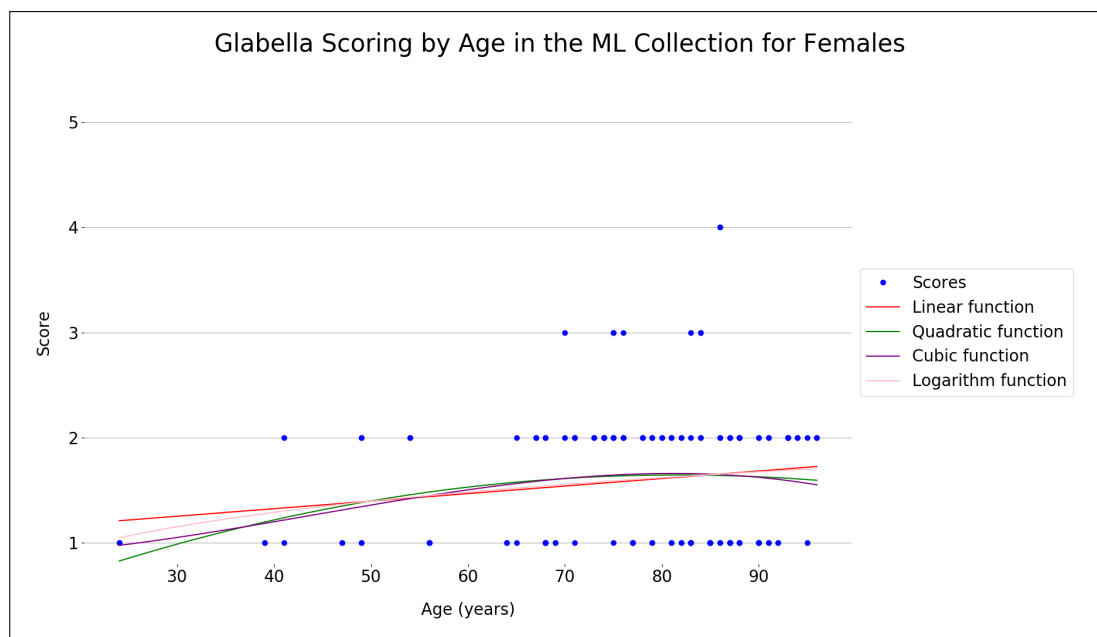
Figure E.15: A scatterplot of age vs. glabella trait scoring for females in the ML collection, with four fitting functions.

Figure E.16: A scatterplot of age vs. glabella trait scoring for males in the ML collection, with four fitting functions.
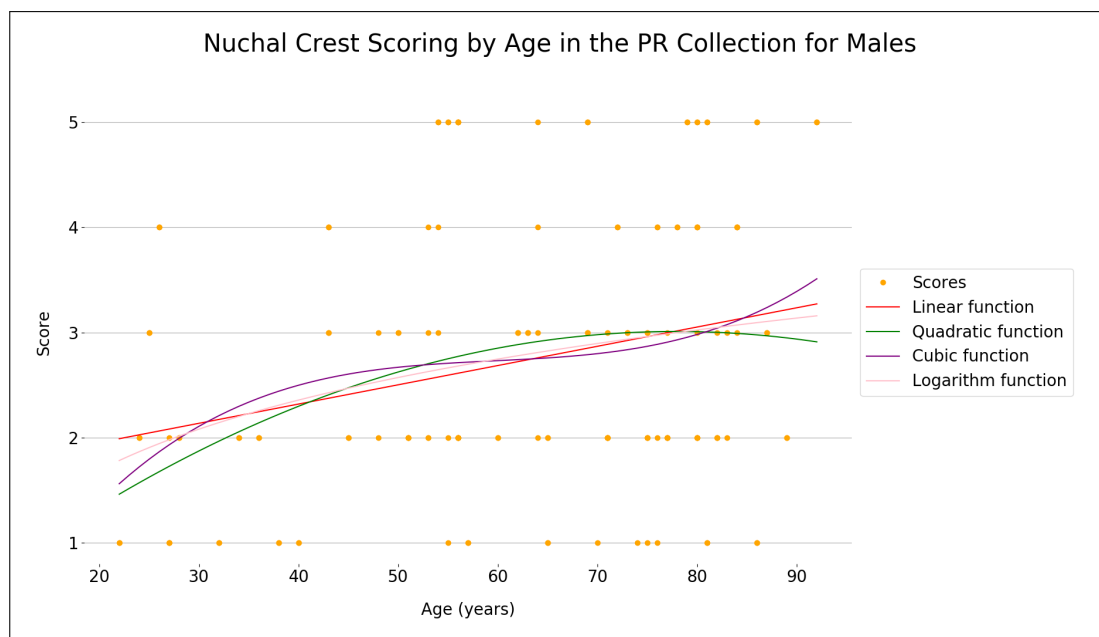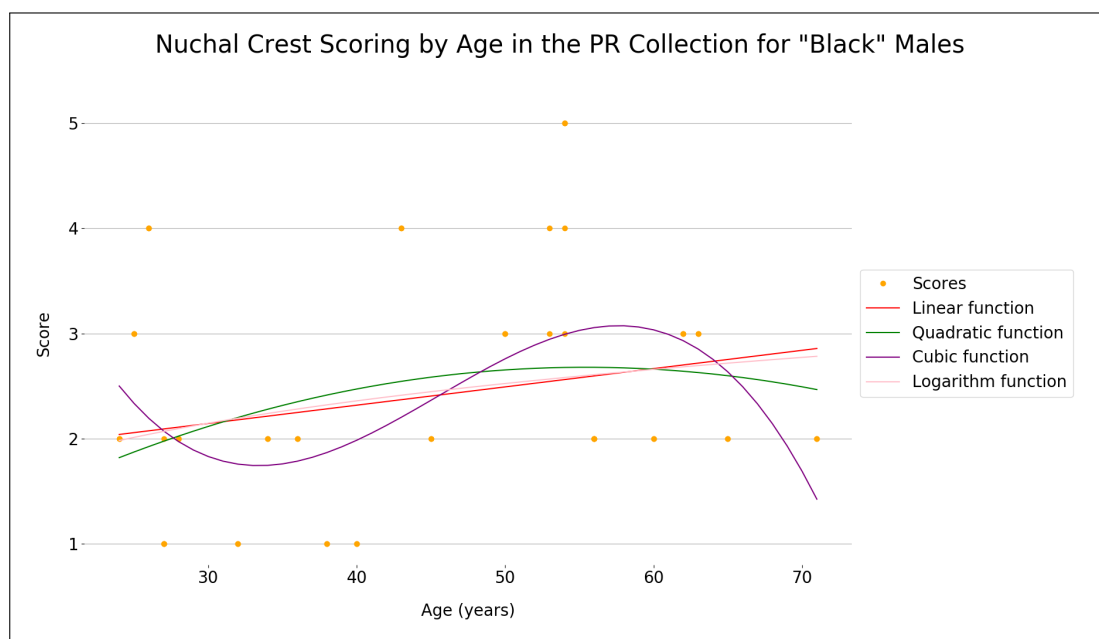
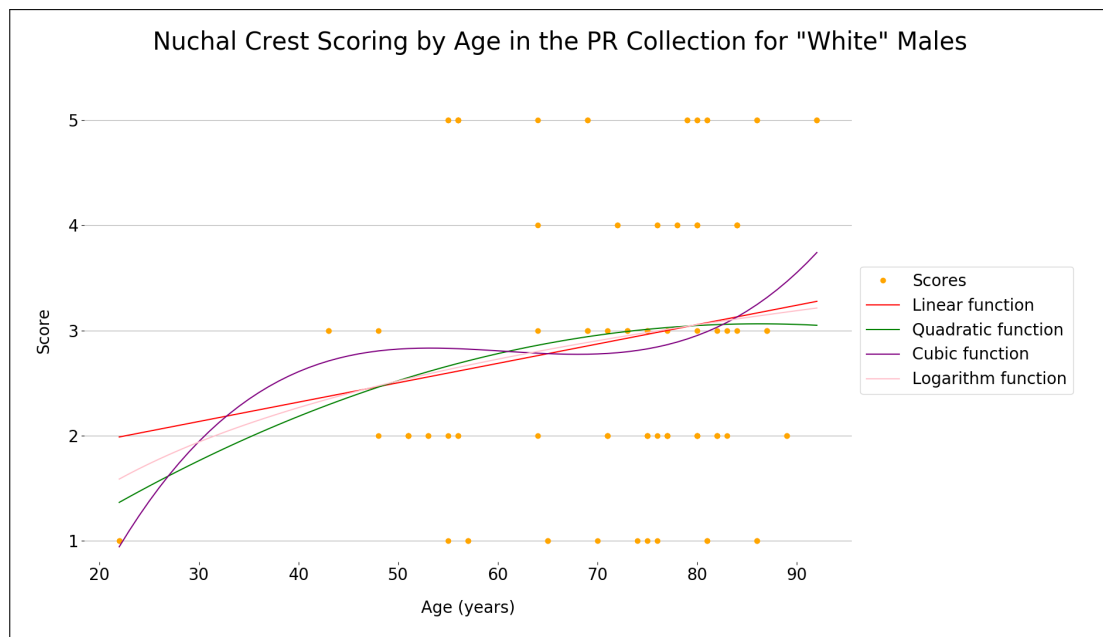## E.5   Zygomatic Extension

Figure E.17: The distribution of the zygomatic extension trait expression in the ML Collection represented using a bar chart. Females are in blue while males are in orange.
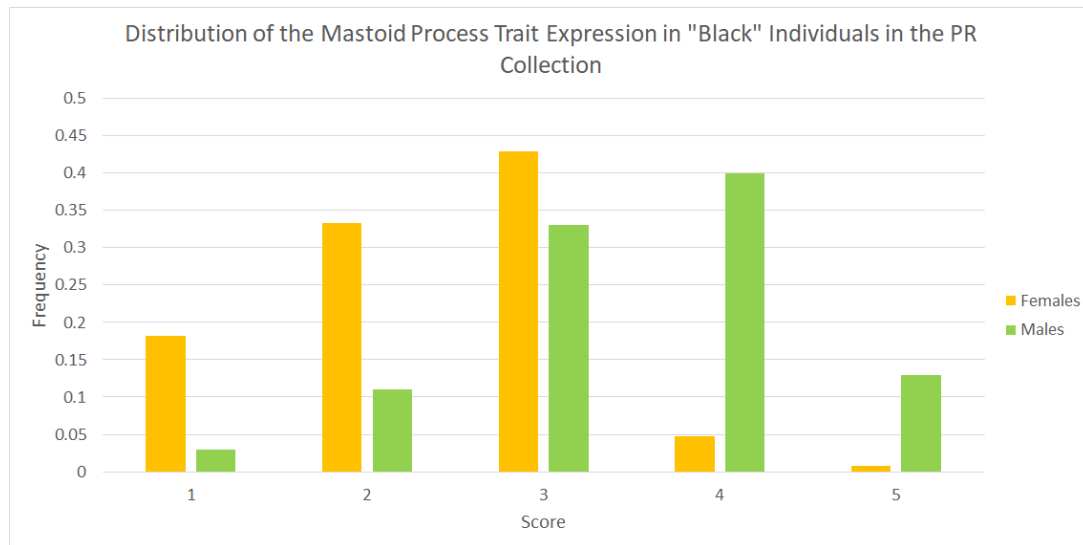

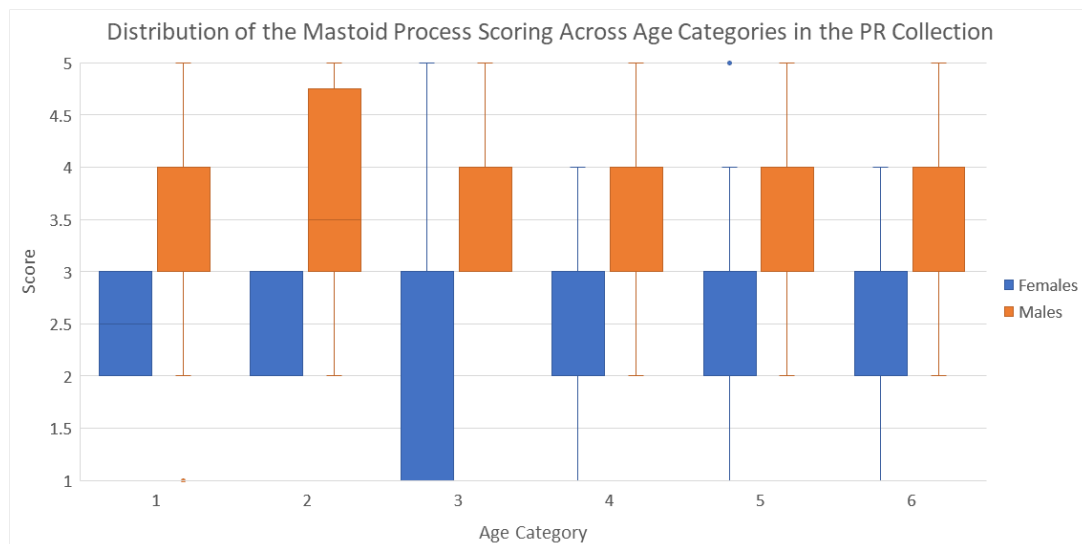**Figure E.18:** A boxplot distribution of zygomatic extension scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.
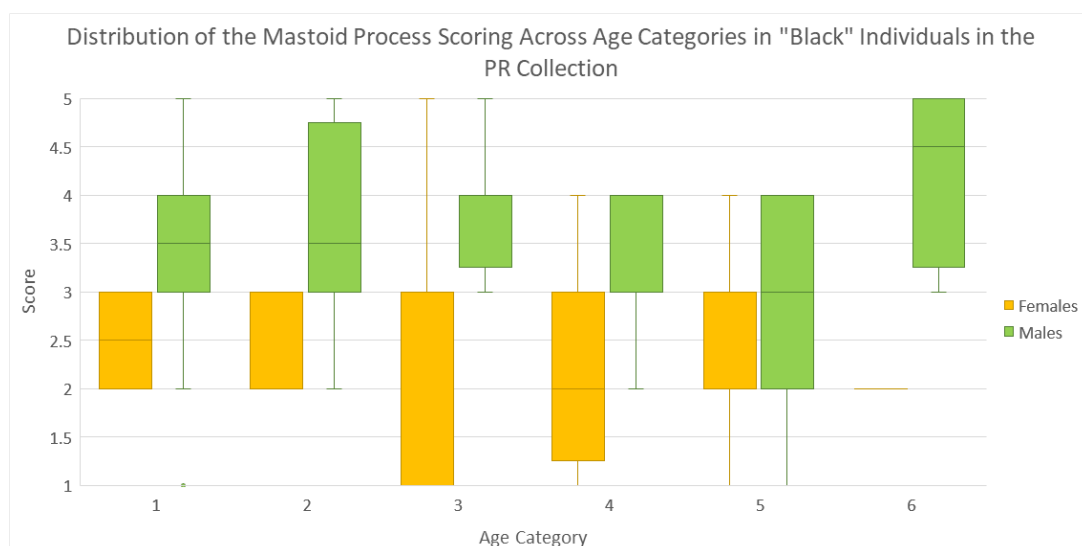
Table E.5: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the ML collection when comparing zygomatic extension trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 8	F = 1.0 M = 5.0	$U = 32.0$ $p = 0.001$ $z = 1.02$ $r = 0.29$	1.000
2	F = 4 M = 24	F = 1.0 M = 1.0	$U = 66.0$ $p = 0.156$ $z = 0.53$ $r = 0.10$	0.375
3	F = 16 M = 12	F = 1.0 M = 1.0	$U = 88.0$ $p = 0.624$ $z = -6.69$ $r = -1.26$	0.333
4	F = 12 M = 20	F = 5.0 M = 5.0	$U = 124.0$ $p = 0.865$ $z = -2.88$ $r = -0.51$	0.433
5	F = 32 M = 52	F = 1.0 M = 3.0	$U = 1014.0$ $p = 0.050$ $z = -3.19$ $r = -0.35$	0.500
6	F = 212 M = 204	F = 1.0 M = 5.0	$U = 25784.0$ $p \ll 0.001$ $z = -15.02$ $r = -0.74$	0.518
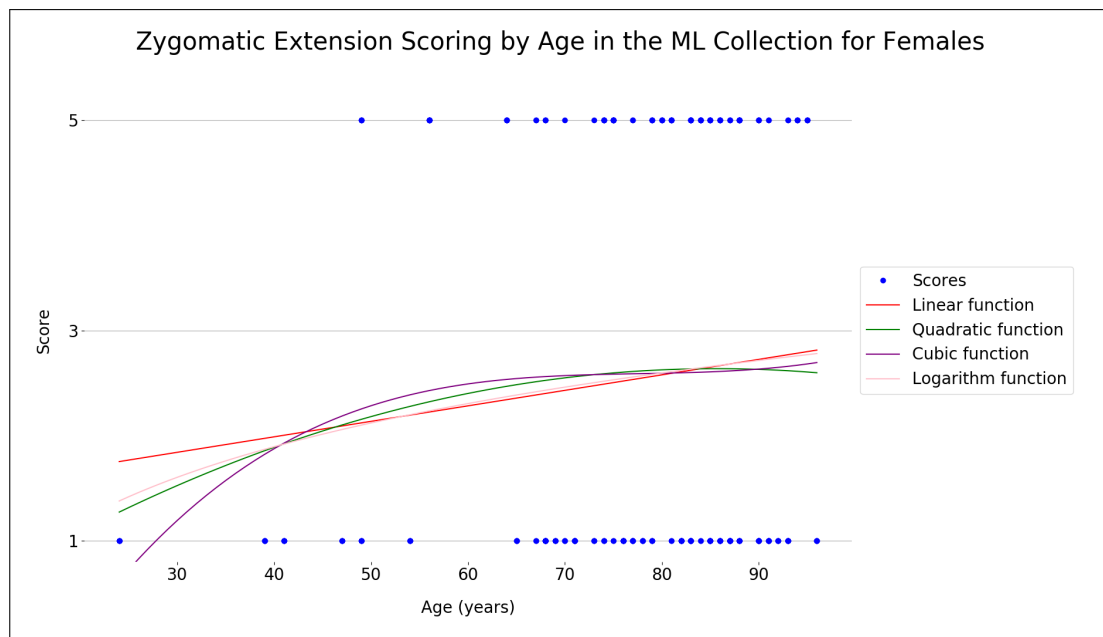
Figure E.19: A scatterplot of age vs. zygomatic extension trait scoring for females in the ML collection, with four fitting functions.

Figure E.20: A scatterplot of age vs. zygomatic extension trait scoring for males in the ML collection, with four fitting functions.

E.6   Nasal Aperture

Figure E.21: The distribution of the nasal aperture trait expression in the ML Collection represented using a bar chart. Females are in blue while males are in orange.

Figure E.22: A boxplot distribution of nasal aperture scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.
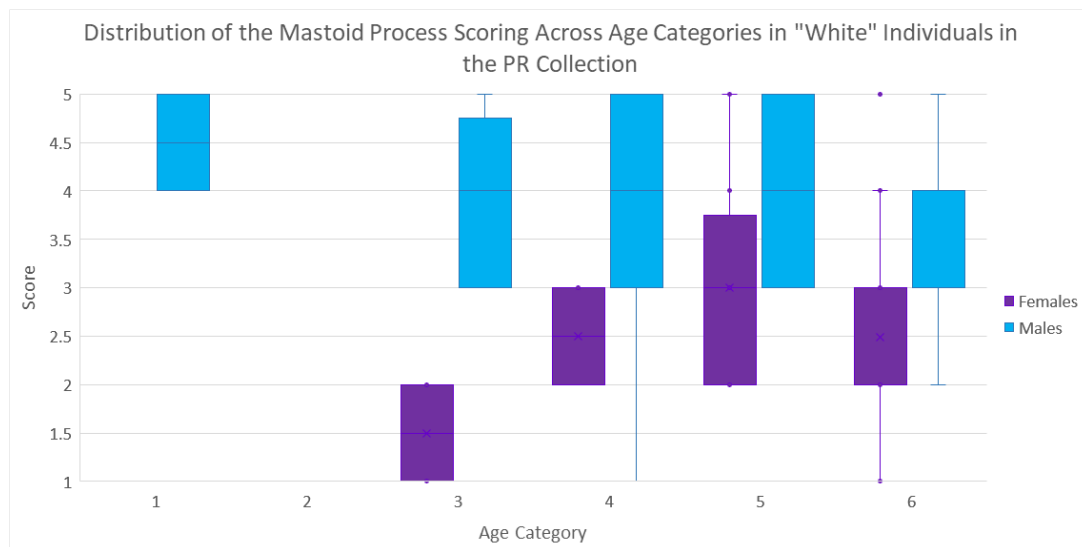
Table E.6: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the ML collection when comparing nasal aperture trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 2 M = 4	F = 5.0 M = 4.0	$U = 2.0$ $p = 0.402$ $z = -2.31$ $r = -0.94$	0.500
2	F = 2 M = 12	F = 3.0 M = 1.0	$U = 9.0$ $p = 0.620$ $z = -1.10$ $r = -0.29$	0.917
3	F = 7 M = 6	F = 1.0 M = 1.0	$U = 17.5$ $p = 0.619$ $z = -4.50$ $r = -1.25$	0.548
4	F = 6 M = 10	F = 3.0 M = 3.0	$U = 33.5$ $p = 0.728$ $z = -1.90$ $r = -0.47$	0.650
5	F = 12 M = 21	F = 1.0 M = 1.0	$U = 163.0$ $p = 0.106$ $z = -1.53$ $r = -0.27$	0.500
6	F = 100 M = 91	F = 1.0 M = 5.0	$U = 6526.0$ $p \ll 0.001$ $z = -8.06$ $r = -0.58$	0.671
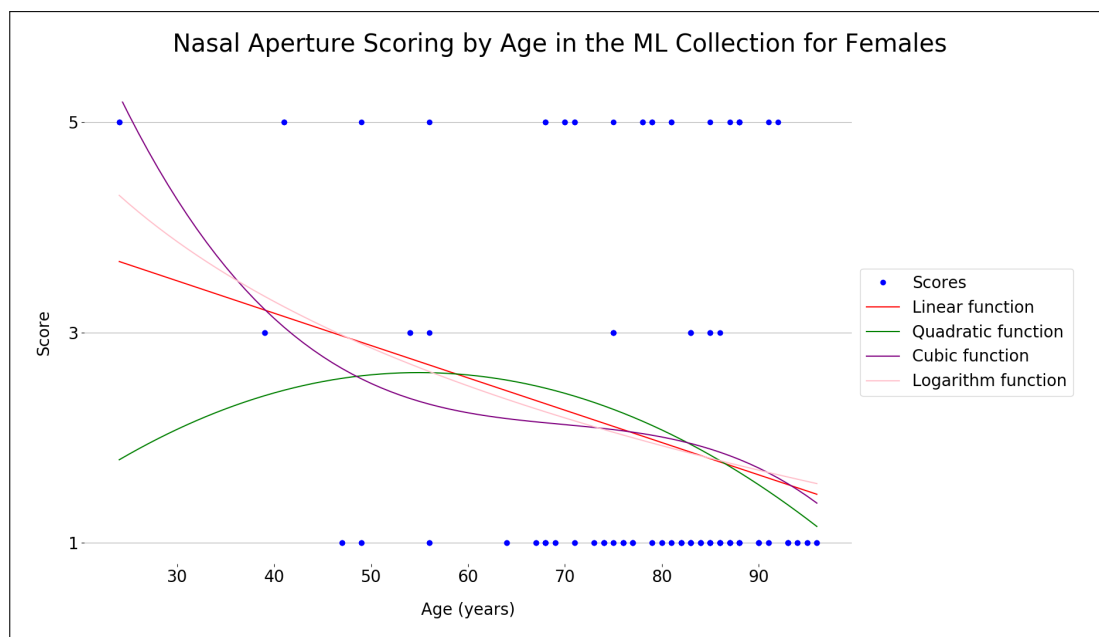
Figure E.23: A scatterplot of age vs. nasal aperture trait scoring for females in the ML collection, with four fitting functions.
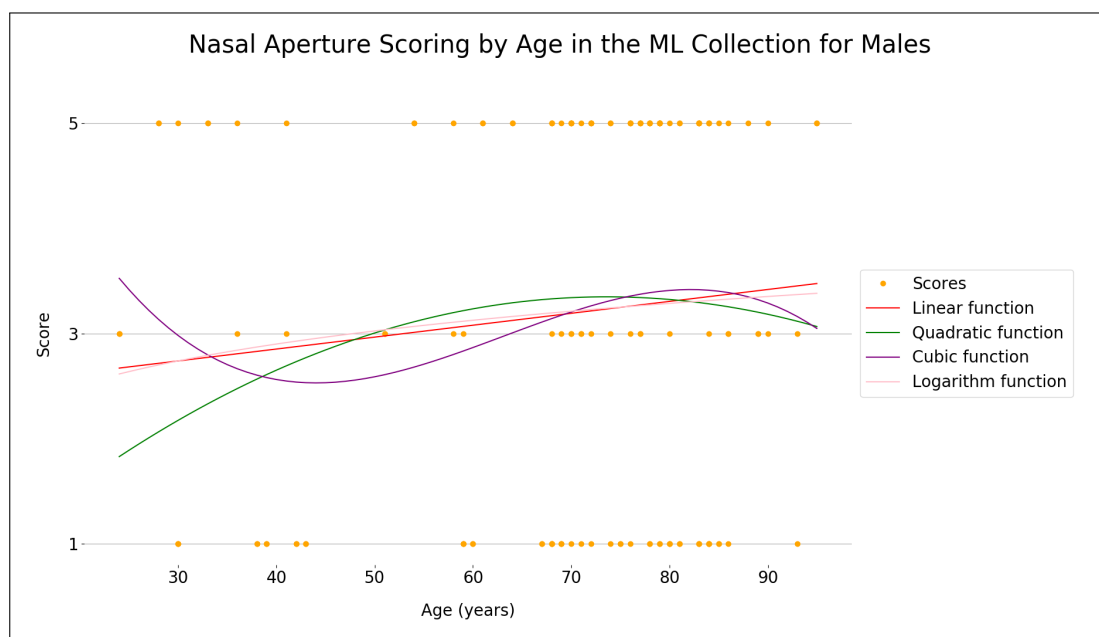
Figure E.24: A scatterplot of age vs. nasal aperture trait scoring for males in the ML collection, with four fitting functions.

E.7 Cranial Size
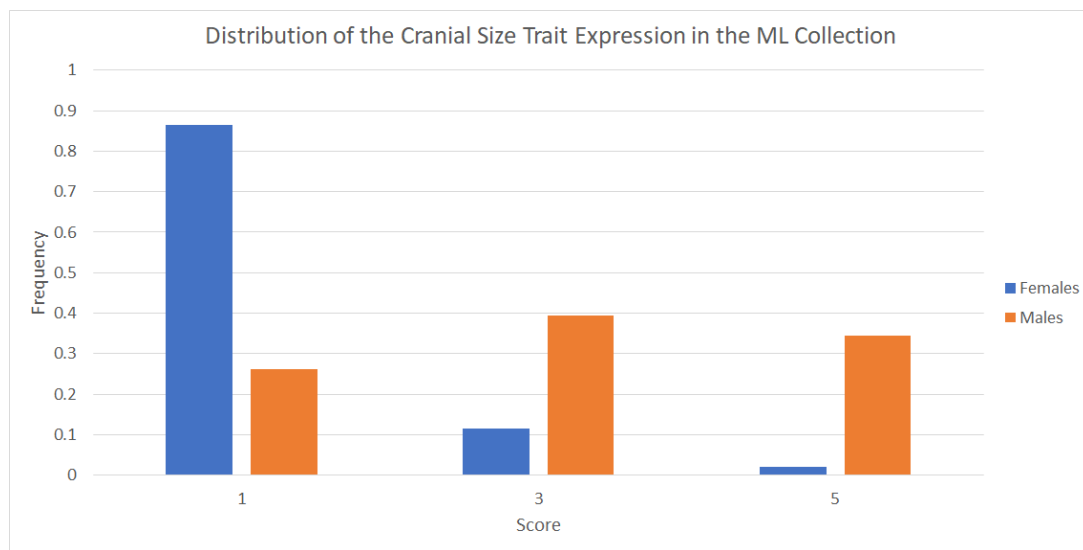
Figure E.25: The distribution of cranial size in the ML Collection represented using a bar chart. Females are in blue while males are in orange.

Figure E.26: A boxplot distribution of cranial size scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Table E.7: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the ML collection when comparing cranial size scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 2 M = 4	F = 1.0 M = 2.0	$U = 6.0$ $p = 0.411$ $z = $ -0.46 $r = $ -0.19	0.500
2	F = 2 M = 12	F = 2.0 M = 3.0	$U = 20.5$ $p = 0.096$ $z = 1.00$ $r = 0.27$	0.708
3	F = 8 M = 6	F = 1.0 M = 1.0	$U = 26.0$ $p = 0.805$ $z = $ -4.39 $r = $ -1.17	0.417
4	F = 6 M = 10	F = 1.0 M = 4.0	$U = 57.0$ $p = 0.002$ $z = 0.65$ $r = 0.16$	0.900
5	F = 16 M = 26	F = 1.0 M = 3.0	$U = 331.0$ $p < 0.001$ $z = $ -0.34 $r = $ -0.05	0.668
6	F = 106 M = 102	F = 1.0 M = 3.0	$U = 8827.5$ $p << 0.001$ $z = $ -5.18 $r = $ -0.36	0.723
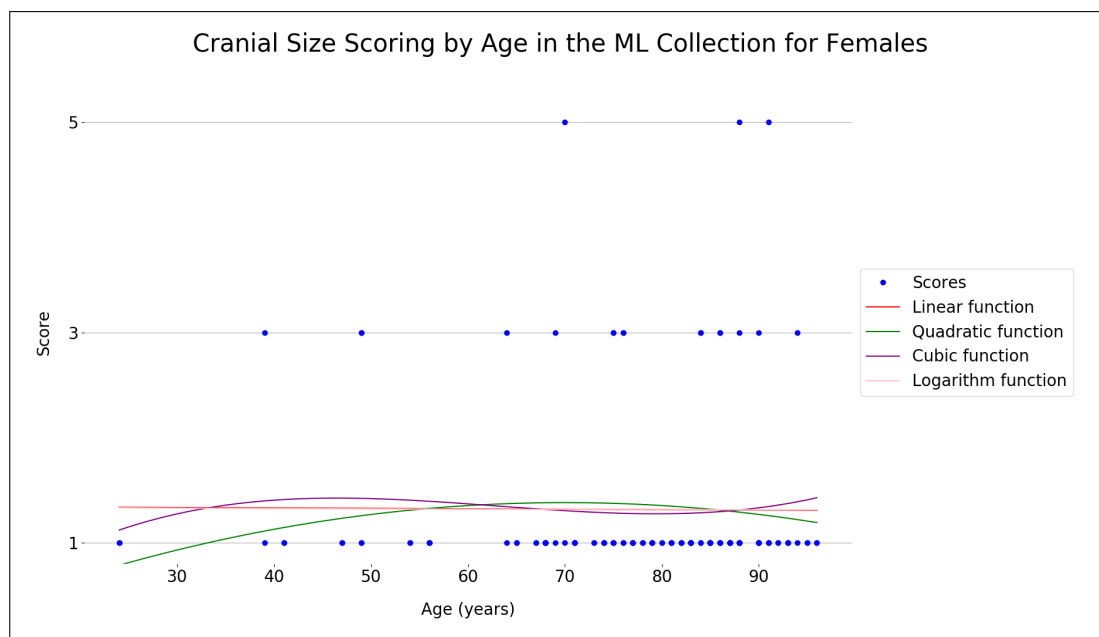
Figure E.27: A scatterplot of age vs. cranial size scoring for females in the ML collection, with four fitting functions.
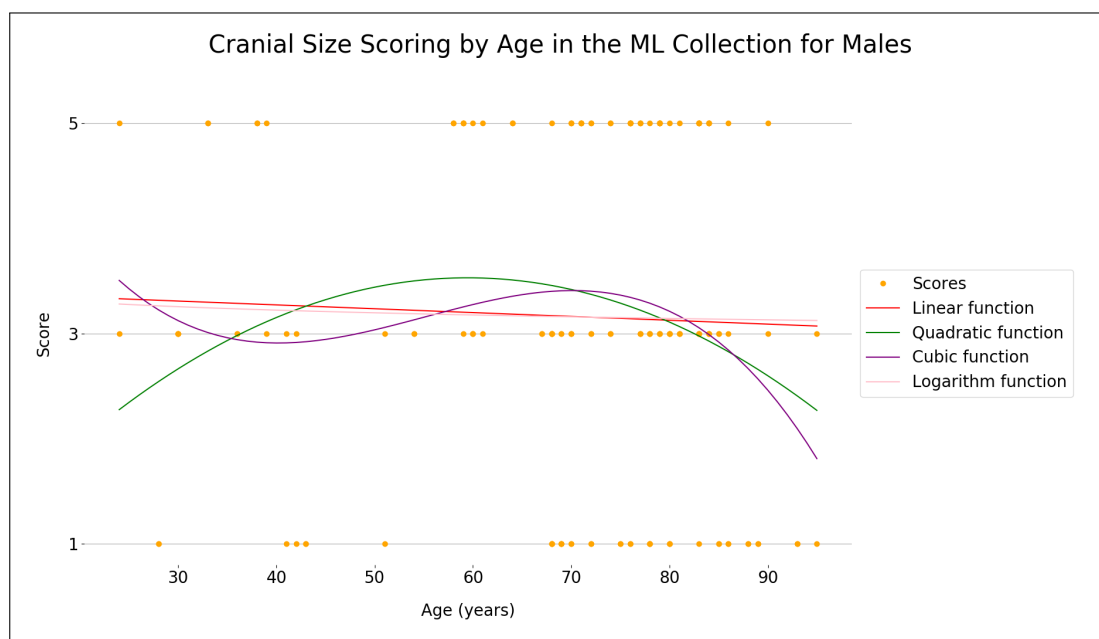
Figure E.28: A scatterplot of age vs. cranial size scoring for males in the ML collection, with four fitting functions.

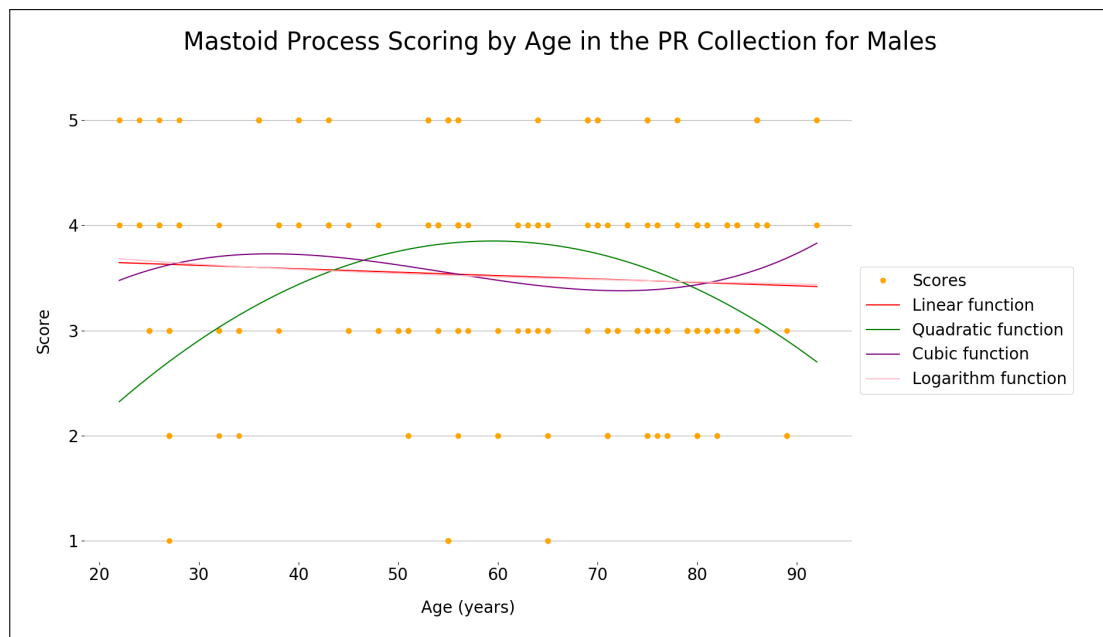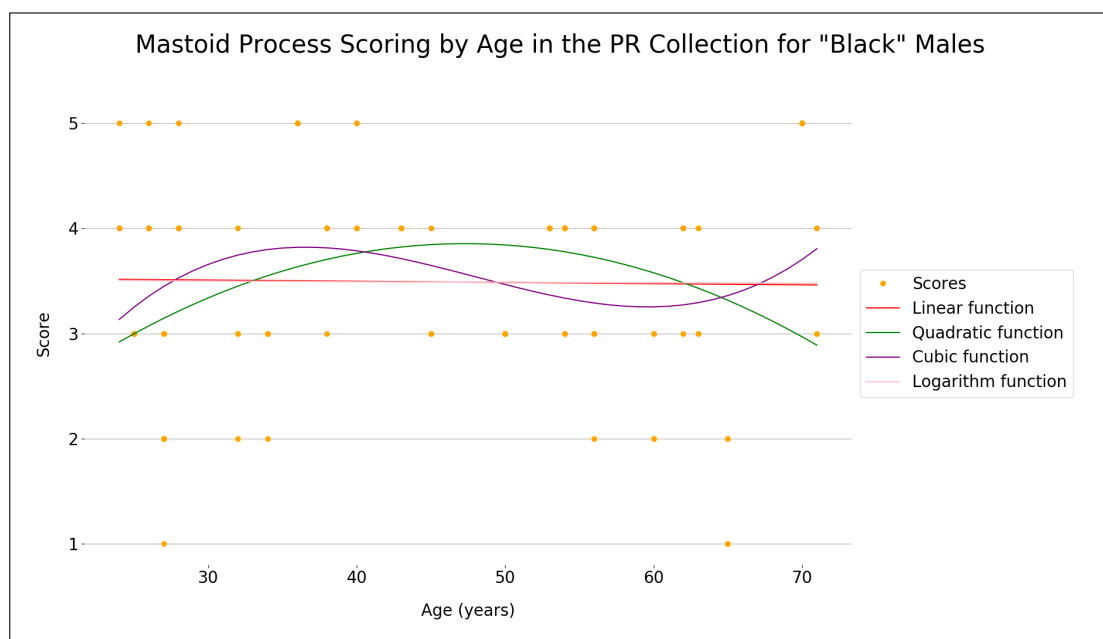# APPENDIX F: Trait Distribution Graphs for

the PR Collection

F.1 Nuchal Crest
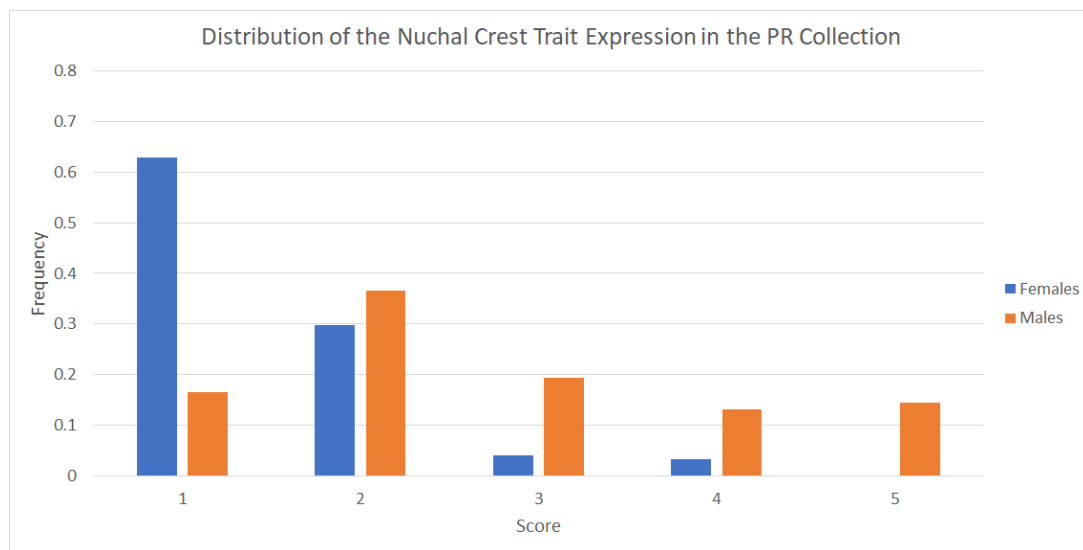
Figure F.1: The distribution of the nuchal crest trait expression in the PR Collection represented using a bar chart. Females are in blue while males are in orange.
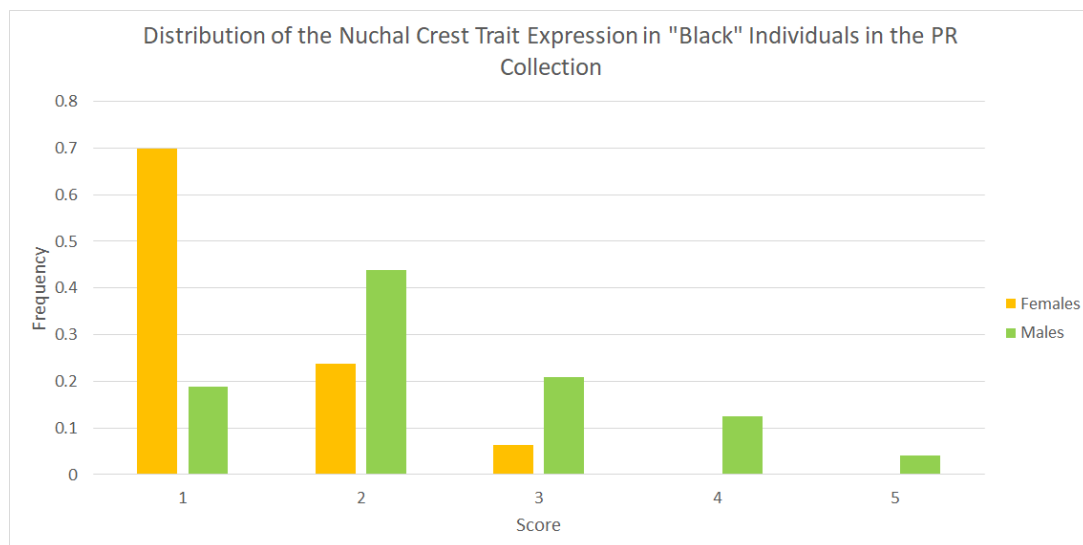
Figure F.2: The distribution of the nuchal crest trait expression in "Black" individuals from the PR Collection represented using a bar chart. Females are in yellow while males are in green.
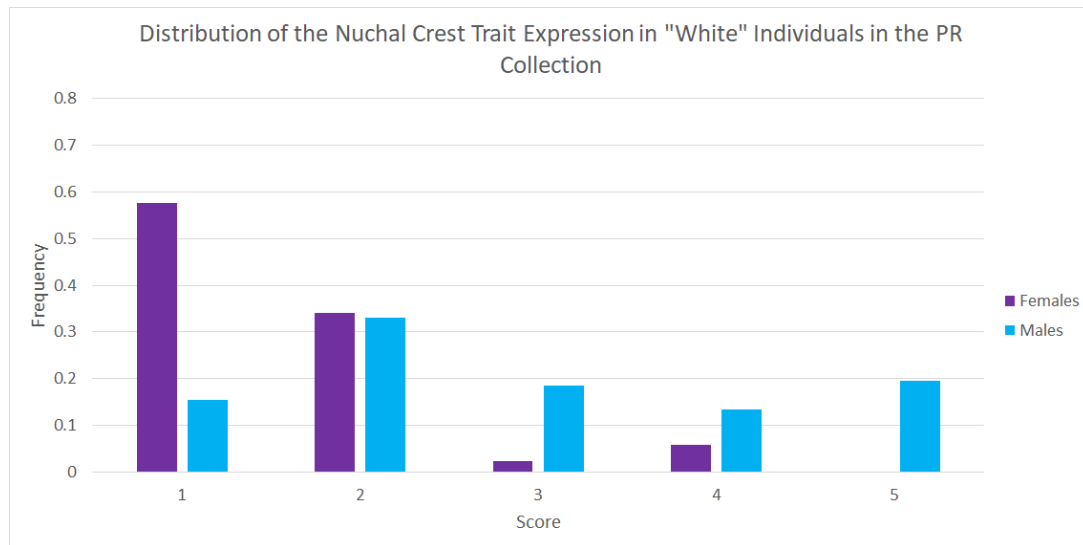
Figure F.3: The distribution of the nuchal crest trait expression in "White" individuals from the PR Collection represented using a bar chart. Females are in purple while males are in cyan.

Figure F.4: A boxplot distribution of nuchal crest scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Figure F.5: A boxplot distribution of nuchal crest scoring across different age categories for "Black" males and females. Females are given in yellow while males are in green. The age categories are defined in Table 2.2.

Figure F.6: A boxplot distribution of nuchal crest scoring across different age categories for "White" males and females. Females are given in purple while males are in cyan. The age categories are defined in Table 2.2.

Table F.1: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the PR collection when comparing nuchal crest trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

| Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 14	F = 1.5 M = 2.0	$U = 36.0$ $p = 0.396$ z = -0.21 r = -0.05	0.643
2	F = 10 M = 8	F = 1.0 M = 1.5	$U = 44.0$ $p = 0.718$ z = -4.53 r = -1.07	0.500
3	F = 20 M = 10	F = 1.0 M = 2.5	$U = 161.5$ $p = 0.003$ z = -6.53 r = -1.19	0.775
4	F = 18 M = 30	F = 1.0 M = 2.0	$U = 489.0$ $p << 0.001$ z = 1.02 r = 0.15	0.844
5	F = 22 M = 18	F = 1.0 M = 3.0	$U = 311.0$ $p = 0.001$ z = -3.81 r = -0.60	0.808
6	F = 74 M = 65	F = 1.0 M = 3.0	$U = 3898.0$ $p << 0.001$ z = -5.41 r = -0.46	0.792

Table F.2: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "Black" individuals from the PR collection when comparing nuchal crest trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 12	F = 1.5 M = 2.0	$U = 34.0$ $p = 0.221$ $z = 0.00$ $r = 0.00$	0.667
2	F = 10 M = 8	F = 1.0 M = 1.5	$U = 44.0$ $p = 0.718$ $z = $ -4.53 $r = $ -1.07	0.500
3	F = 18 M = 6	F = 1.0 M = 2.0	$U = 78.0$ $p = 0.070$ $z = $ -9.80 $r = $ -2.00	0.704
4	F = 15 M = 12	F = 1.0 M = 3.0	$U = 174.0$ $p << 0.001$ $z = $ -1.76 $r = $ -0.34	0.933
5	F = 12 M = 8	F = 1.0 M = 2.5	$U = 78.0$ $p = 0.016$ $z = $ -3.70 $r = $ -0.83	0.792
6	F = 4 M = 2	F = 1.0 M = 2.0	$U = 8.0$ $p = 0.050$ $z = $ -2.78 $r = $ -1.13	1.000

Table F.3: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "White" individuals from the PR collection when comparing nuchal crest trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 0 M = 2	F = N/A M = 1.0	U = N/A p = N/A z = N/A r = N/A	N/A
3	F = 2 M = 4	F = 1.0 M = 3.0	U = 8.0 p = 0.080 z = 0.46 r = 0.19	1.000
4	F = 3 M = 18	F = 1.0 M = 2.0	U = 49.5 p = 0.018 z = 1.66 r = 0.36	0.833
5	F = 10 M = 10	F = 1.5 M = 3.0	U = 76.5 p = 0.043 z = -2.15 r = -0.48	0.830
6	F = 70 M = 63	F = 1.0 M = 3.0	U = 3550.0 $p << 0.001$ z = -5.14 r = -0.45	0.792
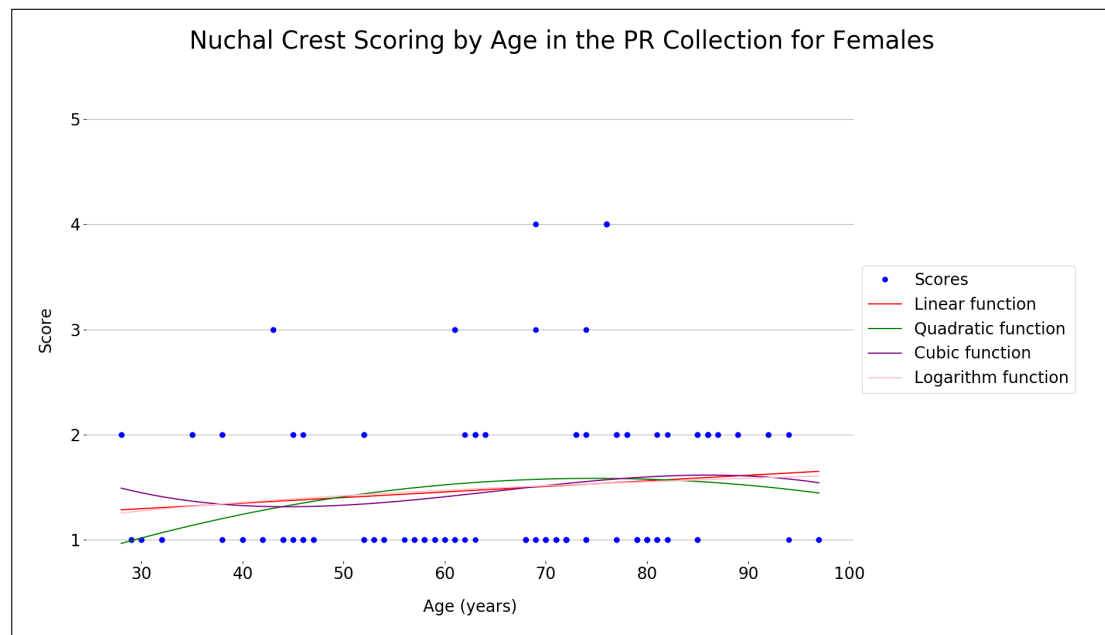
Figure F.7: A scatterplot of age vs. nuchal crest trait scoring for females in the PR collection, with four fitting functions.
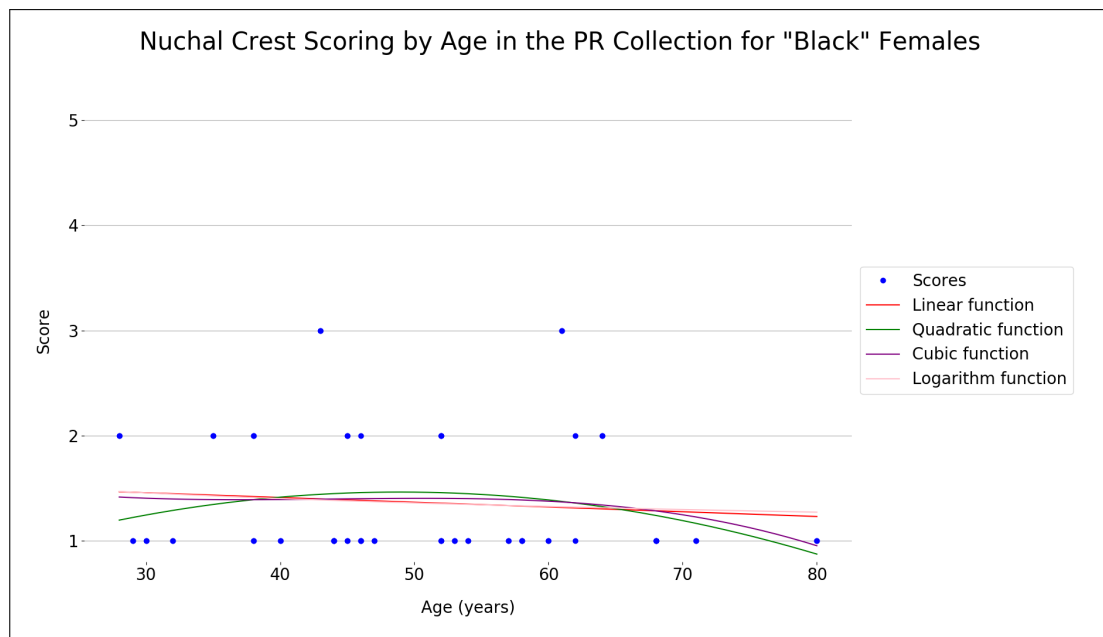
Figure F.8: A scatterplot of age vs. nuchal crest trait scoring for "Black" females in the PR collection, with four fitting functions.
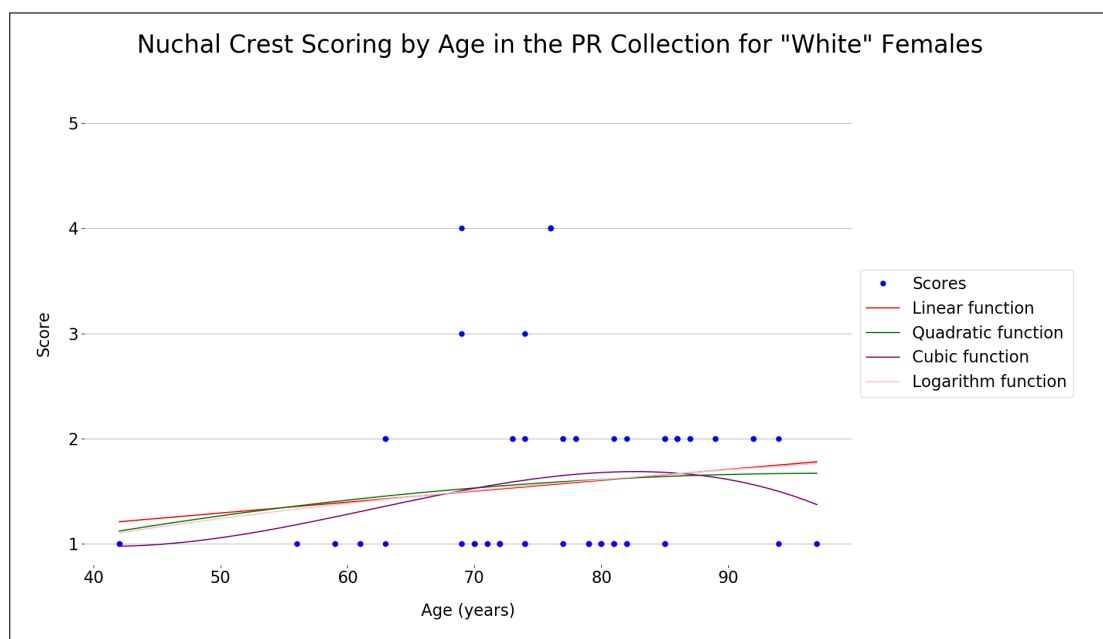
Figure F.9: A scatterplot of age vs. nuchal crest trait scoring for "White" females in the PR collection, with four fitting functions.

Figure F.10: A scatterplot of age vs. nuchal crest trait scoring for males in the PR collection, with four fitting functions.

Figure F.11: A scatterplot of age vs. nuchal crest trait scoring for "Black" males in the PR collection, with four fitting functions.

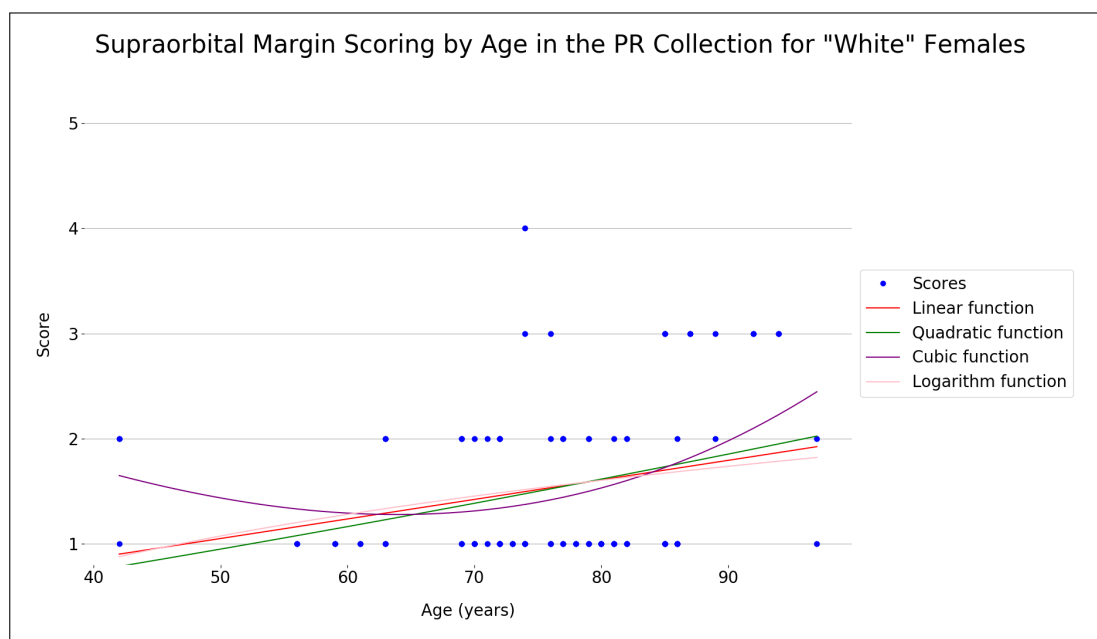**Figure F.12:** A scatterplot of age vs. nuchal crest trait scoring for "White" males in the PR collection, with four fitting functions.

F.2 Mastoid Process

Figure F.13: The distribution of the mastoid process trait expression in the PR Collection represented using a bar chart. Females are in blue while males are in orange.

Figure F.14: The distribution of the mastoid process trait expression in "Black" individuals from the PR Collection represented using a bar chart. Females are in yellow while males are in green.
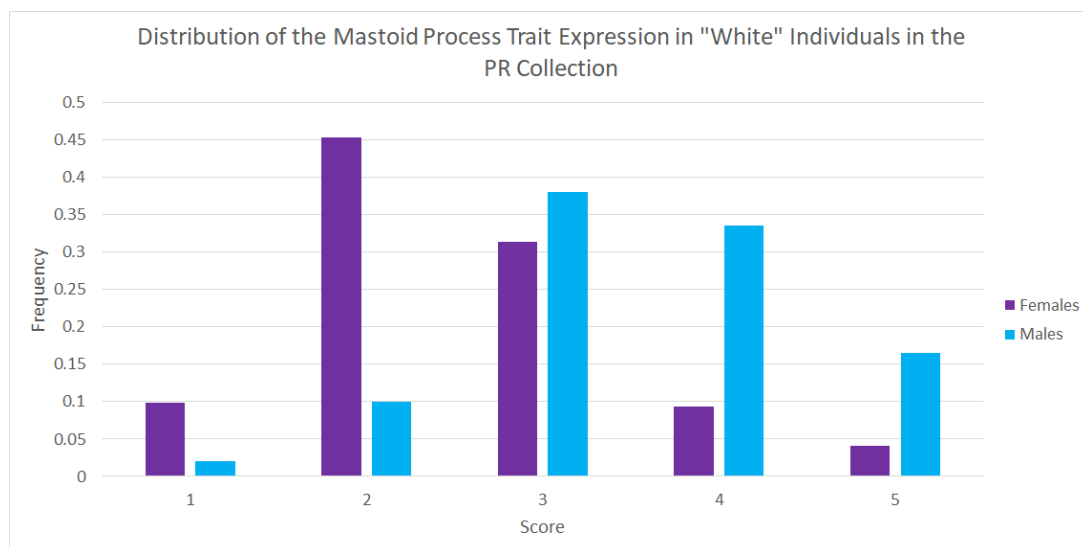
Figure F.15: The distribution of the mastoid process trait expression in "White" individuals from the PR Collection represented using a bar chart. Females are in purple while males are in cyan.

Figure F.16: A boxplot distribution of mastoid process scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Figure F.17: A boxplot distribution of mastoid process scoring across different age categories for "Black" males and females. Females are given in yellow while males are in green. The age categories are defined in Table 2.2.

Figure F.18: A boxplot distribution of mastoid process scoring across different age categories for "White" males and females. Females are given in purple while males are in cyan. The age categories are defined in Table 2.2.

Table F.4: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the PR collection when comparing mastoid process trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 28	F = 2.5 M = 4.0	$U = 178.0$ $p = 0.010$ $z = 1.14$ $r = 0.19$	0.804
2	F = 20 M = 16	F = 2.0 M = 3.5	$U = 268.0$ $p < 0.001$ $z = -3.25$ $r = -0.54$	0.775
3	F = 40 M = 20	F = 3.0 M = 4.0	$U = 690.0$ $p \ll 0.001$ $z = -8.31$ $r = -1.07$	0.825
4	F = 40 M = 60	F = 2.0 M = 4.0	$U = 1998.0$ $p \ll 0.001$ $z = -0.15$ $r = -0.02$	0.832
5	F = 42 M = 36	F = 3.0 M = 3.0	$U = 1034.5$ $p = 0.003$ $z = -6.26$ $r = -0.71$	0.756
6	F = 148 M = 140	F = 2.0 M = 3.0	$U = 16094.5$ $p \ll 0.001$ $z = -7.49$ $r = -0.44$	0.780

Table F.5: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "Black" individuals from the PR collection when comparing mastoid process trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 24	F = 2.5 M = 3.5	$U = 146.0$ $p = 0.025$ $z = 0.61$ $r = 0.11$	0.771
2	F = 20 M = 16	F = 2.0 M = 3.5	$U = 268.0$ $p < 0.001$ $z = -3.25$ $r = -0.54$	0.775
3	F = 36 M = 12	F = 3.0 M = 4.0	$U = 371.0$ $p < 0.001$ $z = -12.17$ $r = -1.76$	0.819
4	F = 32 M = 24	F = 2.0 M = 4.0	$U = 667.5$ $p << 0.001$ $z = -4.05$ $r = -0.54$	0.801
5	F = 22 M = 16	F = 3.0 M = 3.0	$U = 205.0$ $p = 0.368$ $z = -6.62$ $r = -1.07$	0.727
6	F = 8 M = 8	F = 2.0 M = 4.5	$U = 64.0$ $p < 0.001$ $z = -0.42$ $r = -0.11$	1.000

Table F.6: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "White" individuals from the PR collection when comparing mastoid process trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue. Note that there were no individuals belonging to age category 2.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 0 M = 4	F = N/A M = 4.5	U = N/A p = N/A z = N/A r = N/A	N/A
3	F = 4 M = 8	F = 1.5 M = 4.0	U = 32.0 p = 0.007 z = 1.02 r = 0.29	1.000
4	F = 8 M = 36	F = 2.5 M = 4.0	U = 228.0 p = 0.009 z = 1.46 r = 0.22	0.861
5	F = 20 M = 20	F = 3.0 M = 4.0	U = 299.0 p = 0.005 z = -3.00 r = -0.47	0.785
6	F = 140 M = 132	F = 2.0 M = 3.0	U = 14035.5 $p \ll 0.001$ z = -7.83 r = -0.47	0.768

Figure F.19: A scatterplot of age vs. mastoid process trait scoring for females in the PR collection, with four fitting functions.
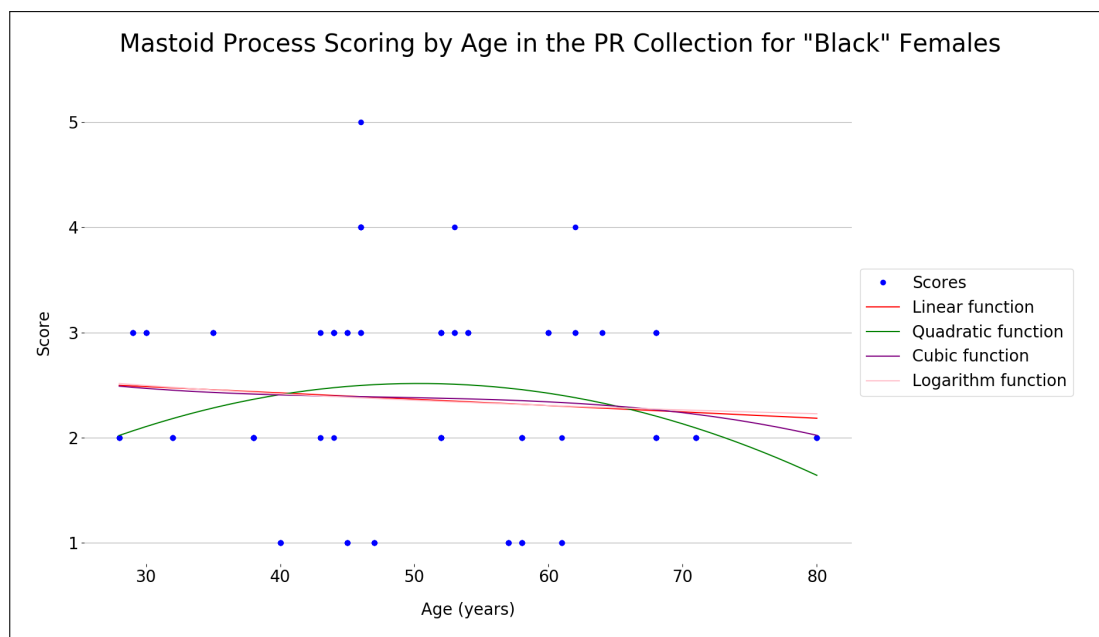
Figure F.20: A scatterplot of age vs. mastoid process trait scoring for "Black" females in the PR collection, with four fitting functions.
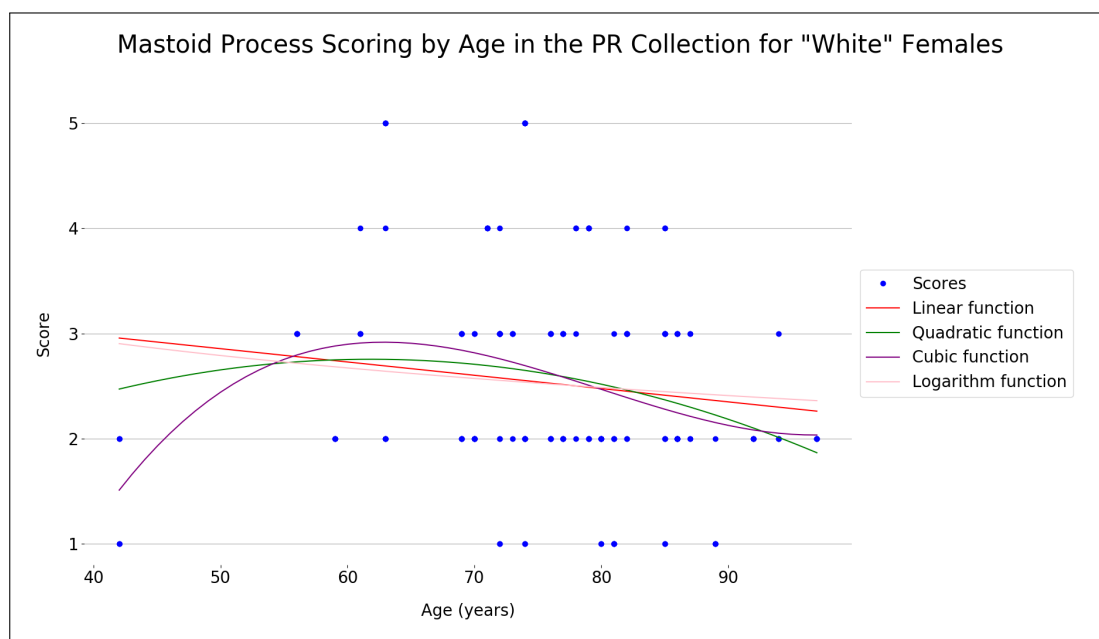
Figure F.21: A scatterplot of age vs. mastoid process trait scoring for "White" females in the PR collection, with four fitting functions.

Figure F.22: A scatterplot of age vs. mastoid process trait scoring for males in the PR collection, with four fitting functions.

Figure F.23: A scatterplot of age vs. mastoid process trait scoring for "Black" males in the PR collection, with four fitting functions.
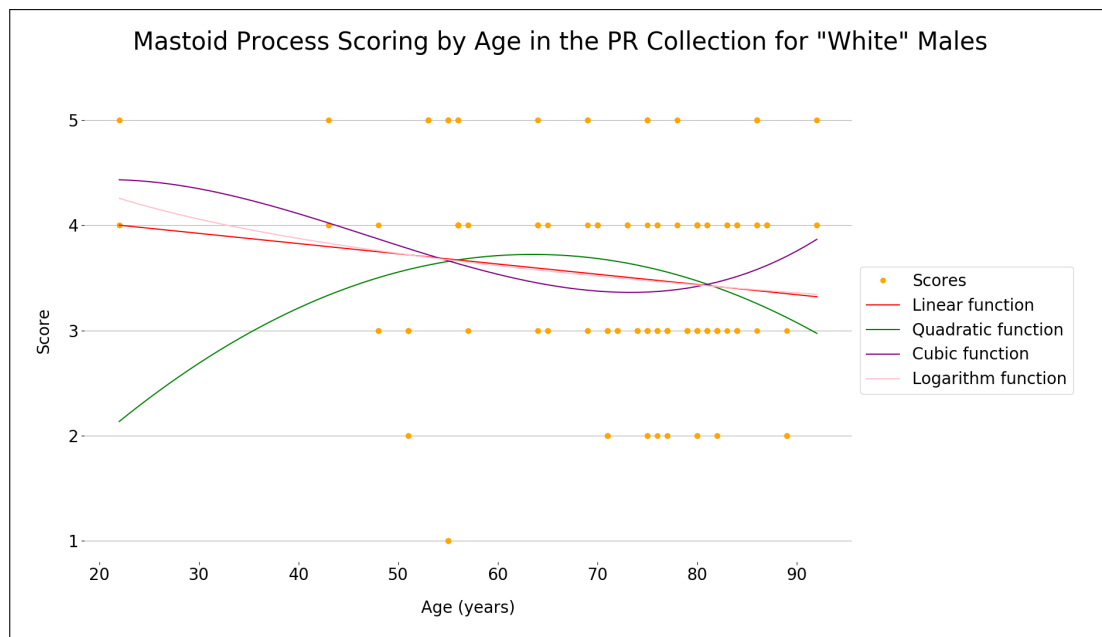
Figure F.24: A scatterplot of age vs. mastoid process trait scoring for "White" males in the PR collection, with four fitting functions.

F.3 Supraorbital Margin
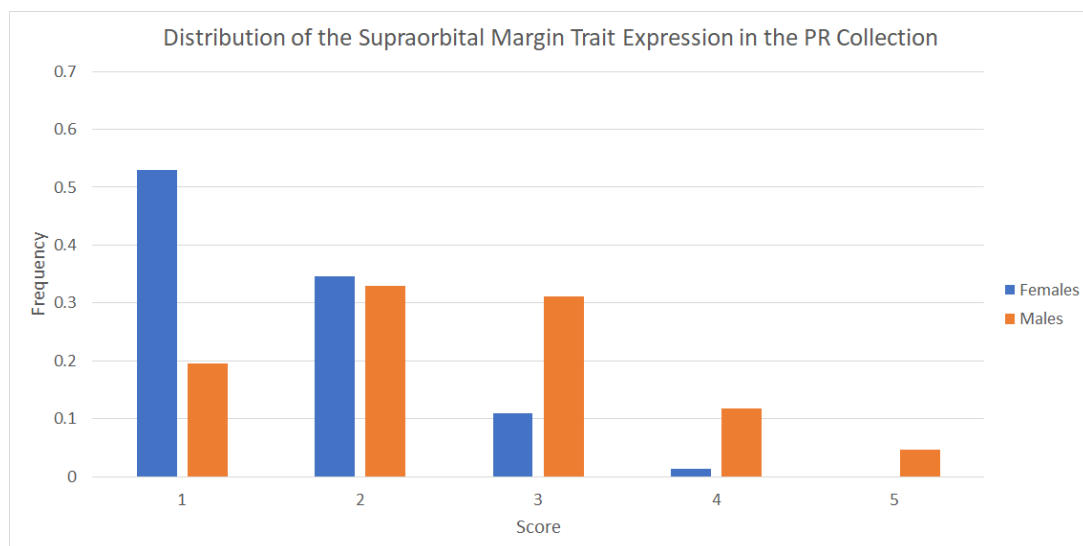
Figure F.25: The distribution of the supraorbital margin trait expression in the PR Collection represented using a bar chart. Females are in blue while males are in orange.
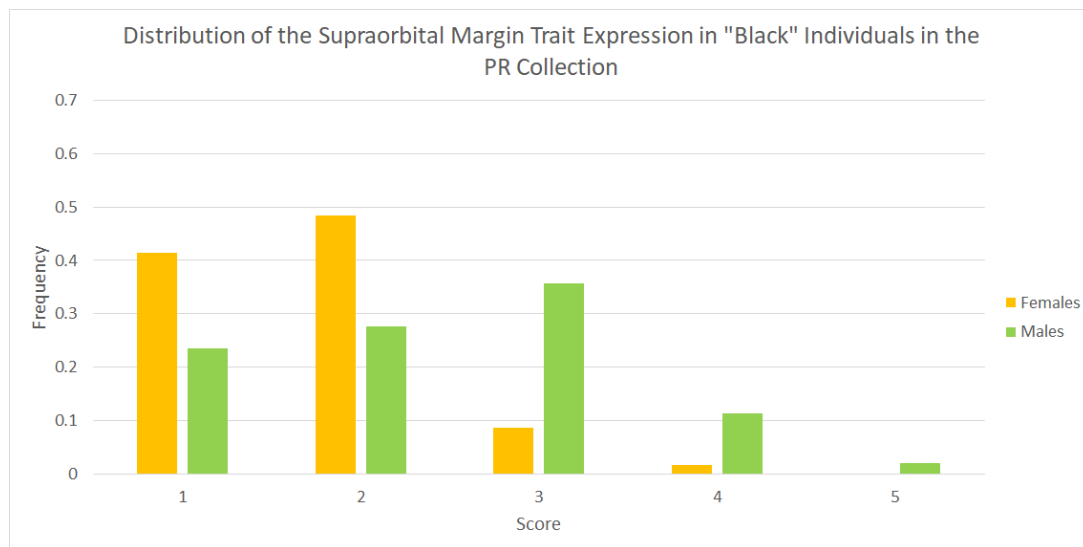
Figure F.26: The distribution of the supraorbital margin trait expression in "Black" individuals from the PR Collection represented using a bar chart. Females are in yellow while males are in green.
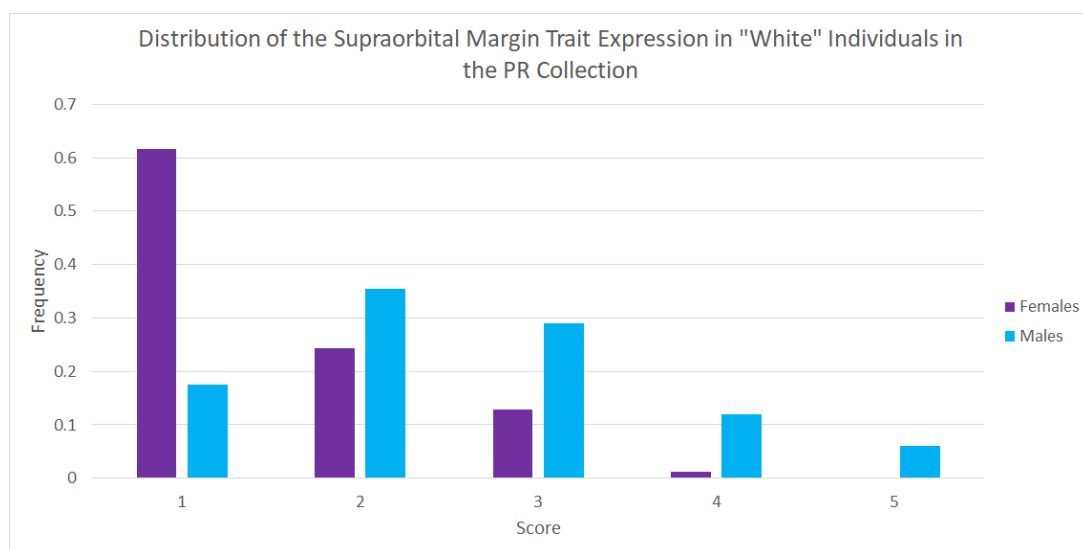
Figure F.27: The distribution of the supraorbital margin trait expression in "White" individuals from the PR Collection represented using a bar chart. Females are in purple while males are in cyan.
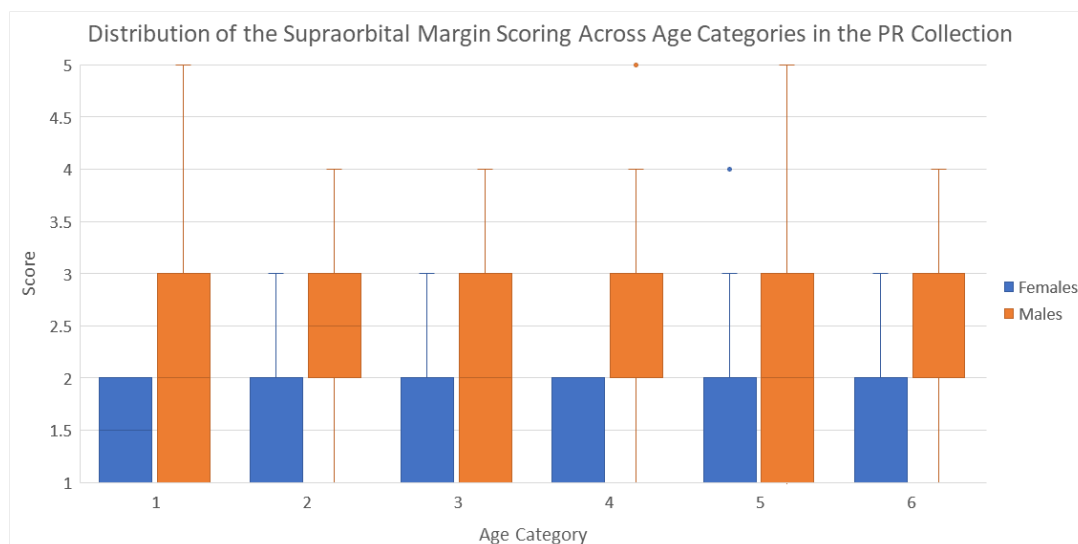
Figure F.28: A boxplot distribution of supraorbital margin scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.
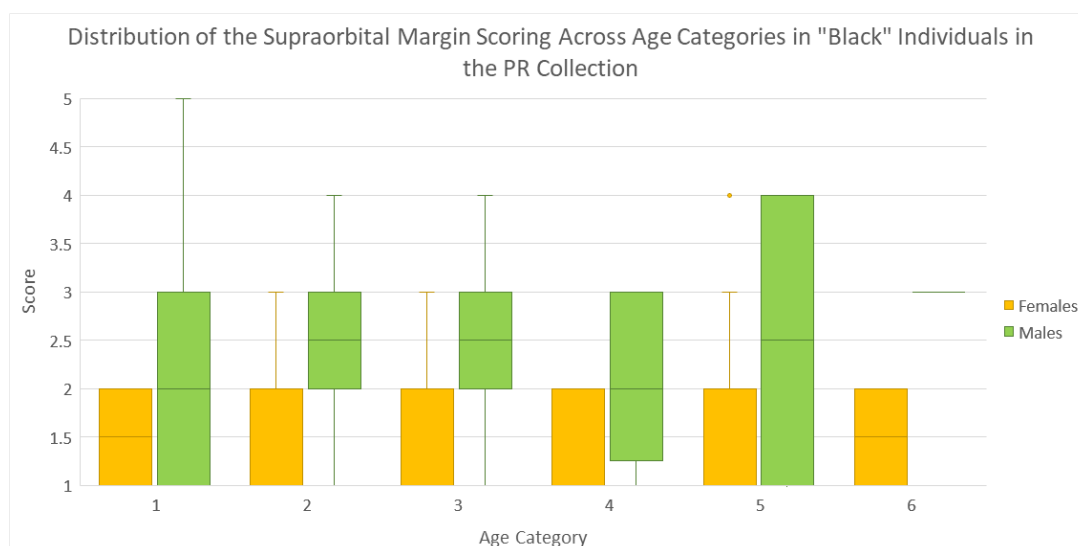
Figure F.29: A boxplot distribution of supraorbital margin scoring across different age categories for "Black" males and females. Females are given in yellow while males are in green. The age categories are defined in Table 2.2.
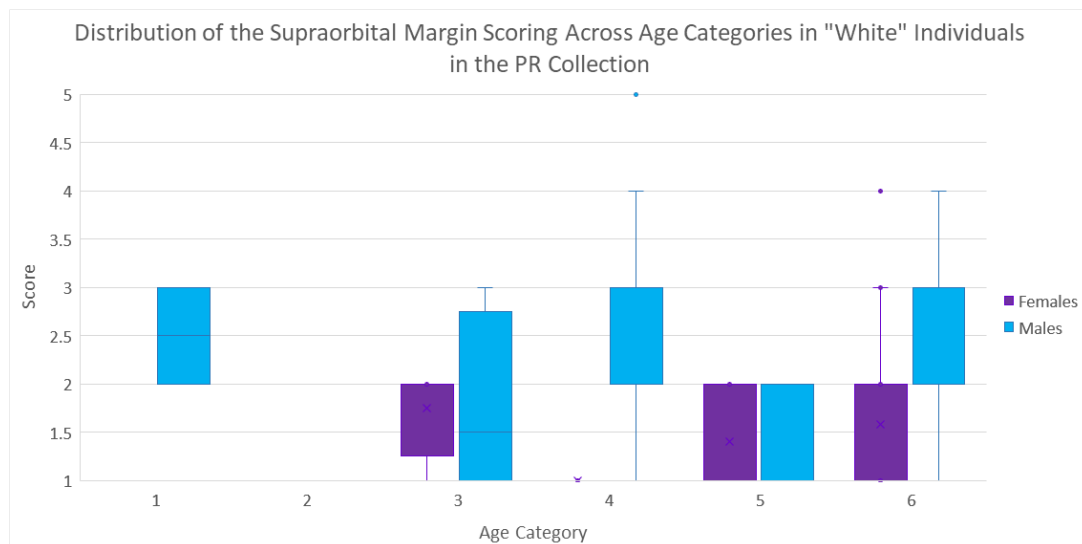
Figure F.30: A boxplot distribution of supraorbital margin scoring across different age categories for "White" males and females. Females are given in purple while males are in cyan. The age categories are defined in Table 2.2.

Table F.7: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the PR collection when comparing supraorbital margin trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 28	F = 1.5 M = 2.0	$U = 152.0$ $p = 0.115$ $z = 0.15$ $r = 0.03$	0.714
2	F = 20 M = 16	F = 2.0 M = 2.5	$U = 236.0$ $p = 0.011$ $z = -4.27$ $r = -0.71$	0.725
3	F = 40 M = 20	F = 2.0 M = 2.0	$U = 506.0$ $p = 0.075$ $z = -11.20$ $r = -1.45$	0.700
4	F = 40 M = 60	F = 1.0 M = 2.0	$U = 1902.5$ $p << 0.001$ $z = -0.83$ $r = -0.08$	0.727
5	F = 44 M = 34	F = 2.0 M = 2.0	$U = 898.5$ $p = 0.105$ $z = -8.46$ $r = -0.96$	0.674
6	F = 148 M = 140	F = 1.0 M = 3.0	$U = 16303.5$ $p << 0.001$ $z = -7.19$ $r = -0.42$	0.789

Table F.8: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "Black" individuals from the PR collection when comparing supraorbital margin trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 24	F = 1.5 M = 2.0	$U = 124.0$ $p = 0.206$ $z = -0.35$ $r = -0.06$	0.708
2	F = 20 M = 16	F = 2.0 M = 2.5	$U = 236.0$ $p = 0.011$ $z = -4.27$ $r = -0.71$	0.725
3	F = 36 M = 12	F = 2.0 M = 2.5	$U = 314.0$ $p = 0.012$ $z = -13.52$ $r = -1.95$	0.731
4	F = 32 M = 24	F = 2.0 M = 2.0	$U = 536.0$ $p = 0.006$ $z = -6.23$ $r = -0.83$	0.661
5	F = 24 M = 14	F = 2.0 M = 3.0	$U = 219.5$ $p = 0.107$ $z = -7.52$ $r = -1.22$	0.783
6	F = 8 M = 8	F = 1.5 M = 3.0	$U = 64.0$ $p < 0.001$ $z = -0.42$ $r = -0.11$	1.000

Table F.9: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "White" individuals from the PR collection when comparing supraorbital margin trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 0￼ M = 4	F = N/A￼ M = 2.5	$U = N/A$￼ $p = N/A$￼ $z = N/A$￼ $r = N/A$	N/A
3	F = 4￼ M = 8	F = 2.0￼ M = 1.5	$U = 15.0$￼ $p = 0.927$￼ $z = -1.87$￼ $r = -0.54$	0.688
4	F = 8￼ M = 36	F = 1.0￼ M = 2.0	$U = 272.0$￼ $p \ll 0.001$￼ $z = 2.80$￼ $r = 0.42$	0.889
5	F = 20￼ M = 20	F = 1.0￼ M = 2.0	$U = 248.0$￼ $p = 0.147$￼ $z = -4.38$￼ $r = -0.69$	0.560
6	F = 140￼ M = 132	F = 1.0￼ M = 3.0	$U = 14365.5$￼ $p \ll 0.001$￼ $z = -7.32$￼ $r = -0.44$	0.787
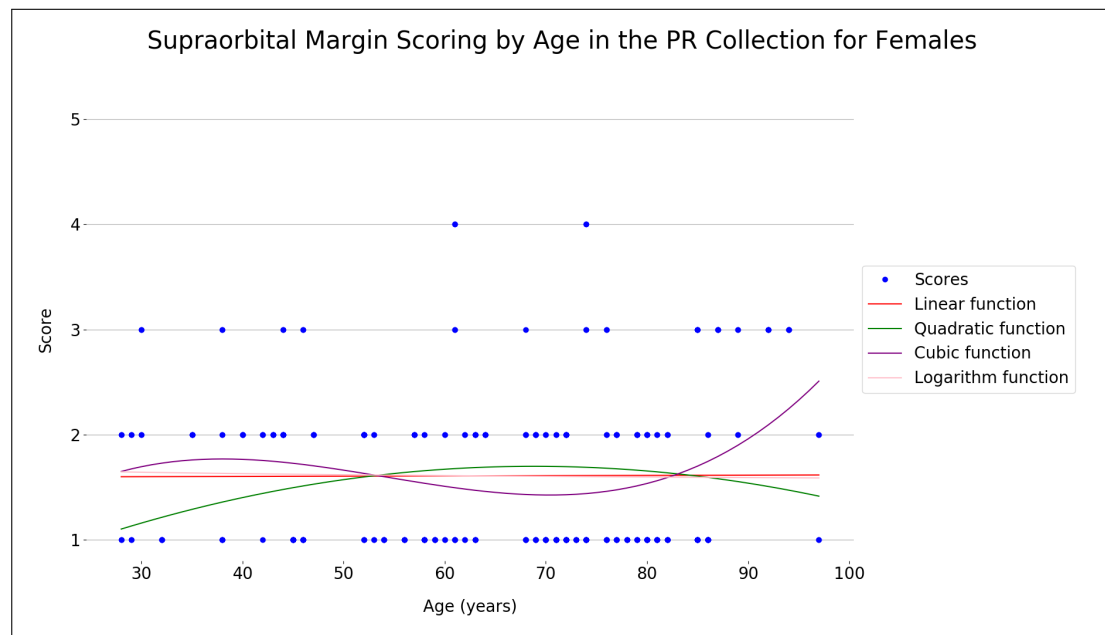
Figure F.31: A scatterplot of age vs. supraorbital margin trait scoring for females in the PR collection, with four fitting functions.
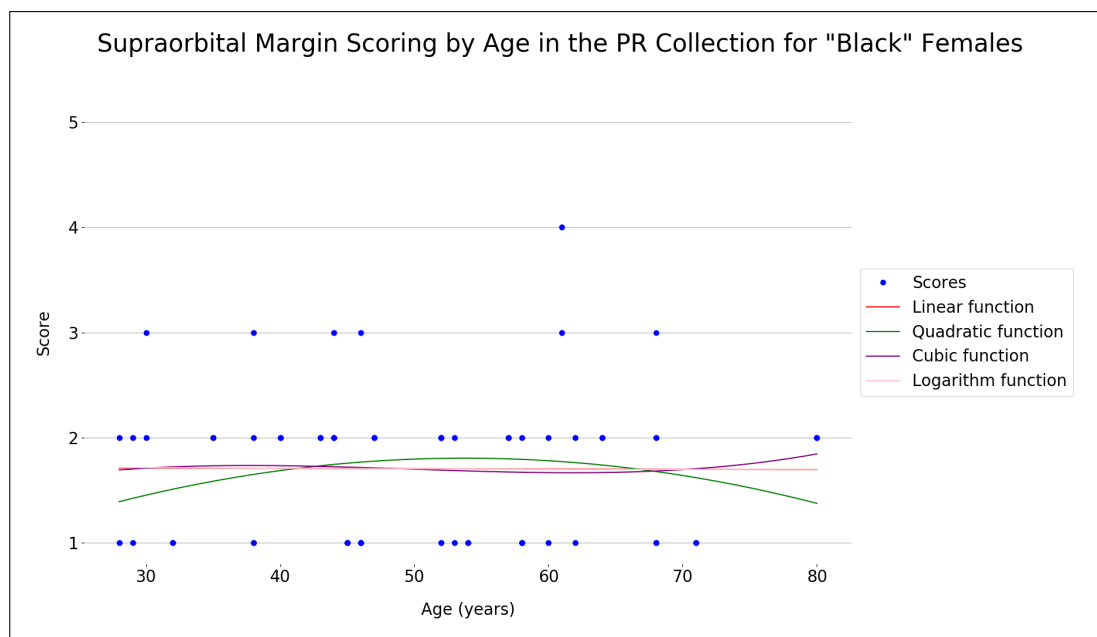
Figure F.32: A scatterplot of age vs. supraorbital margin trait scoring for "Black" females in the PR collection, with four fitting functions.

Figure F.33: A scatterplot of age vs. supraorbital margin trait scoring for "White" females in the PR collection, with four fitting functions.
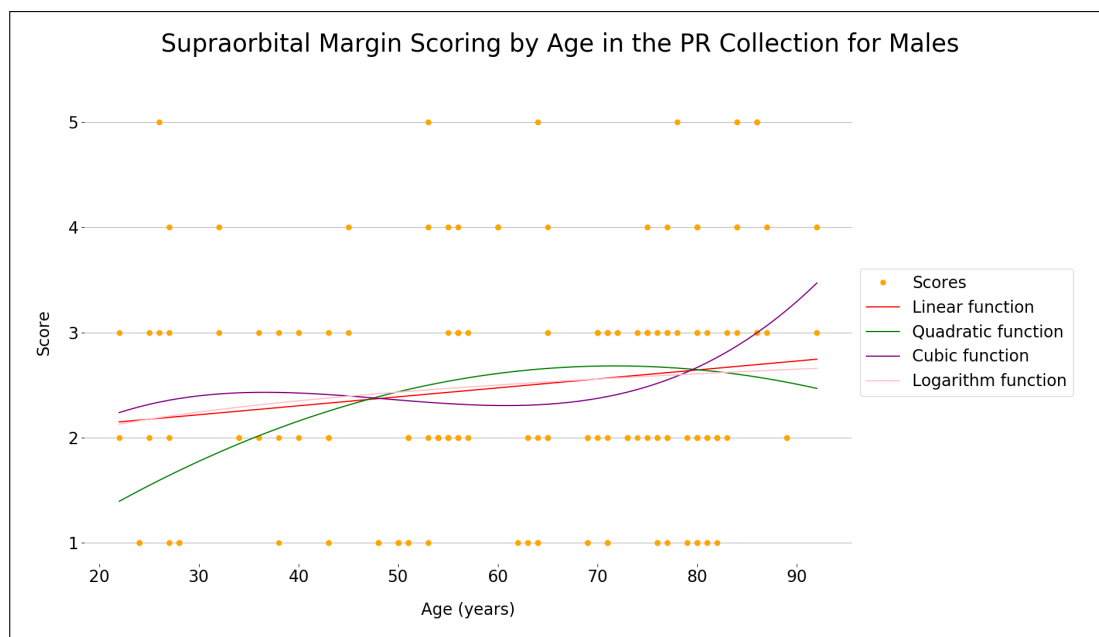
Figure F.34: A scatterplot of age vs. supraorbital margin trait scoring for males in the PR collection, with four fitting functions.
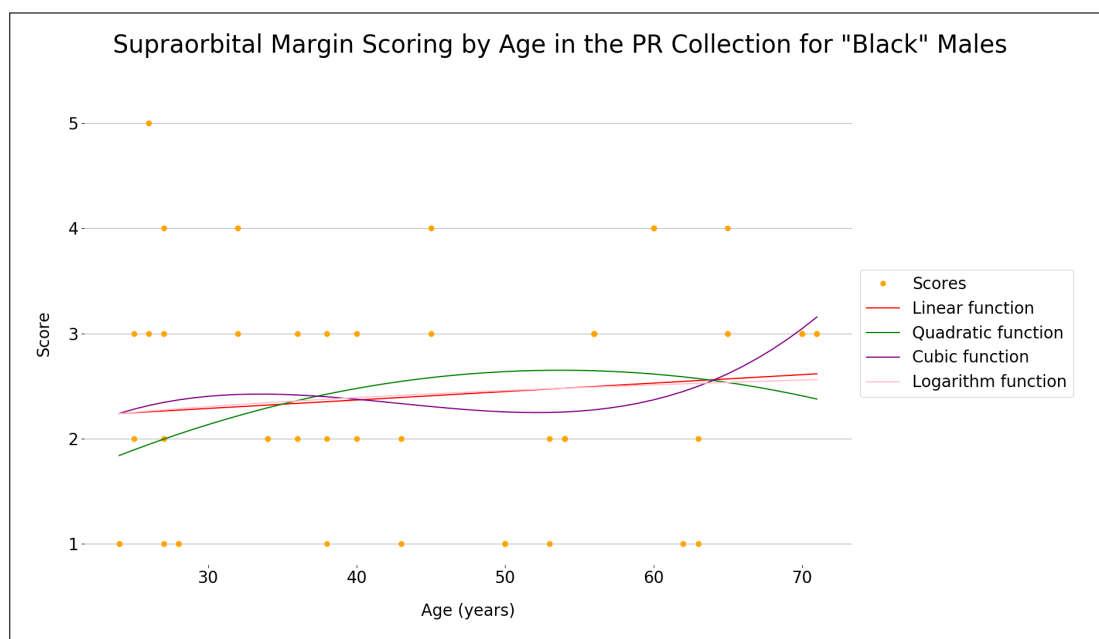
Figure F.35: A scatterplot of age vs. supraorbital margin trait scoring for "Black" males in the PR collection, with four fitting functions.
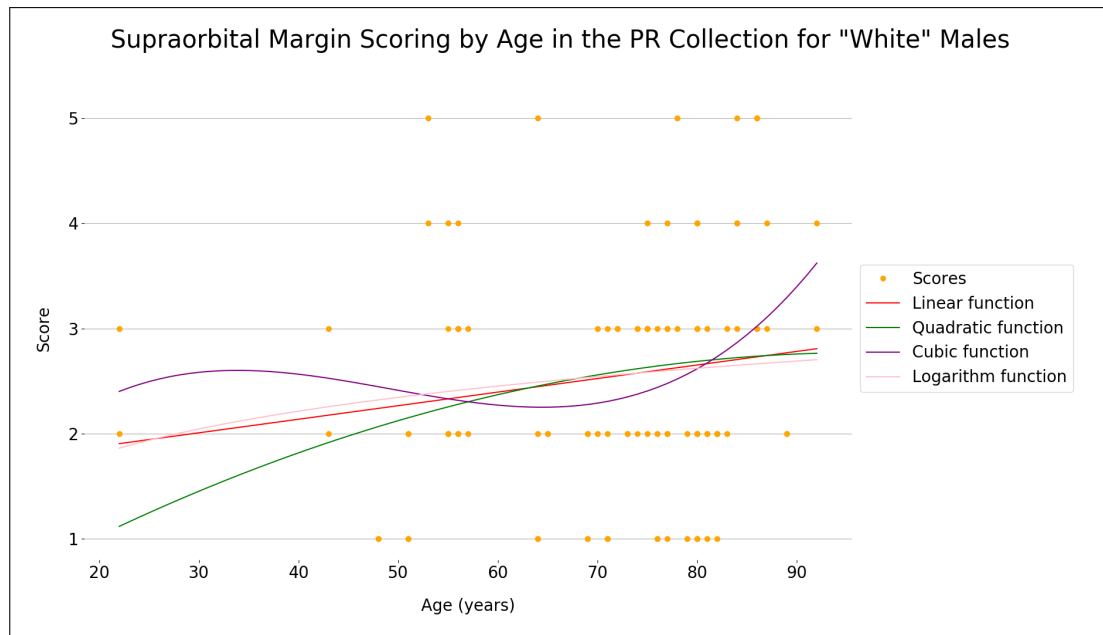
Figure F.36: A scatterplot of age vs. supraorbital margin trait scoring for "White" males in the PR collection, with four fitting functions.

F.4 Glabella
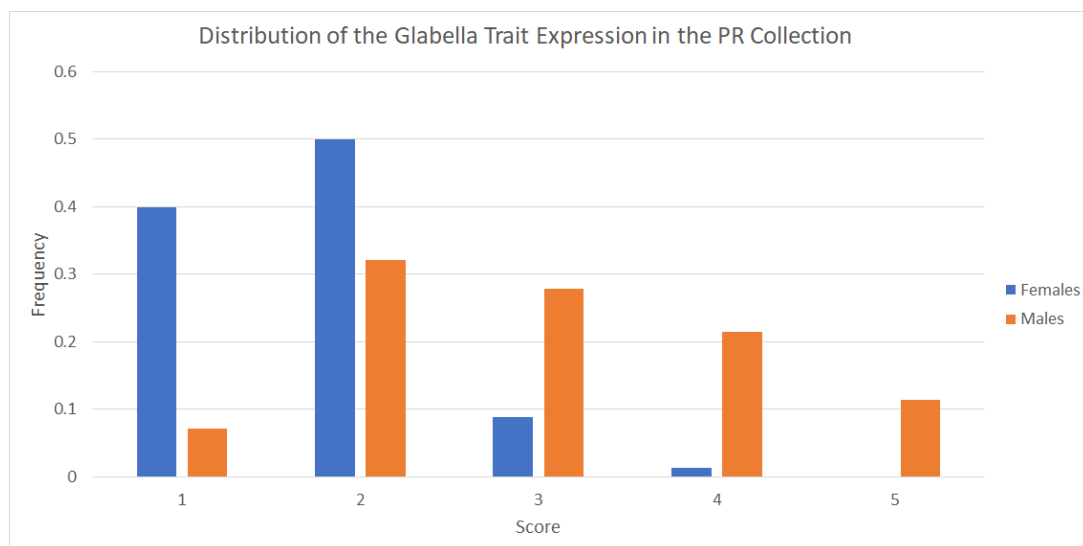
Figure F.37: The distribution of the glabella trait expression in the PR Collection represented using a bar chart. Females are in blue while males are in orange.
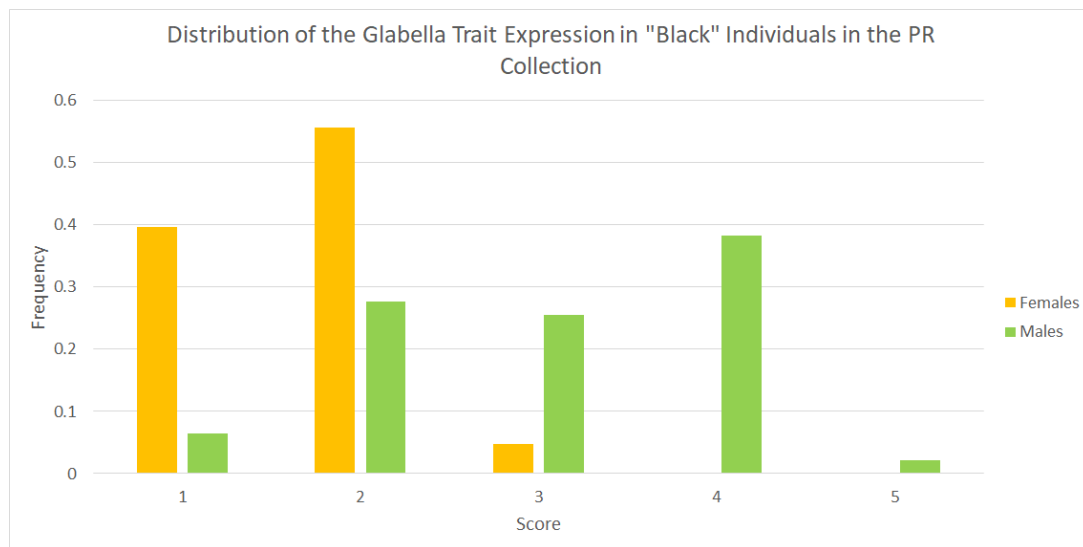
Figure F.38: The distribution of the glabella trait expression in "Black" individuals from the PR Collection represented using a bar chart. Females are in yellow while males are in green.
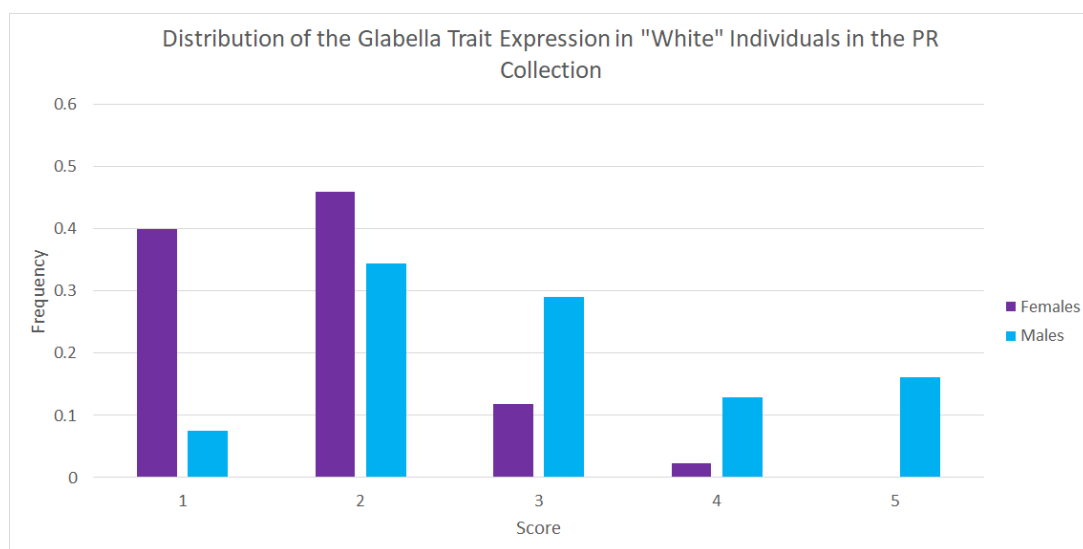
Figure F.39: The distribution of the glabella trait expression in "White" individuals from the PR Collection represented using a bar chart. Females are in purple while males are in cyan.
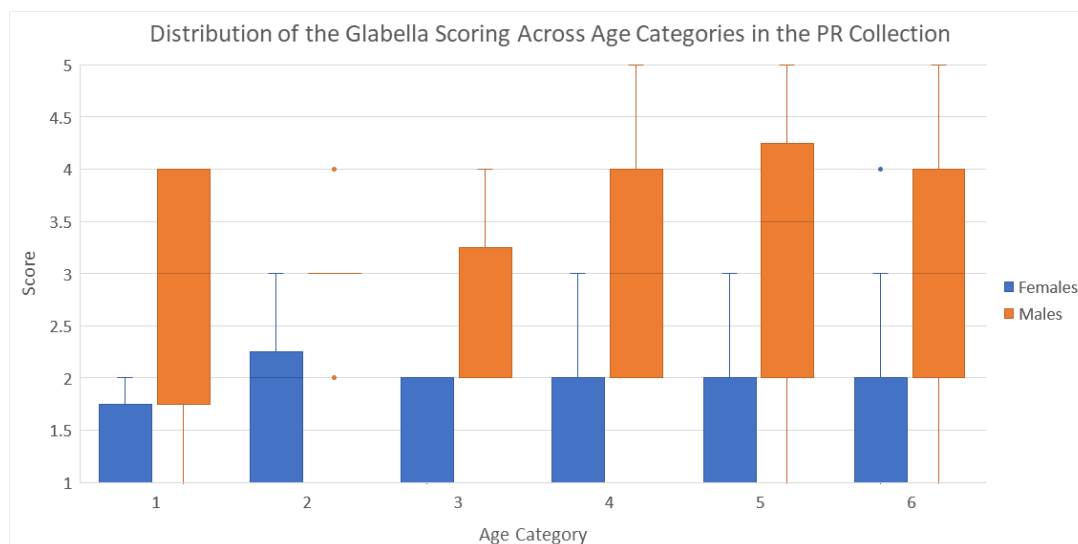
Figure F.40: A boxplot distribution of glabella scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.
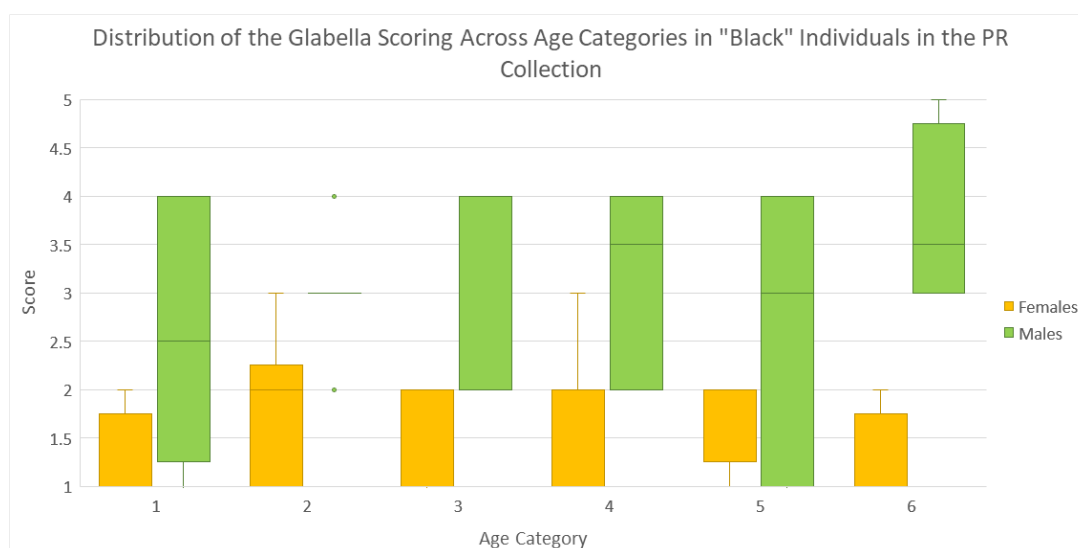
Figure F.41: A boxplot distribution of glabella scoring across different age categories for "Black" males and females. Females are given in yellow while males are in green. The age categories are defined in Table 2.2.
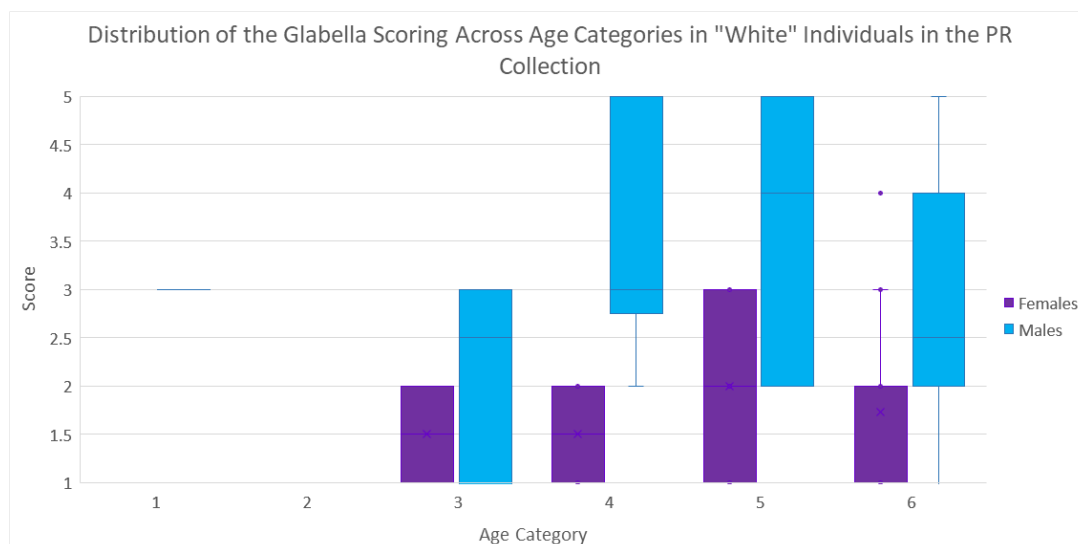
Figure F.42: A boxplot distribution of glabella scoring across different age categories for "White" males and females. Females are given in purple while males are in cyan. The age categories are defined in Table 2.2.

Table F.10: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the PR collection when comparing glabella trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 12	F = 1.0 M = 3.0	$U = 44.0$ $p = 0.014$ $z = 1.21$ $r = 0.30$	0.875
2	F = 10 M = 8	F = 2.0 M = 3.0	$U = 70.0$ $p = 0.005$ $z = -2.22$ $r = -0.52$	0.800
3	F = 19 M = 9	F = 2.0 M = 2.0	$U = 143.5$ $p = 0.001$ $z = -6.49$ $r = -1.23$	0.678
4	F = 20 M = 30	F = 2.0 M = 3.0	$U = 533.0$ $p << 0.001$ $z = 0.46$ $r = 0.06$	0.807
5	F = 22 M = 17	F = 2.0 M = 4.0	$U = 295.0$ $p = 0.001$ $z = -4.11$ $r = -0.66$	0.813
6	F = 73 M = 67	F = 2.0 M = 3.0	$U = 3791.0$ $p << 0.001$ $z = -5.65$ $r = -0.48$	0.770

Table F.11: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "Black" individuals from the PR collection when comparing glabella trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 10	F = 1.0 M = 3.5	$U = 36.0$ $p = 0.022$ $z = 0.85$ $r = 0.23$	0.850
2	F = 10 M = 8	F = 2.0 M = 3.0	$U = 70.0$ $p = 0.005$ $z = -2.22$ $r = -0.52$	0.800
3	F = 17 M = 6	F = 2.0 M = 2.0	$U = 82.0$ $p = 0.013$ $z = -8.54$ $r = -1.78$	0.608
4	F = 16 M = 12	F = 2.0 M = 3.5	$U = 161.5$ $p < 0.001$ $z = -3.27$ $r = -0.62$	0.734
5	F = 12 M = 7	F = 2.0 M = 3.0	$U = 63.0$ $p = 0.064$ $z = -4.82$ $r = -1.11$	0.929
6	F = 4 M = 4	F = 1.0 M = 3.5	$U = 16.0$ $p = 0.026$ $z = -0.58$ $r = -0.20$	1.000

Table F.12: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "White" individuals from the PR collection when comparing glabella trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 0 M = 2	F = *N/A* M = 3.0	$U = $ *N/A* $p = $ *N/A* $z = $ *N/A* $r = $ *N/A*	*N/A*
3	F = 2 M = 3	F = 1.5 M = 3.0	$U = 5.5$ $p = 0.224$ $z = $ -0.29 $r = $ -0.13	0.833
4	F = 4 M = 18	F = 1.5 M = 3.0	$U = 68.0$ $p = 0.006$ $z = 1.87$ $r = 0.40$	0.889
5	F = 10 M = 10	F = 2.0 M = 4.0	$U = 80.0$ $p = 0.020$ $z = $ -1.89 $r = $ -0.42	0.840
6	F = 69 M = 63	F = 2.0 M = 3.0	$U = 3288.0$ $p << 0.001$ $z = $ -5.92 $r = $ -0.52	0.755
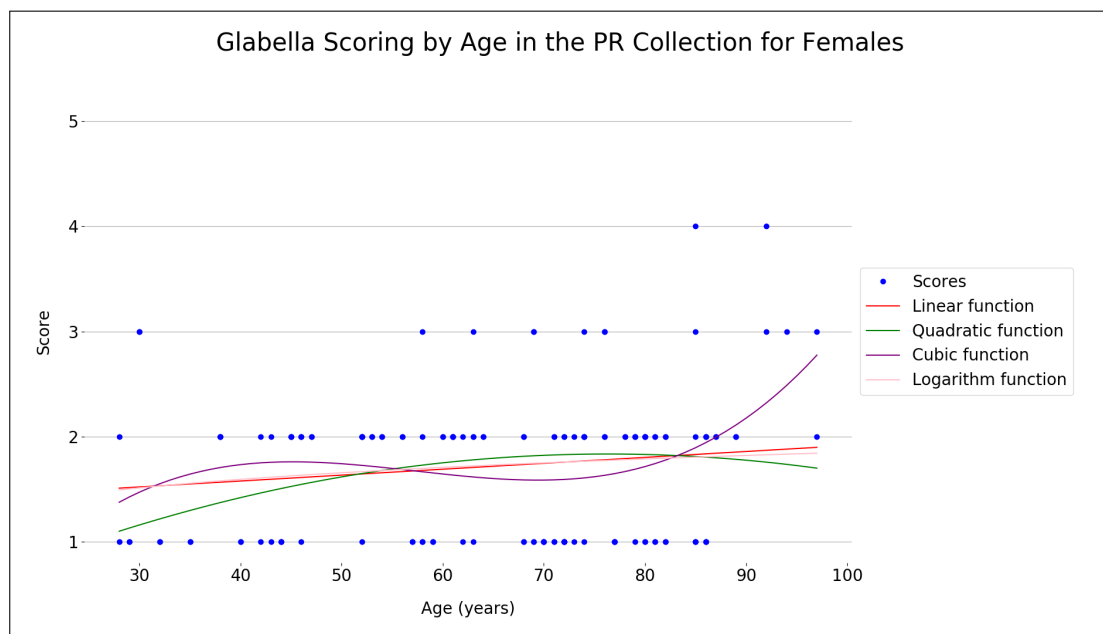
Figure F.43: A scatterplot of age vs. glabella trait scoring for females in the PR collection, with four fitting functions.
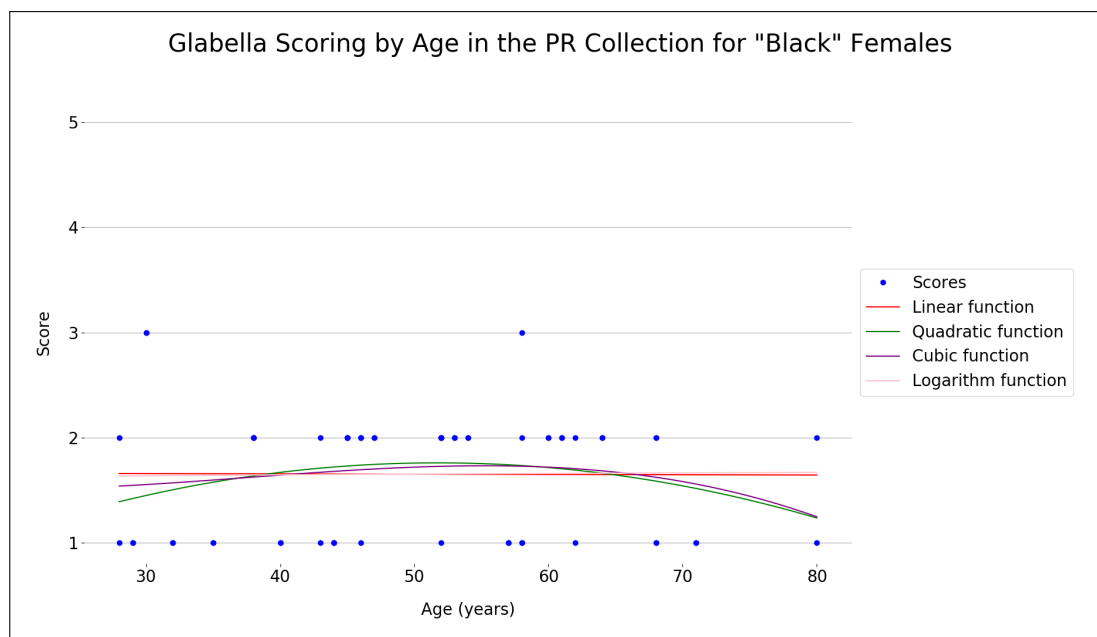
Figure F.44: A scatterplot of age vs. glabella trait scoring for "Black" females in the PR collection, with four fitting functions.
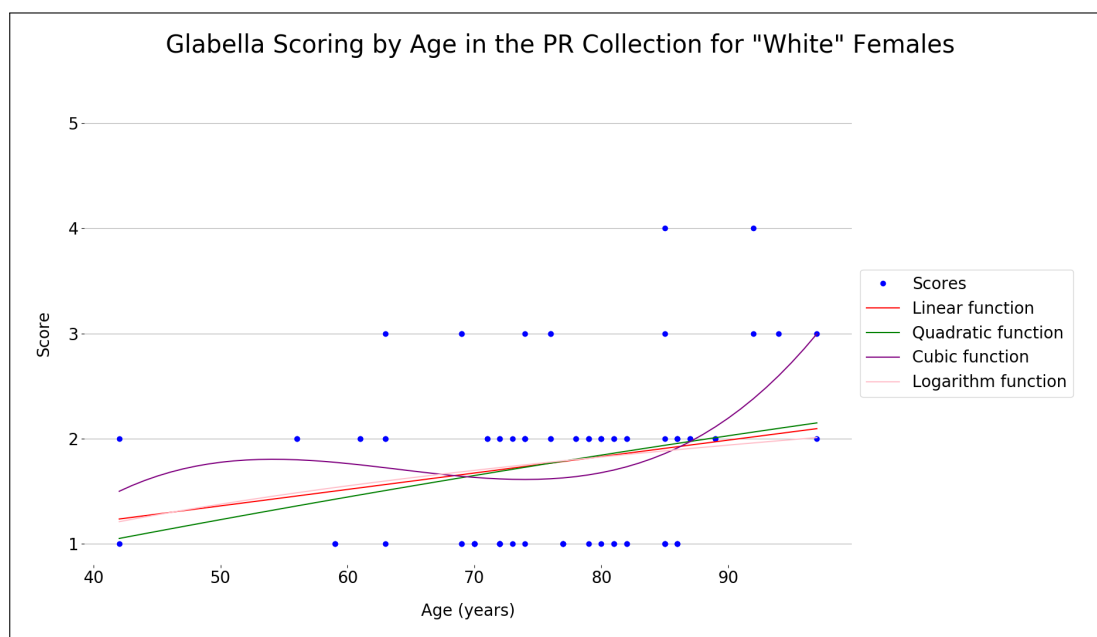
Figure F.45: A scatterplot of age vs. glabella trait scoring for "White" females in the PR collection, with four fitting functions.
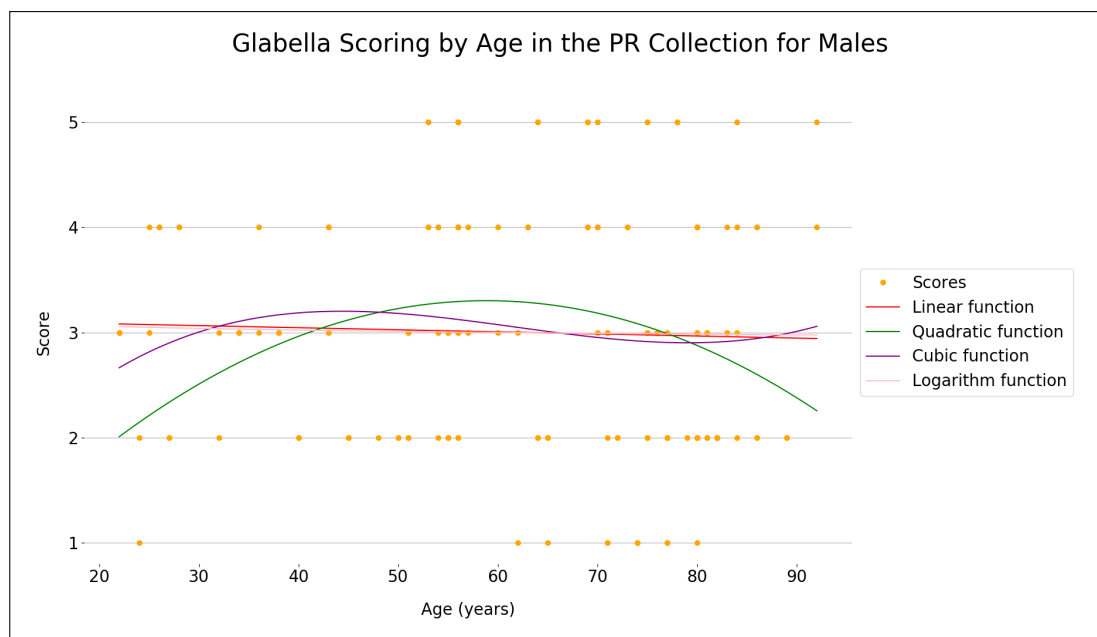
Figure F.46: A scatterplot of age vs. glabella trait scoring for males in the PR collection, with four fitting functions.
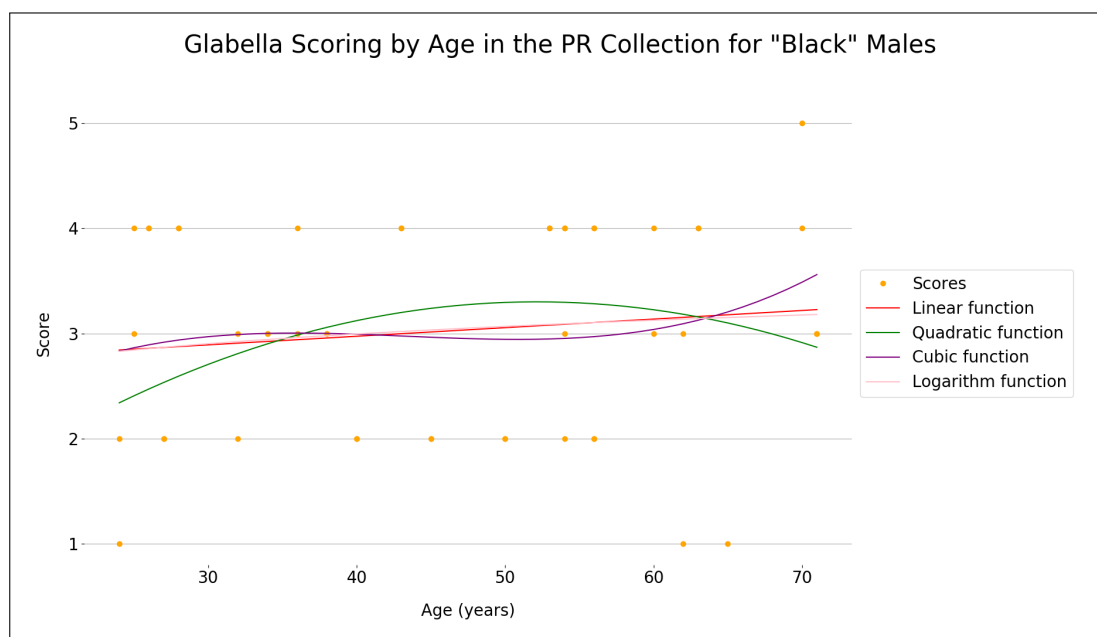
Figure F.47: A scatterplot of age vs. glabella trait scoring for "Black" males in the PR collection, with four fitting functions.
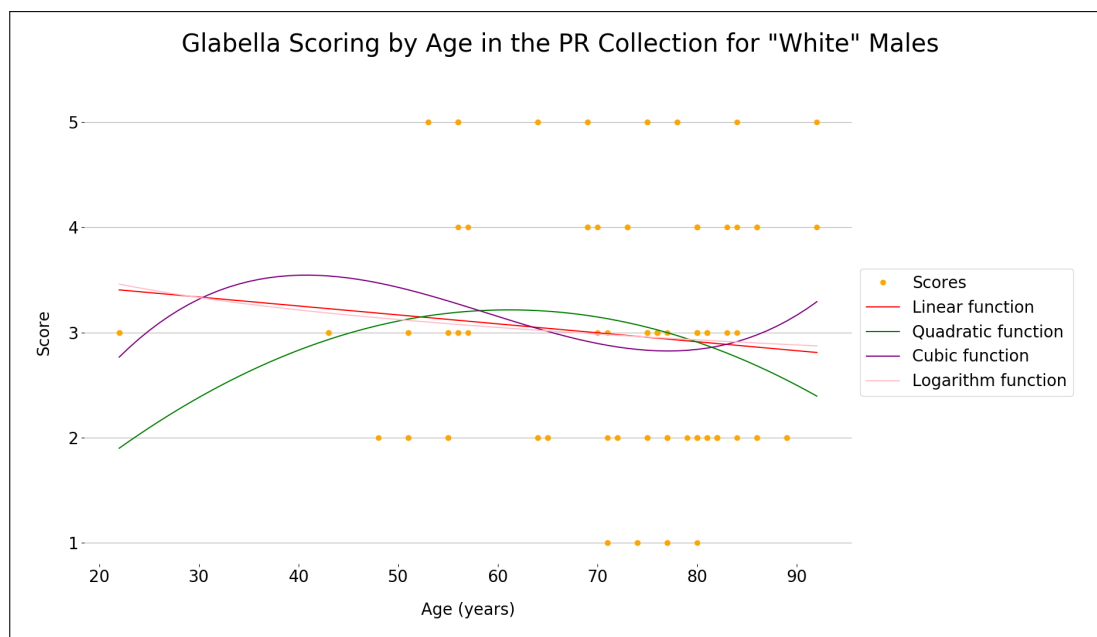
Figure F.48: A scatterplot of age vs. glabella trait scoring for "White" males in the PR collection, with four fitting functions.

F.5 Zygomatic Extension
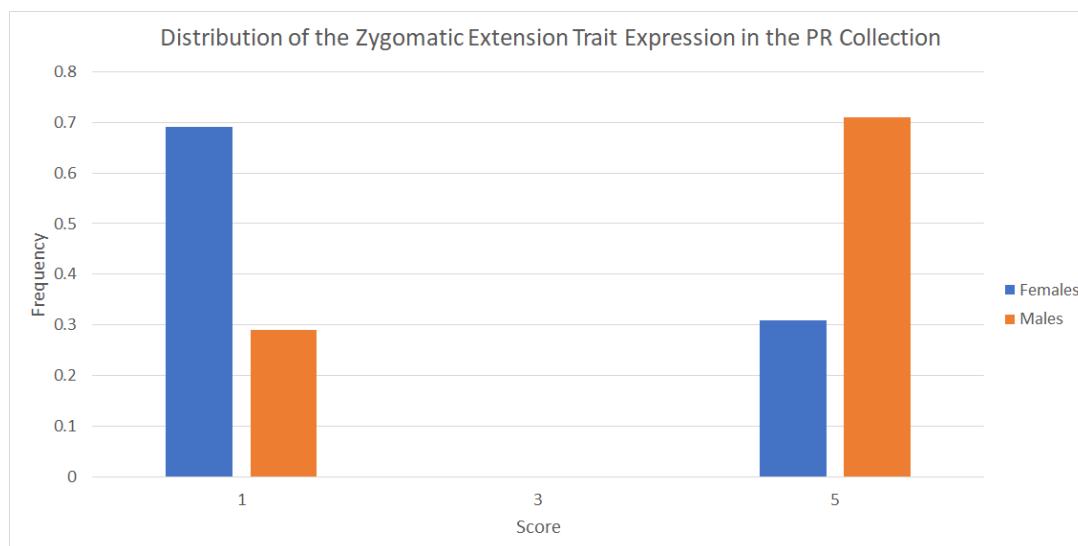
Figure F.49: The distribution of the zygomatic extension trait expression in the PR Collection represented using a bar chart. Females are in blue while males are in orange.
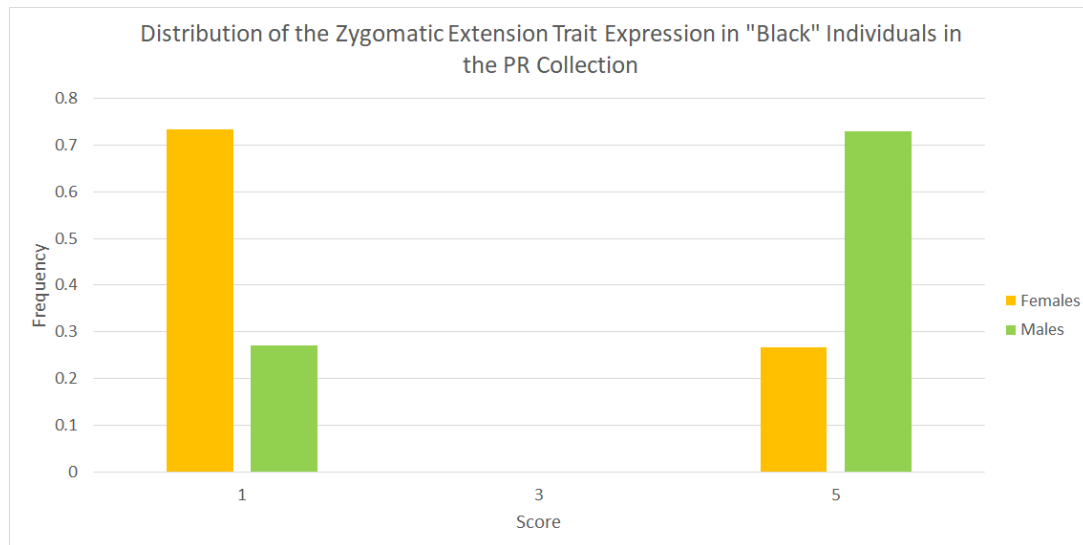
Figure F.50: The distribution of the zygomatic extension trait expression in "Black" individuals from the PR Collection represented using a bar chart. Females are in yellow while males are in green.
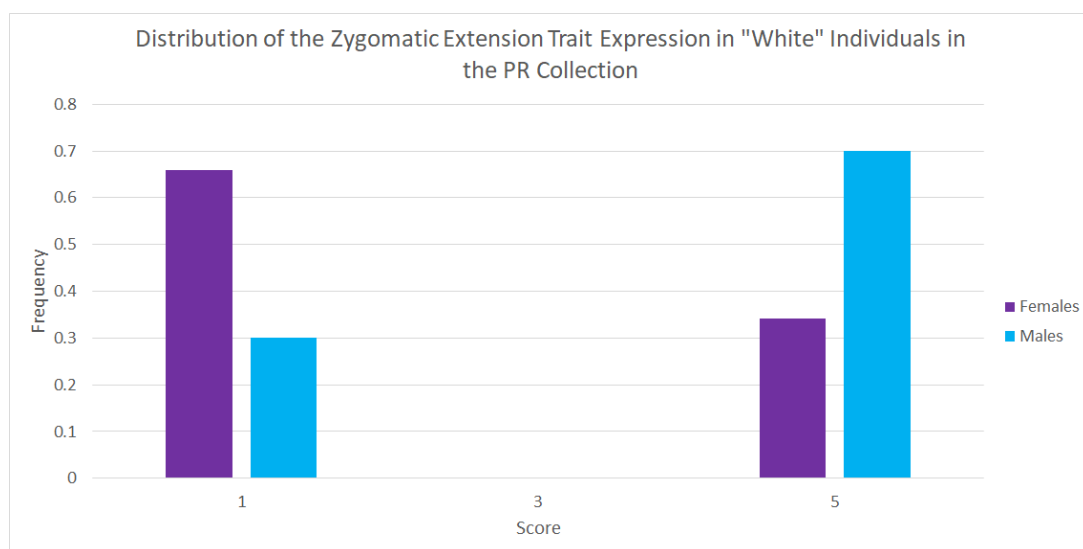
Figure F.51: The distribution of the zygomatic extension trait expression in "White" individuals from the PR Collection represented using a bar chart. Females are in purple while males are in cyan.
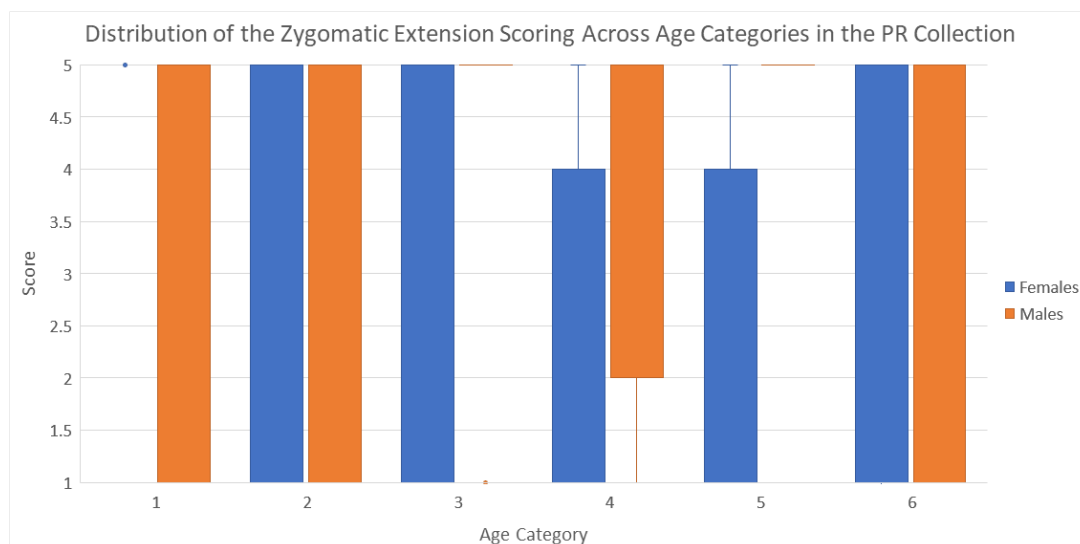
Figure F.52: A boxplot distribution of zygomatic extension scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.
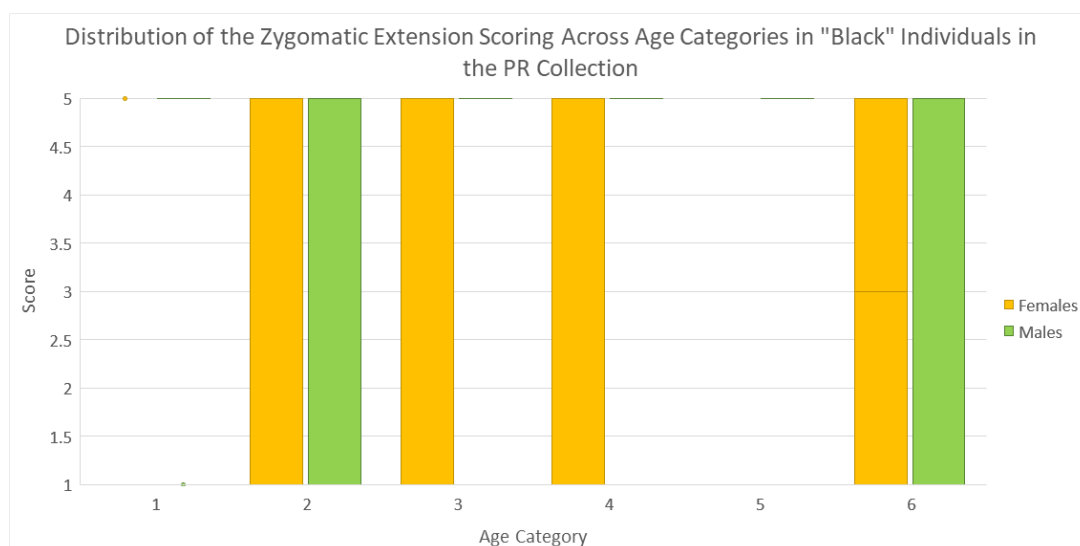
Figure F.53: A boxplot distribution of zygomatic extension scoring across different age categories for "Black" males and females. Females are given in yellow while males are in green. The age categories are defined in Table 2.2.
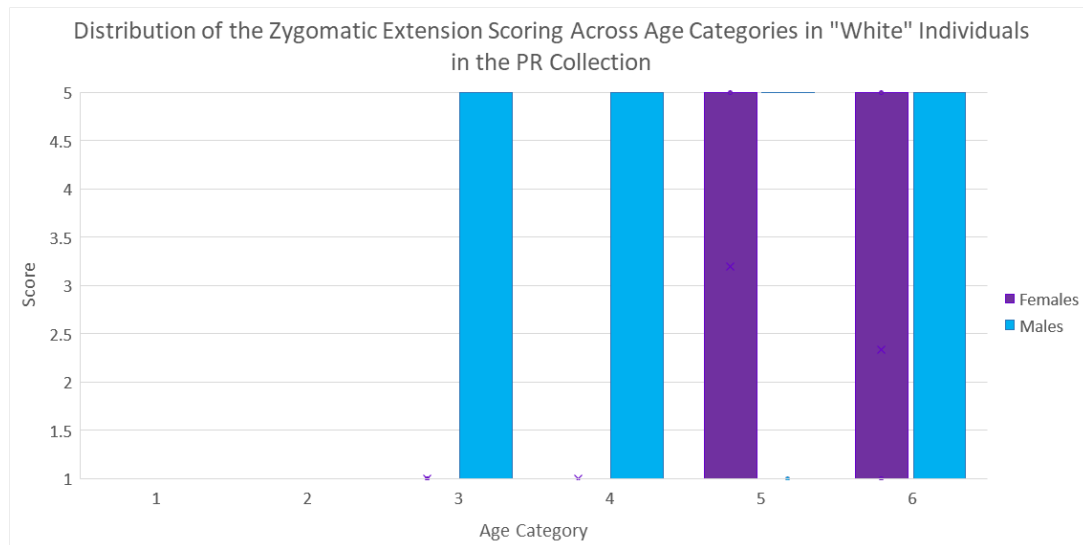
Figure F.54: A boxplot distribution of zygomatic extension scoring across different age categories for "White" males and females. Females are given in purple while males are in cyan. The age categories are defined in Table 2.2.

Table F.13: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the PR collection when comparing zygomatic extension trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 28	F = 1.0 M = 5.0	$U = 174.0$ $p = 0.007$ $z = 0.99$ $r = 0.16$	0.634
2	F = 20 M = 16	F = 1.0 M = 1.0	$U = 162.0$ $p = 0.952$ $z = -6.62$ $r = -1.10$	0.425
3	F = 40 M = 20	F = 1.0 M = 5.0	$U = 590.0$ $p < 0.001$ $z = -9.88$ $r = -1.28$	0.605
4	F = 40 M = 60	F = 1.0 M = 5.0	$U = 1800.0$ $p \ll 0.001$ $z = -1.55$ $r = -0.15$	0.625
5	F = 44 M = 36	F = 1.0 M = 5.0	$U = 1298.0$ $p \ll 0.001$ $z = -4.68$ $r = -0.52$	0.694
6	F = 146 M = 140	F = 1.0 M = 5.0	$U = 13658.0$ $p \ll 0.001$ $z = -10.43$ $r = -0.62$	0.556

Table F.14: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "Black" individuals from the PR collection when comparing zygomatic extension trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 8 M = 24	F = 1.0 M = 5.0	$U = 160.0$ $p < 0.001$ $z = 1.22$ $r = 0.22$	0.719
2	F = 20 M = 16	F = 1.0 M = 1.0	$U = 162.0$ $p = 0.952$ $z = -6.62$ $r = -1.10$	0.425
3	F = 36 M = 12	F = 1.0 M = 5.0	$U = 336.0$ $p = 0.001$ $z = -13.00$ $r = -1.88$	0.616
4	F = 32 M = 24	F = 1.0 M = 5.0	$U = 584.0$ $p < 0.001$ $z = -5.43$ $r = -0.73$	0.625
5	F = 24 M = 16	F = 1.0 M = 5.0	$U = 372.0$ $p << 0.001$ $z = -3.31$ $r = -0.52$	0.938
6	F = 8 M = 8	F = 3.0 M = 1.0	$U = 28.0$ $p = 0.669$ $z = -4.20$ $r = -1.05$	0.500

Table F.15: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "White" individuals from the PR collection when comparing zygomatic extension trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 0 M = 4	F = N/A M = 1.0	$U = N/A$ $p = N/A$ $z = N/A$ $r = N/A$	N/A
3	F = 4 M = 8	F = 1.0 M = 5.0	$U = 26.0$ $p = 0.060$ $z = 0.00$ $r = 0.00$	0.625
4	F = 8 M = 36	F = 1.0 M = 5.0	$U = 244.0$ $p < 0.001$ $z = 1.95$ $r = 0.29$	0.694
5	F = 20 M = 20	F = 5.0 M = 5.0	$U = 260.0$ $p = 0.043$ $z = -4.06$ $r = -0.64$	0.465
6	F = 138 M = 132	F = 1.0 M = 5.0	$U = 12423.0$ $p \ll 0.001$ $z = -9.79$ $r = -0.60$	0.565
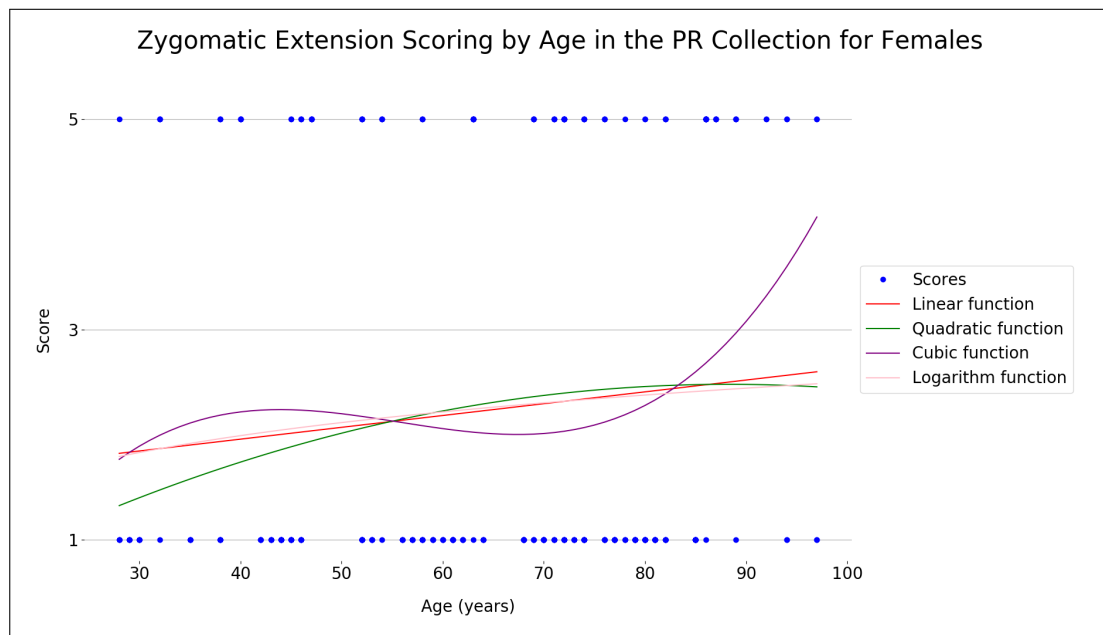
Figure F.55: A scatterplot of age vs. zygomatic extension trait scoring for females in the PR collection, with four fitting functions.
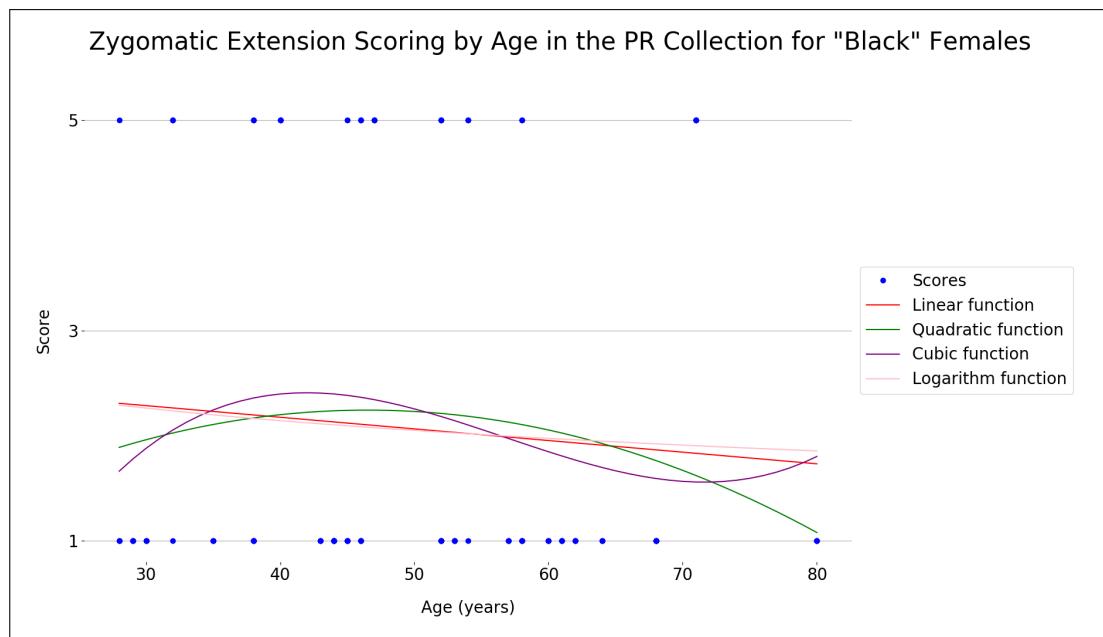
Figure F.56: A scatterplot of age vs. zygomatic extension trait scoring for "Black" females in the PR collection, with four fitting functions.
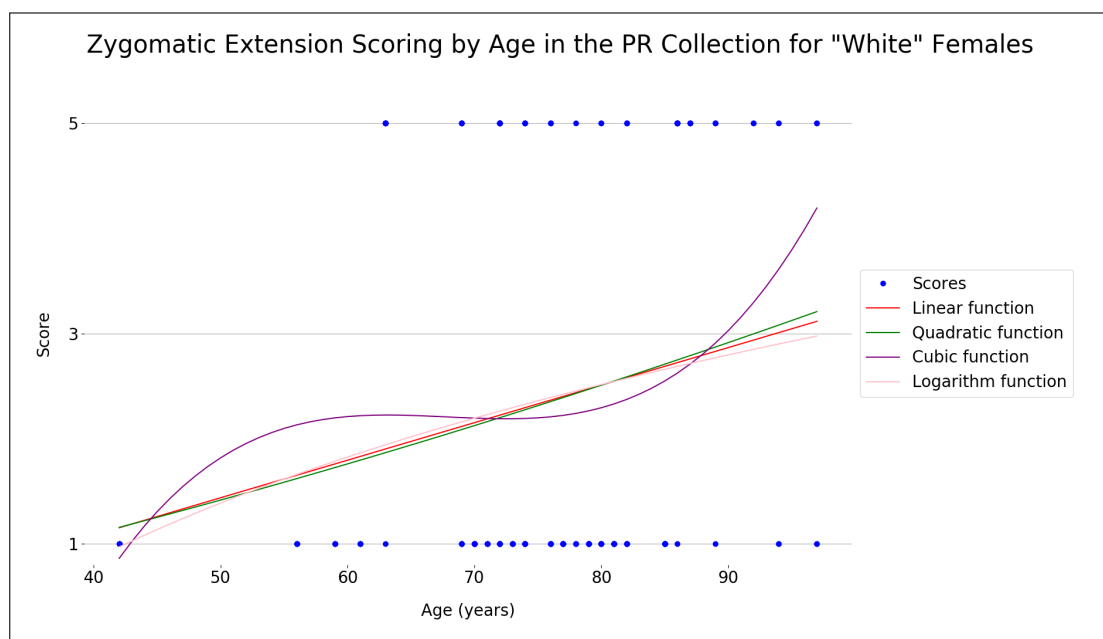
Figure F.57: A scatterplot of age vs. zygomatic extension trait scoring for "White" females in the PR collection, with four fitting functions.
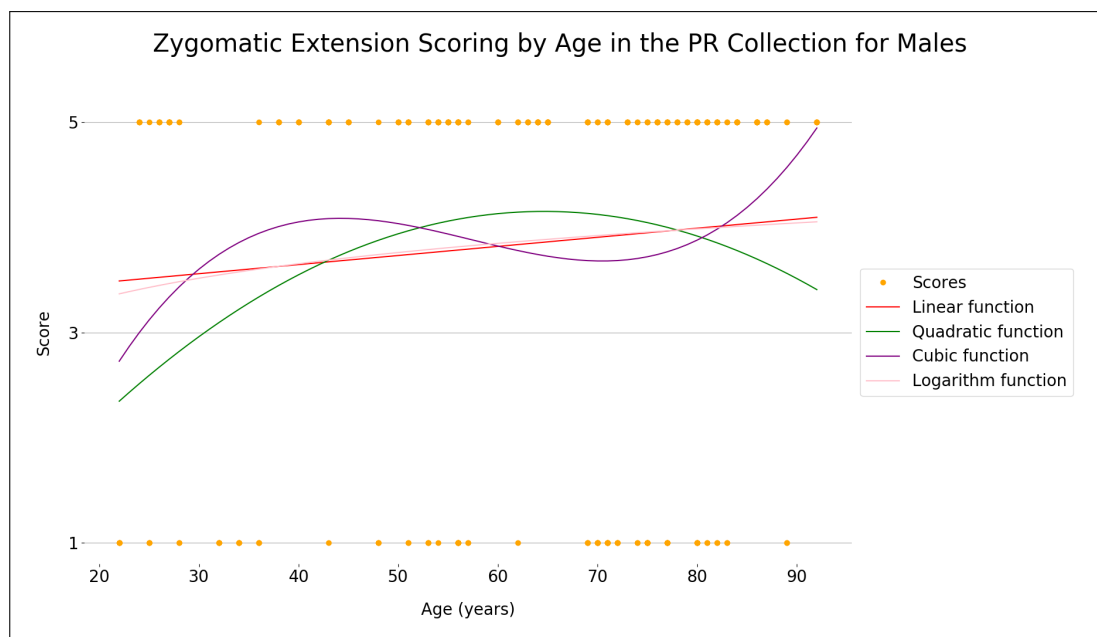
Figure F.58: A scatterplot of age vs. zygomatic extension trait scoring for males in the PR collection, with four fitting functions.
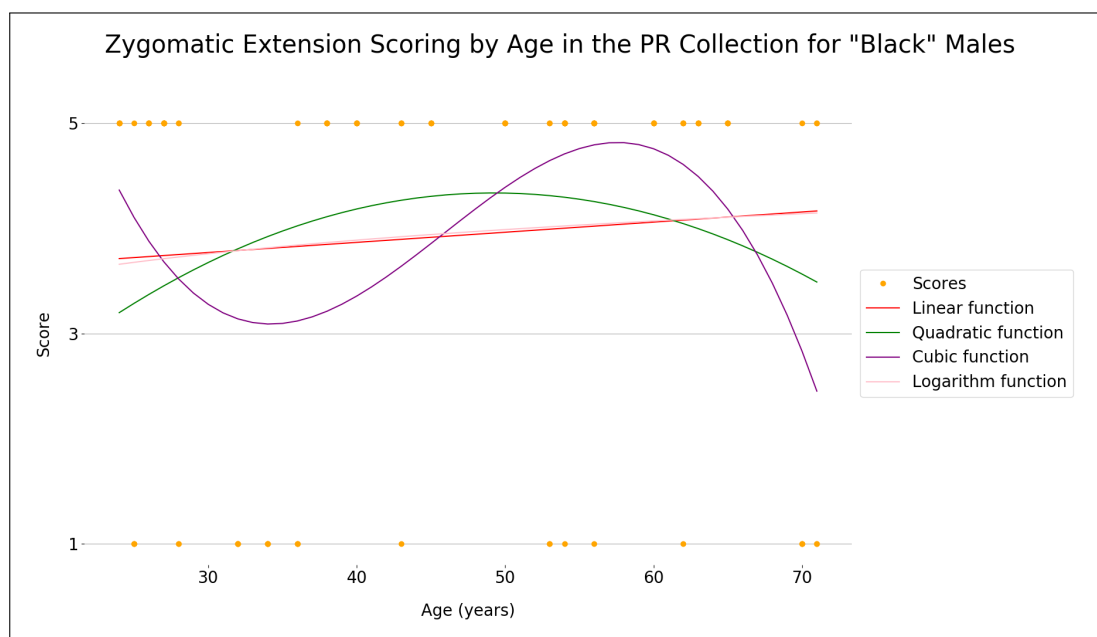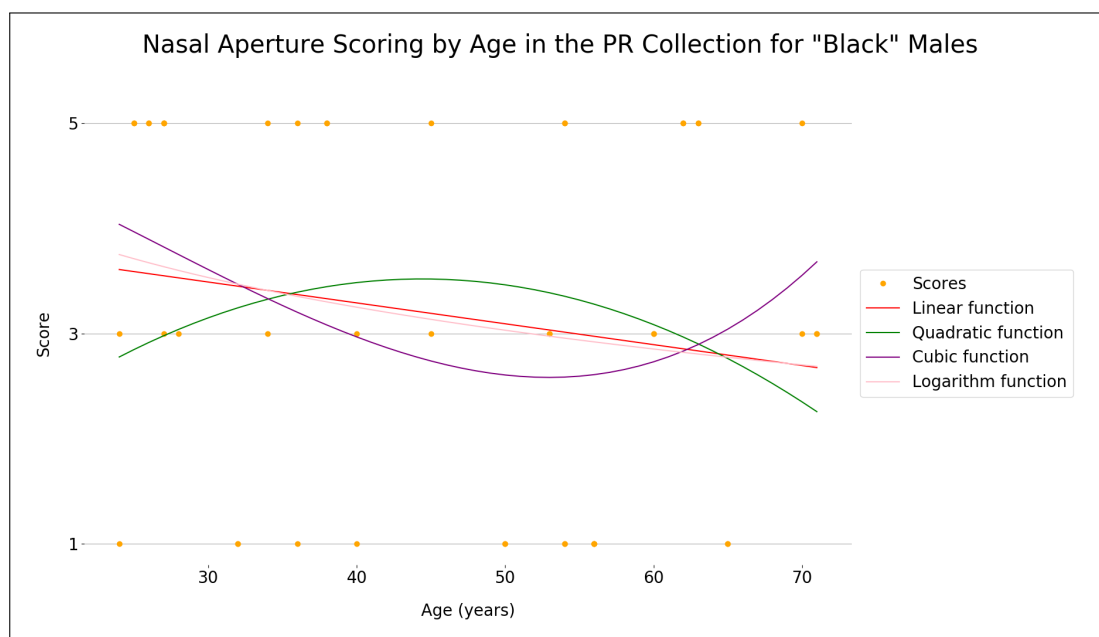

Figure F.59: A scatterplot of age vs. zygomatic extension trait scoring for "Black" males in the PR collection, with four fitting functions.
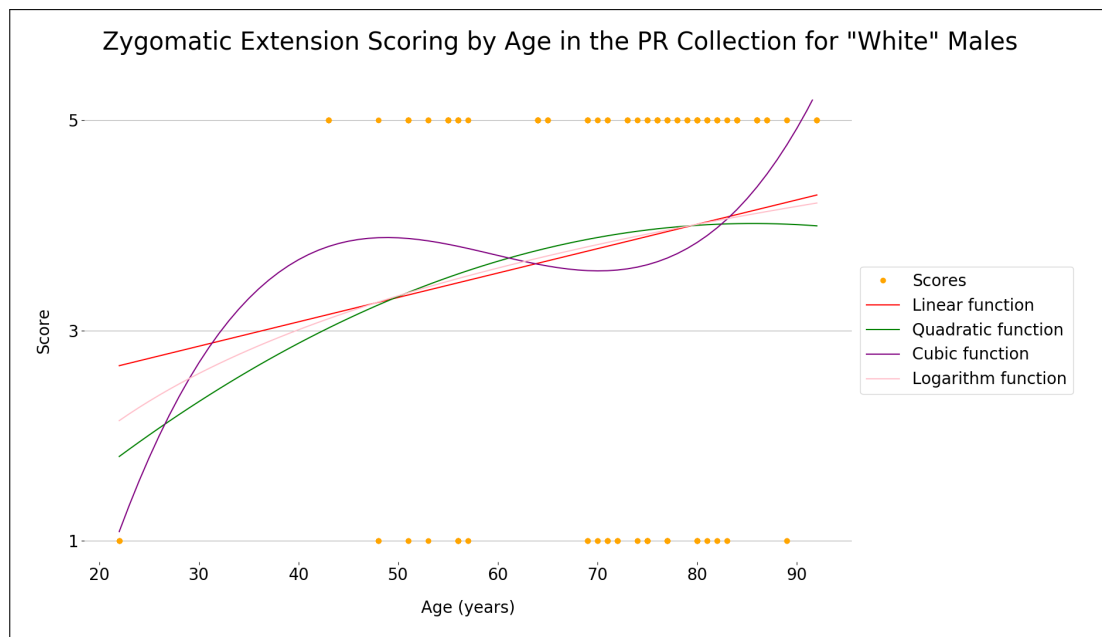
Figure F.60: A scatterplot of age vs. zygomatic extension trait scoring for "White" males in the PR collection, with four fitting functions.

F.6 Nasal Aperture
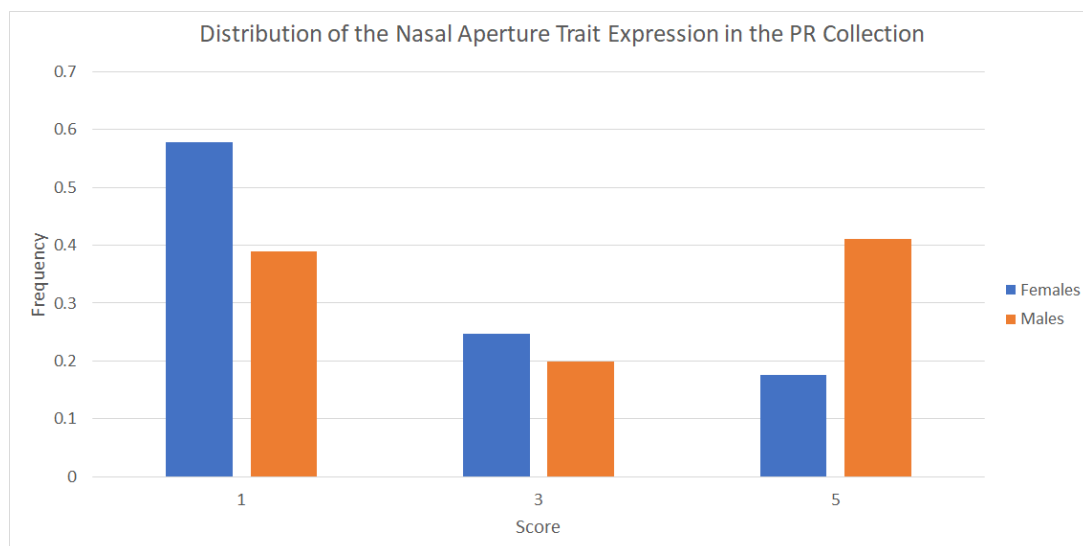
Figure F.61: The distribution of the nasal aperture trait expression in the PR Collection represented using a bar chart. Females are in blue while males are in orange.
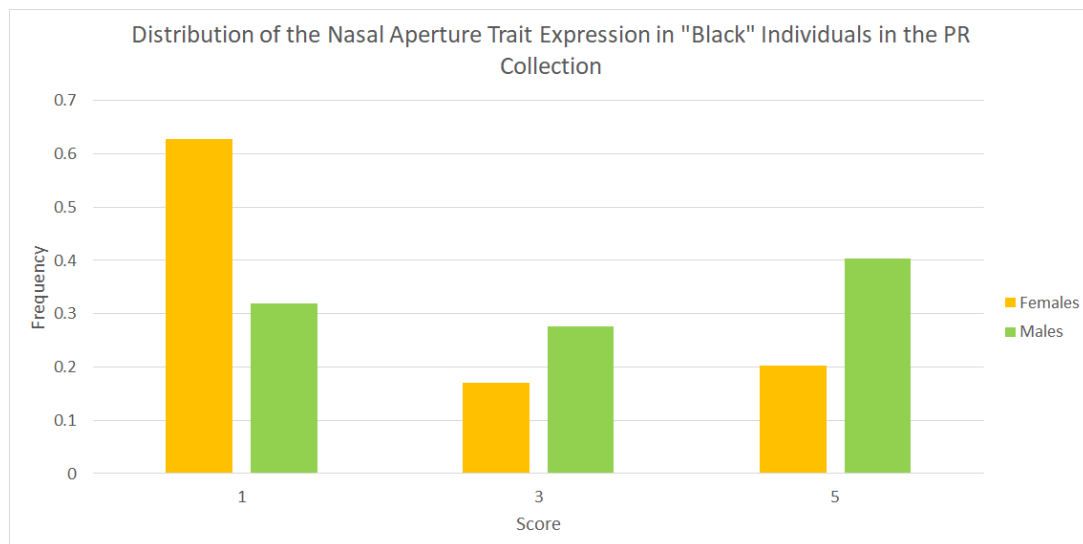
Figure F.62: The distribution of the nasal aperture trait expression in "Black" individuals from the PR Collection represented using a bar chart. Females are in yellow while males are in green.
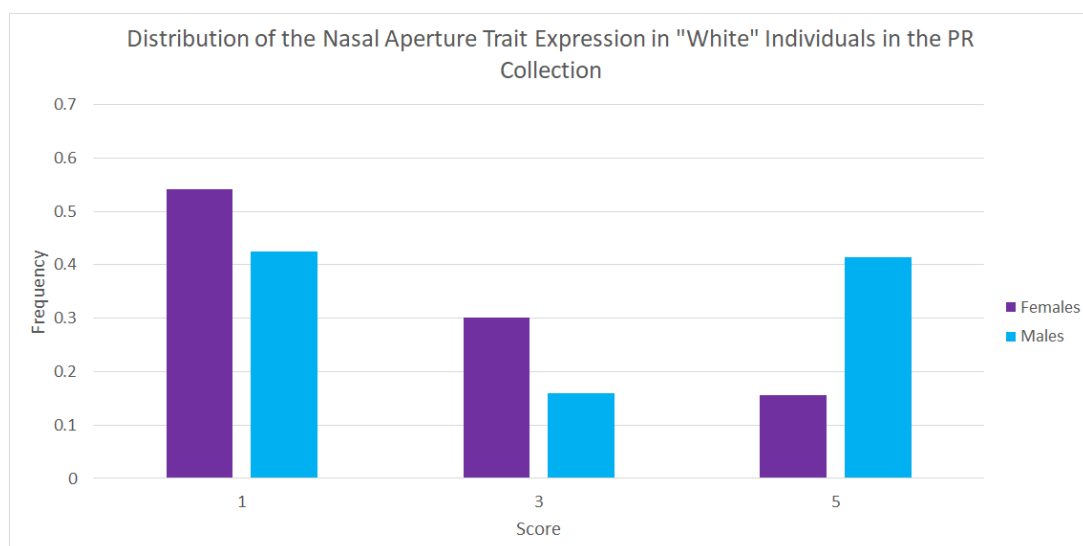
Figure F.63: The distribution of the nasal aperture trait expression in "White" individuals from the PR Collection represented using a bar chart. Females are in purple while males are in cyan.
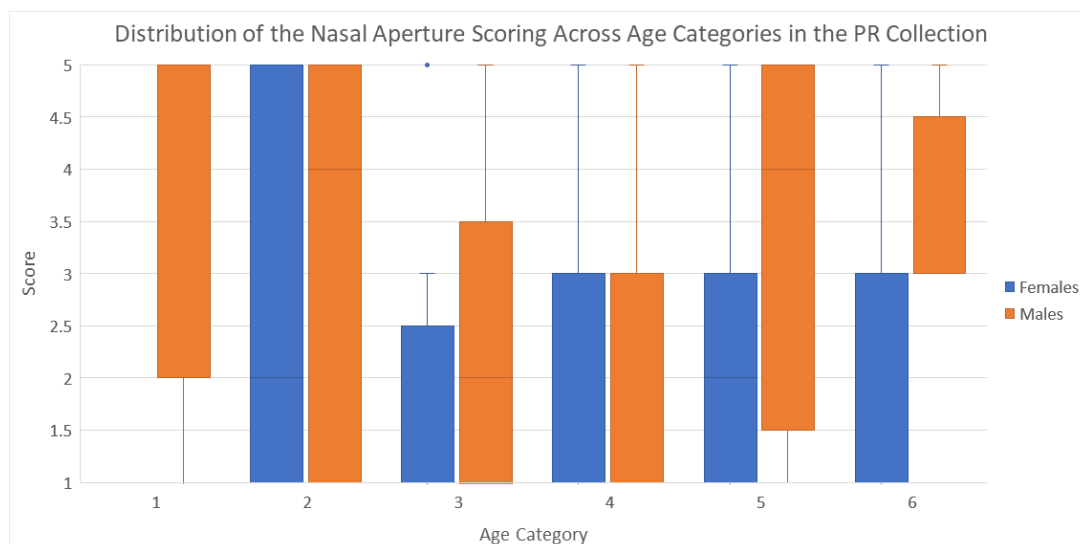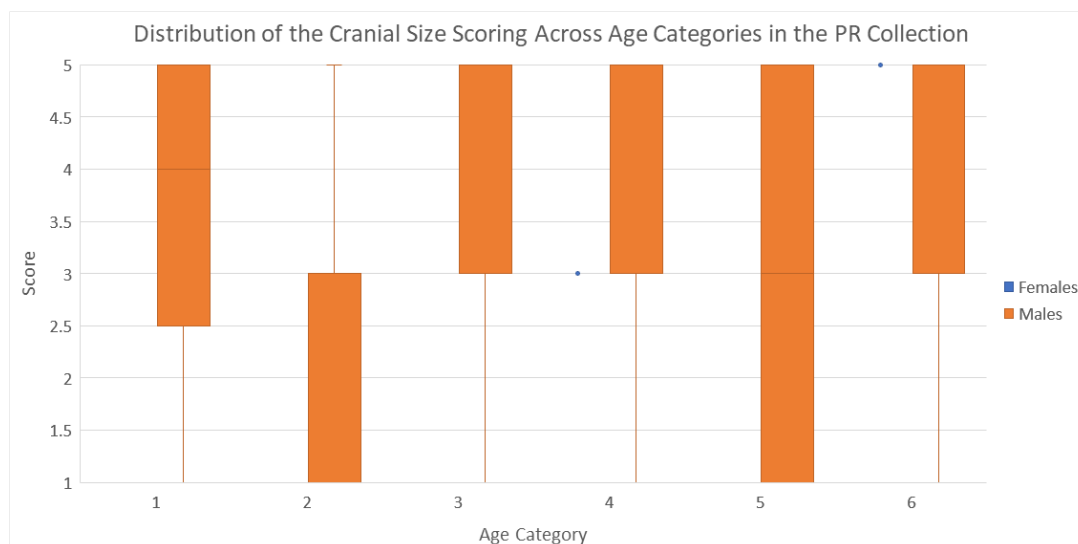
Figure F.64: A boxplot distribution of nasal aperture scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.
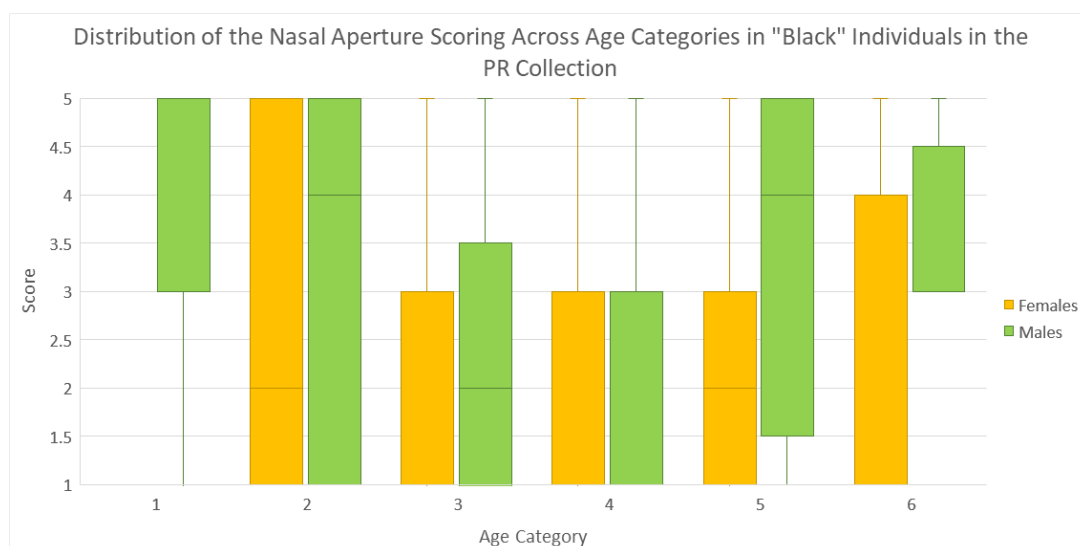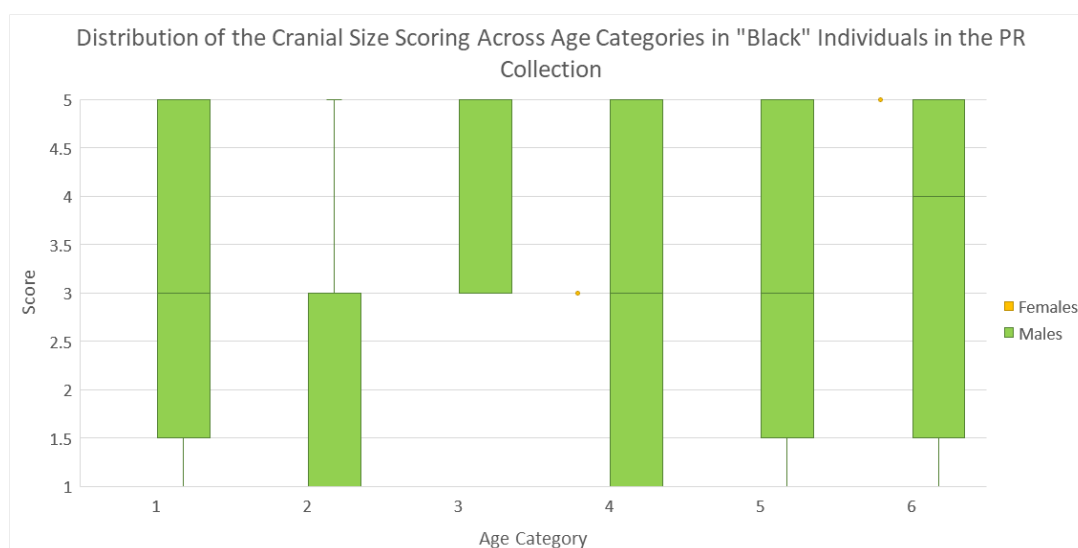

Figure F.65: A boxplot distribution of nasal aperture scoring across different age categories for "Black" males and females. Females are given in yellow while males are in green. The age categories are defined in Table 2.2.
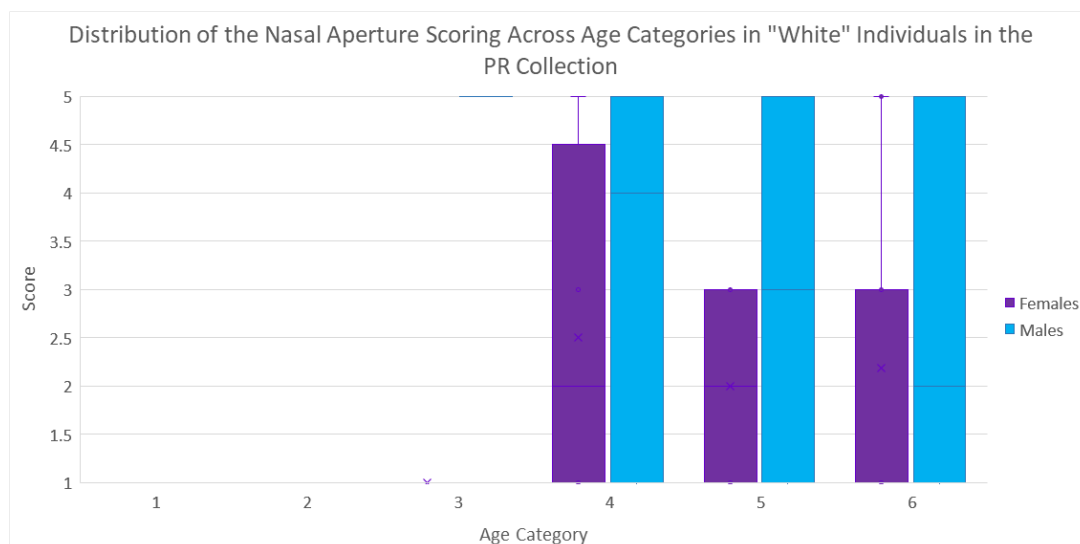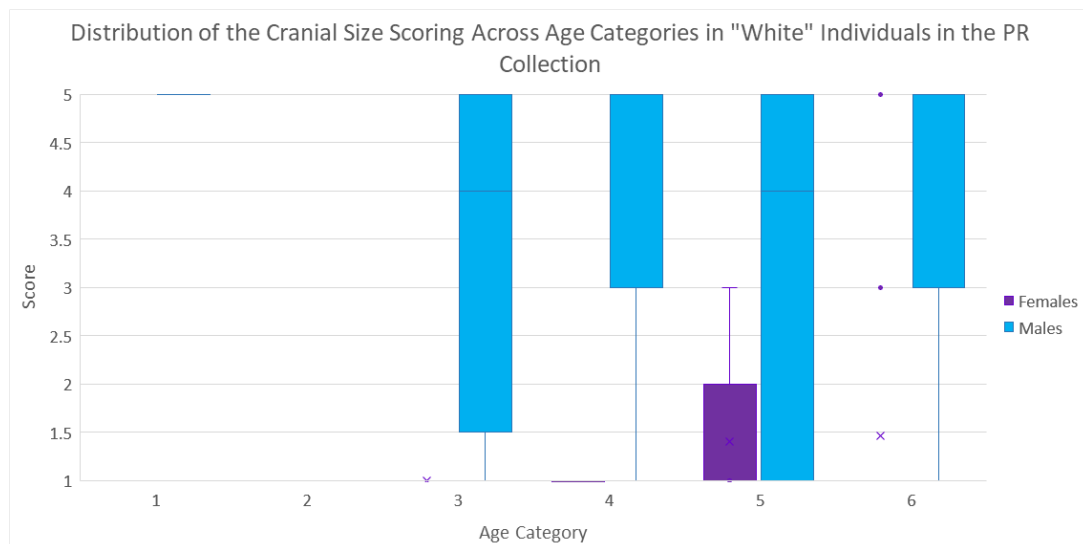
Figure F.66: A boxplot distribution of nasal aperture scoring across different age categories for "White" males and females. Females are given in purple while males are in cyan. The age categories are defined in Table 2.2.

Table F.16: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the PR collection when comparing nasal aperture trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 13	F = 1.0 M = 5.0	$U = 46.0$ $p = 0.017$ $z = 1.13$ $r = 0.27$	0.769
2	F = 10 M = 8	F = 2.0 M = 4.0	$U = 47.5$ $p = 0.500$ $z = -4.22$ $r = -0.99$	0.637
3	F = 18 M = 8	F = 1.0 M = 5.0	$U = 117.0$ $p = 0.006$ $z = -7.00$ $r = -1.37$	0.778
4	F = 19 M = 29	F = 1.0 M = 3.0	$U = 332.5$ $p = 0.189$ $z = -2.80$ $r = -0.40$	0.610
5	F = 20 M = 17	F = 3.0 M = 5.0	$U = 225.0$ $p = 0.077$ $z = -4.72$ $r = -0.78$	0.735
6	F = 71 M = 66	F = 1.0 M = 3.0	$U = 2712.0$ $p = 0.084$ $z = -9.42$ $r = -0.80$	0.637

Table F.17: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "Black" individuals from the PR collection when comparing nasal aperture trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 11	F = 1.0 M = 5.0	$U = 42.0$ $p = 0.006$ $z = 1.31$ $r = 0.34$	0.909
2	F = 10 M = 8	F = 2.0 M = 4.0	$U = 47.5$ $p = 0.500$ $z = -4.22$ $r = -0.99$	0.637
3	F = 16 M = 4	F = 1.0 M = 3.0	$U = 44.0$ $p = 0.216$ $z = -11.72$ $r = -2.62$	0.719
4	F = 15 M = 12	F = 1.0 M = 1.0	$U = 89.0$ $p = 0.977$ $z = -5.90$ $r = -1.14$	0.500
5	F = 10 M = 8	F = 3.0 M = 4.0	$U = 52.0$ $p = 0.279$ $z = -3.82$ $r = -0.90$	0.700
6	F = 4 M = 4	F = 1.0 M = 3.0	$U = 12.5$ $p = 0.222$ $z = -1.59$ $r = -0.56$	0.938

Table F.18: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "White" individuals from the PR collection when comparing nasal aperture trait scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 0 M = 2	F = N/A M = 1.0	U = N/A p = N/A z = N/A r = N/A	N/A
3	F = 2 M = 4	F = 1.0 M = 5.0	U = 8.0 p = 0.050 z = 0.46 r = 0.19	1.000
4	F = 4 M = 17	F = 2.0 M = 5.0	U = 44.0 p = 0.356 z = 0.00 r = 0.00	0.676
5	F = 10 M = 9	F = 2.0 M = 5.0	U = 60.0 p = 0.202 z = -3.27 r = -0.75	0.778
6	F = 67 M = 62	F = 1.0 M = 3.0	U = 2349.0 p = 0.163 z = -9.46 r = -0.83	0.633
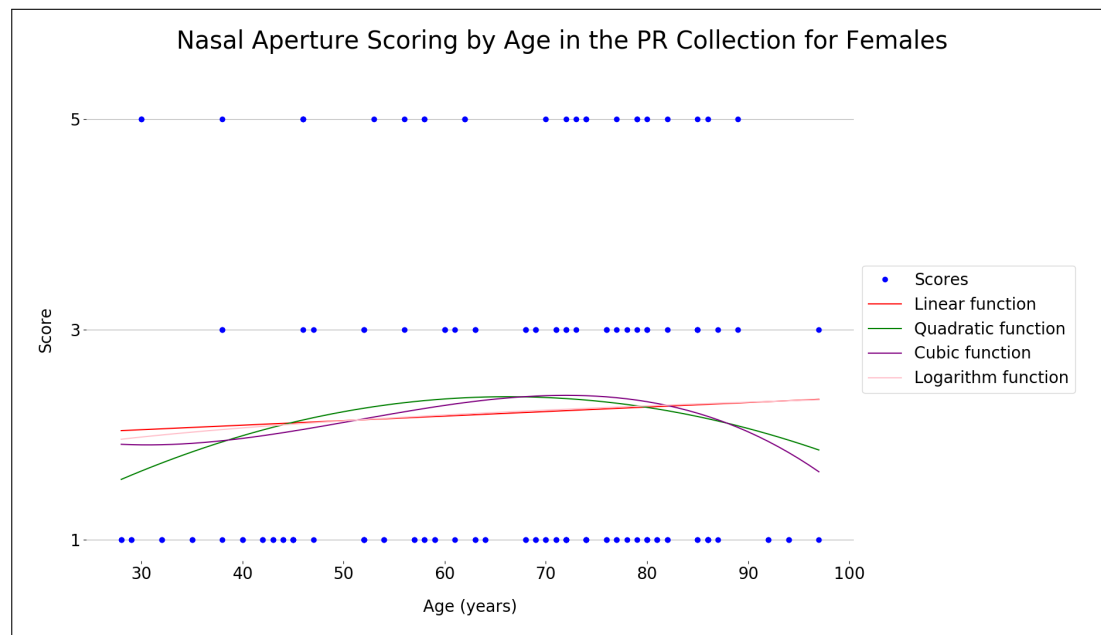
Figure F.67: A scatterplot of age vs. nasal aperture trait scoring for females in the PR collection, with four fitting functions.
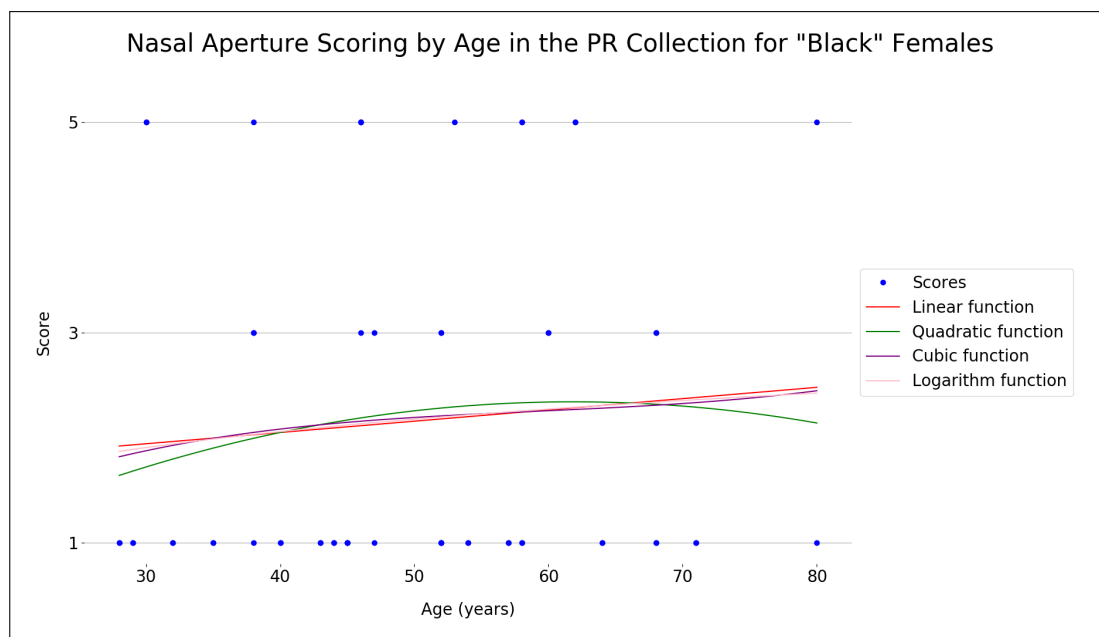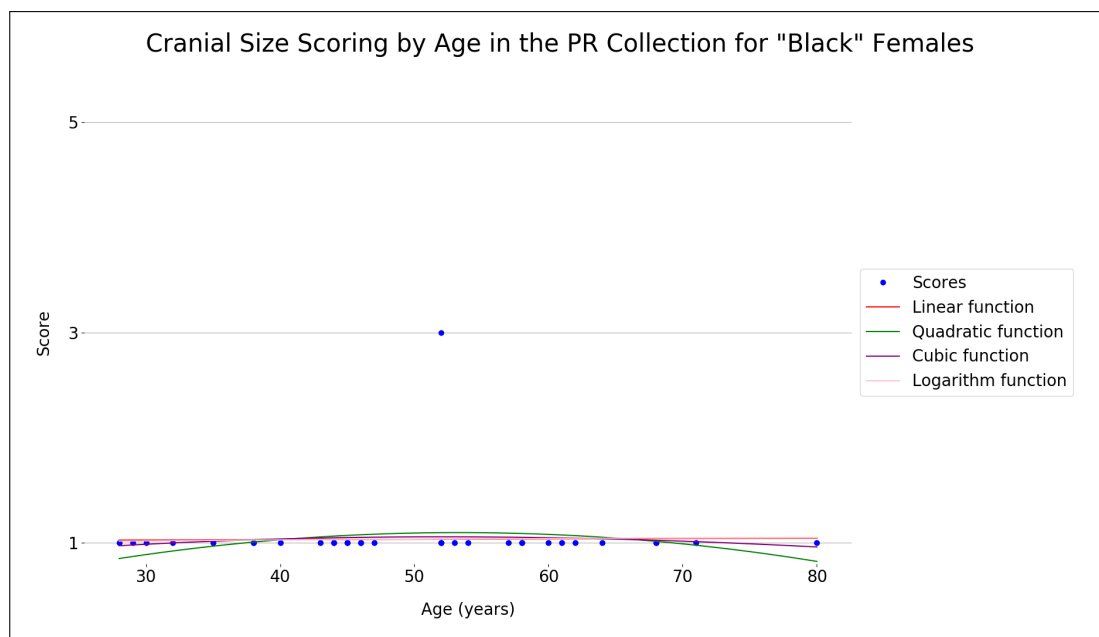
Figure F.68: A scatterplot of age vs. nasal aperture trait scoring for "Black" females in the PR collection, with four fitting functions.
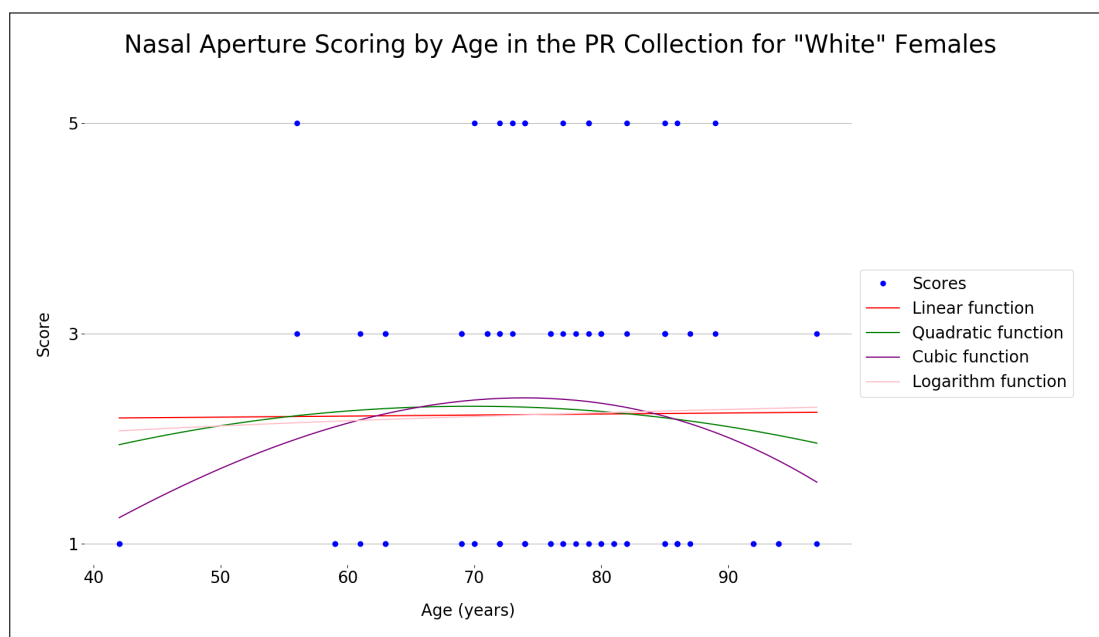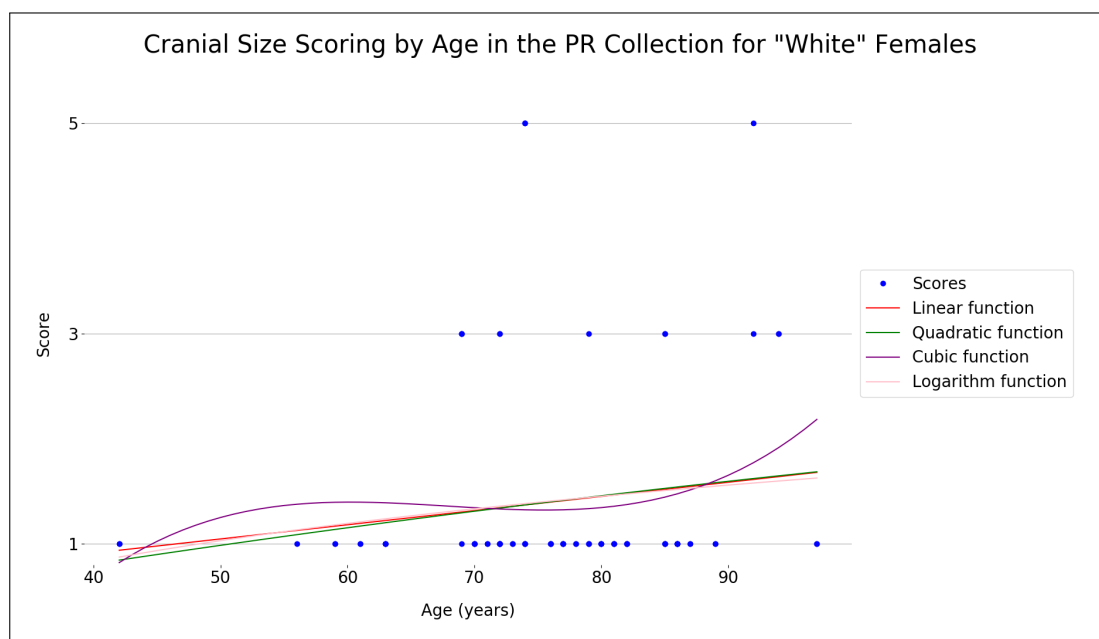

Figure F.69: A scatterplot of age vs. nasal aperture trait scoring for "White" females in the PR collection, with four fitting functions.
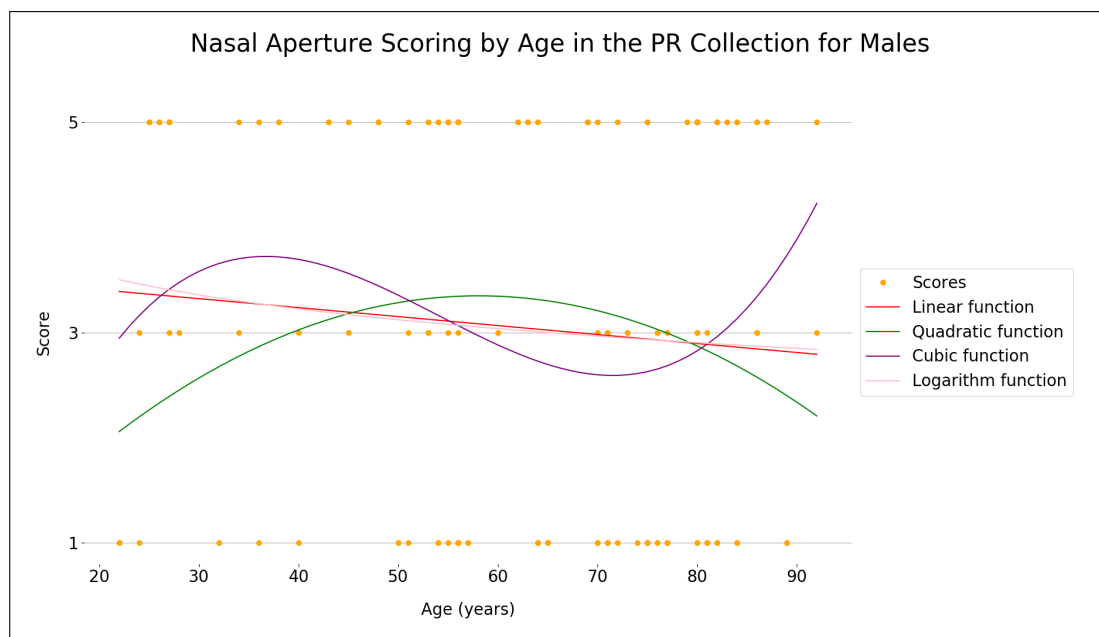
Figure F.70: A scatterplot of age vs. nasal aperture trait scoring for males in the PR collection, with four fitting functions.
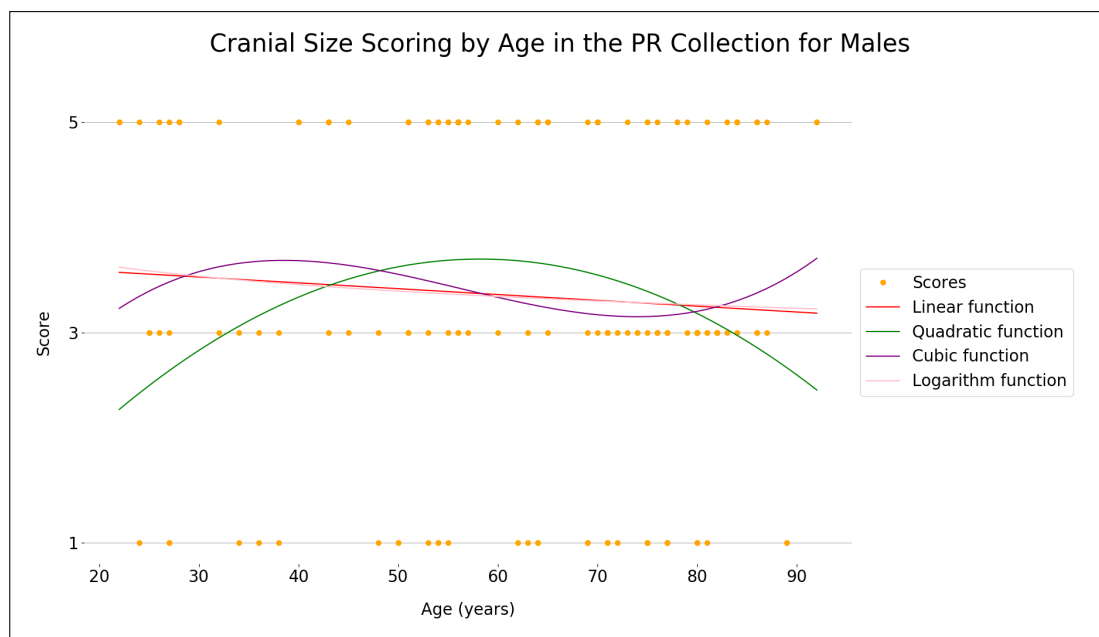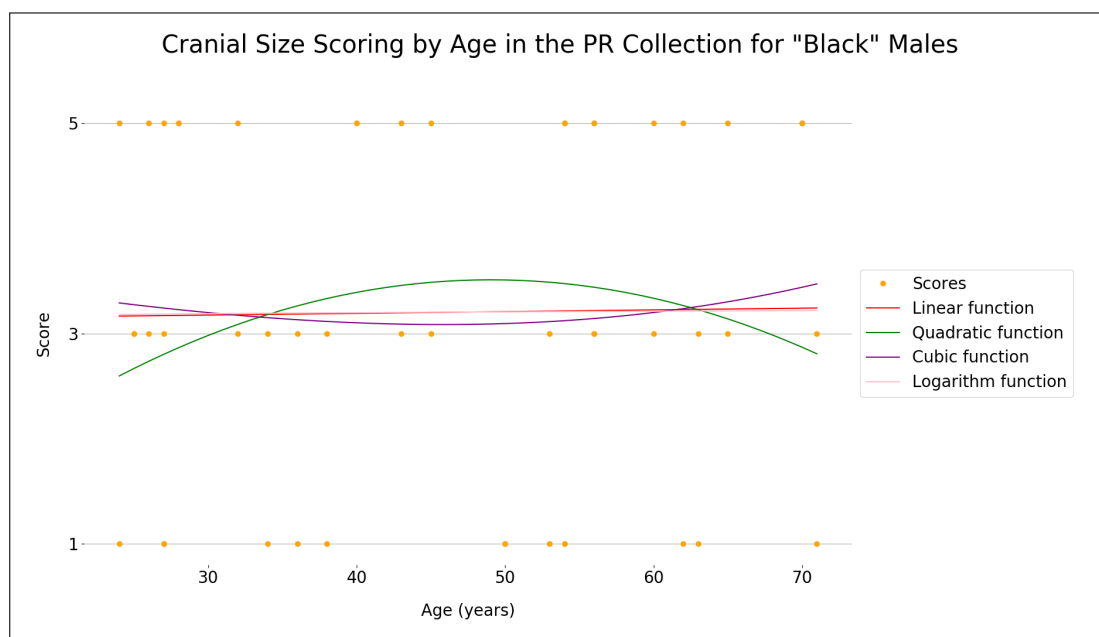

Figure F.71: A scatterplot of age vs. nasal aperture trait scoring for "Black" males in the PR collection, with four fitting functions.
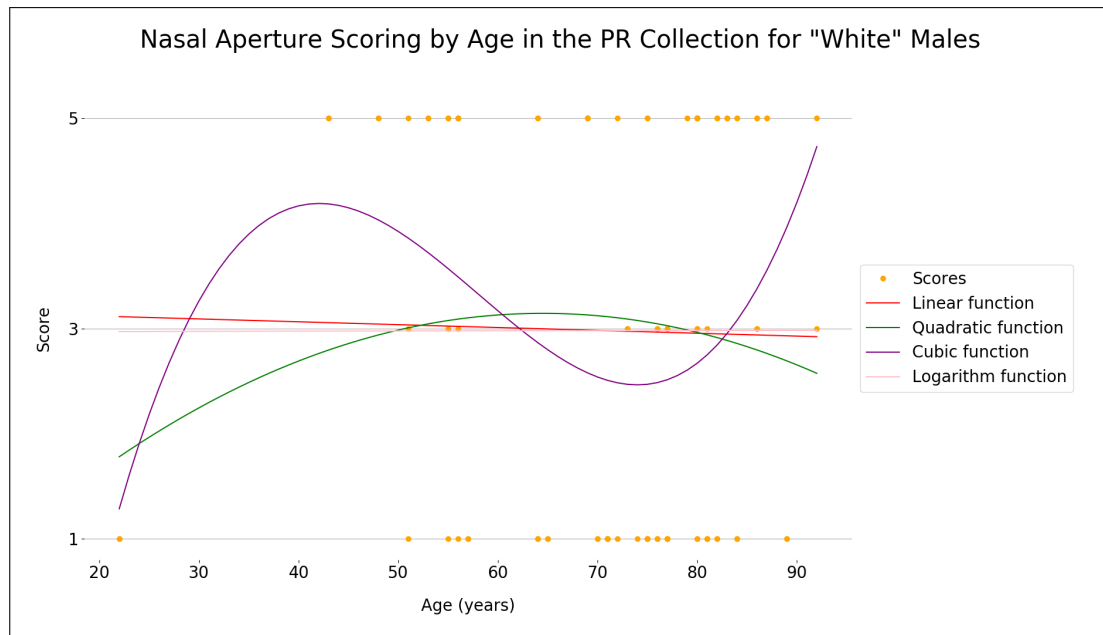
Figure F.72: A scatterplot of age vs. nasal aperture trait scoring for "White" males in the PR collection, with four fitting functions.

F.7   Cranial Size
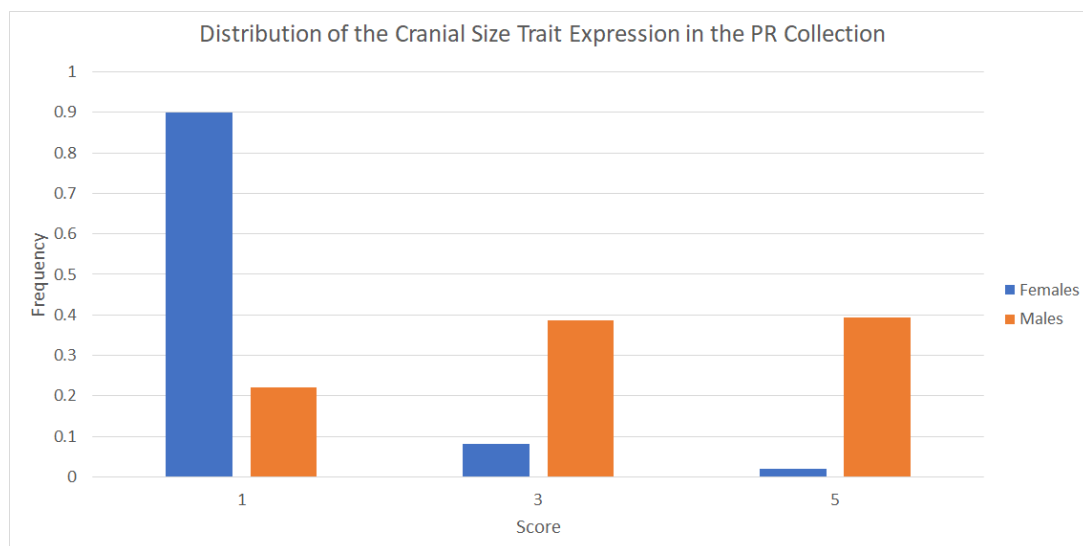
Figure F.73: The distribution of the cranial size in the PR Collection represented using a bar chart. Females are in blue while males are in orange.
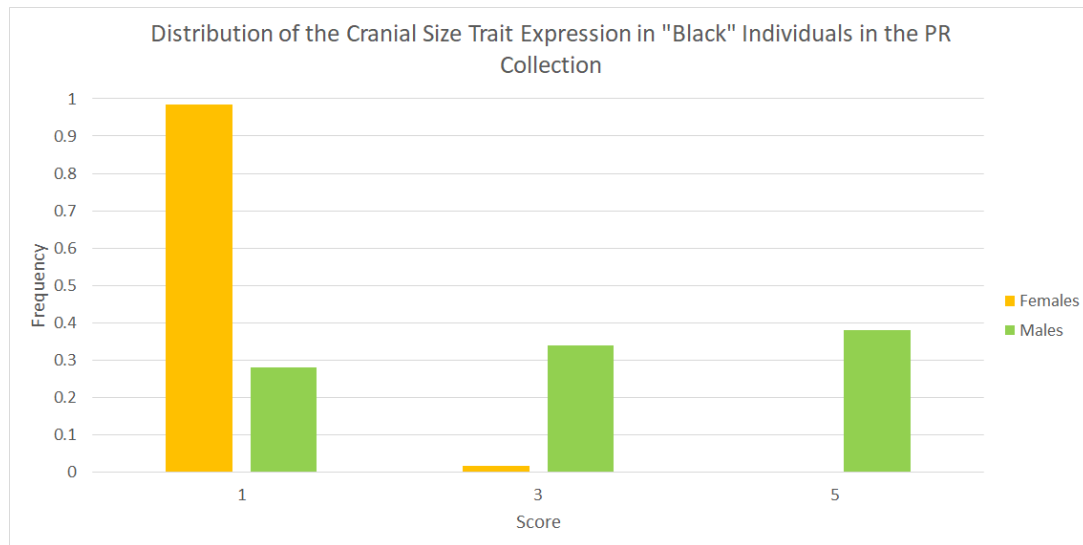
Figure F.74: The distribution of the cranial size in "Black" individuals from the PR Collection represented using a bar chart. Females are in yellow while males are in green.
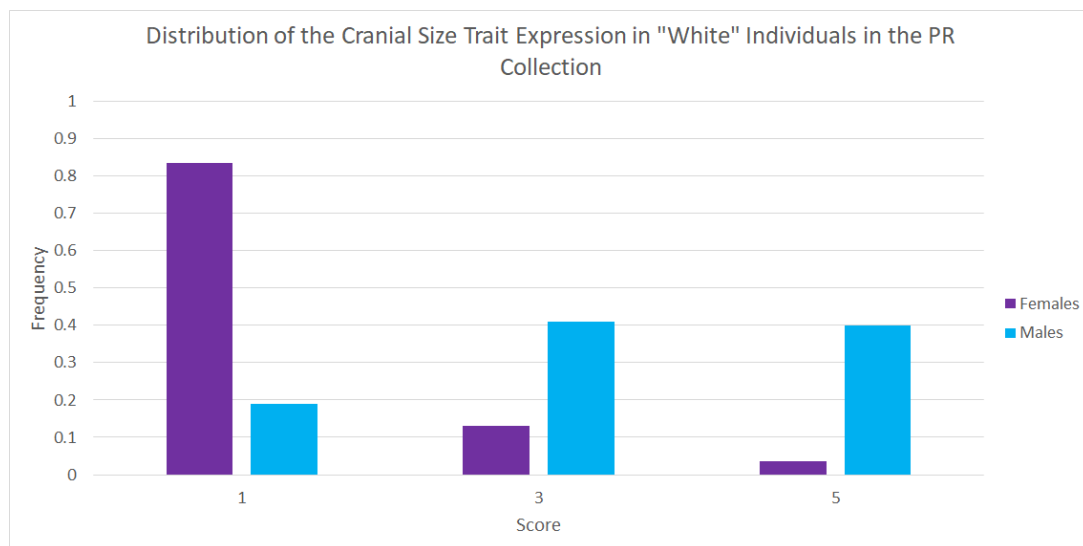
Figure F.75: The distribution of the cranial size in "White" individuals from the PR Collection represented using a bar chart. Females are in purple while males are in cyan.

Figure F.76: A boxplot distribution of cranial size scoring across different age categories for males and females. Females are given in blue while males are in orange. The age categories are defined in Table 2.2.

Figure F.77: A boxplot distribution of cranial size scoring across different age categories for "Black" males and females. Females are given in yellow while males are in green. The age categories are defined in Table 2.2.

Figure F.78: A boxplot distribution of cranial size scoring across different age categories for "White" males and females. Females are given in purple while males are in cyan. The age categories are defined in Table 2.2.

Table F.19: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in the PR collection when comparing cranial size scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 14	F = 1.0 M = 4.0	$U = 50.0$ $p = 0.015$ $z = 1.27$ $r = 0.30$	0.786
2	F = 10 M = 8	F = 1.0 M = 3.0	$U = 65.0$ $p = 0.005$ $z = -2.67$ $r = -0.63$	0.625
3	F = 20 M = 10	F = 1.0 M = 5.0	$U = 190.0$ $p << 0.001$ $z = -5.28$ $r = -0.96$	0.900
4	F = 19 M = 30	F = 1.0 M = 5.0	$U = 506.5$ $p << 0.001$ $z = 0.65$ $r = 0.09$	0.798
5	F = 22 M = 18	F = 1.0 M = 3.0	$U = 315.0$ $p < 0.001$ $z = -3.70$ $r = -0.58$	0.682
6	F = 74 M = 70	F = 1.0 M = 3.0	$U = 4303.0$ $p << 0.001$ $z = -4.24$ $r = -0.35$	0.763

Table F.20: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "Black" individuals from the PR collection when comparing cranial size scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 4 M = 12	F = 1.0 M = 3.0	$U = 42.0$ $p = 0.023$ $z = 0.97$ $r = 0.24$	0.750
2	F = 10 M = 8	F = 1.0 M = 3.0	$U = 65.0$ $p = 0.005$ $z = -2.67$ $r = -0.63$	0.625
3	F = 18 M = 6	F = 1.0 M = 5.0	$U = 108.0$ $p \ll 0.001$ $z = -7.80$ $r = -1.59$	1.000
4	F = 16 M = 12	F = 1.0 M = 3.0	$U = 148.0$ $p = 0.003$ $z = -3.90$ $r = -0.74$	0.594
5	F = 12 M = 8	F = 1.0 M = 3.0	$U = 84.0$ $p < 0.001$ $z = -3.24$ $r = -0.72$	0.750
6	F = 4 M = 4	F = 1.0 M = 4.0	$U = 14.0$ $p = 0.067$ $z = -1.15$ $r = -0.41$	0.750

Table F.21: The results of the Mann-Whitney statistical tests and the discrimination factor d for each age category in "White" individuals from the PR collection when comparing cranial size scoring. Results that indicate a statistically significant difference ($p < 0.05$) are given in blue.

Age Category	# of observations	Median Score	Mann-Whitney results	Discrimination Factor (d)
1	F = 0 M = 2	F = N/A M = 5.0	$U = N/A$ $p = N/A$ $z = N/A$ $r = N/A$	*N/A*
3	F = 2 M = 4	F = 1.0 M = 4.0	$U = 7.0$ $p = 0.211$ $z = 0.00$ $r = 0.00$	0.750
4	F = 3 M = 18	F = 1.0 M = 5.0	$U = 52.5$ $p = 0.004$ $z = 1.96$ $r = 0.43$	0.944
5	F = 10 M = 10	F = 1.0 M = 4.0	$U = 72.5$ $p = 0.065$ $z = -2.46$ $r = -0.55$	0.690
6	F = 70 M = 66	F = 1.0 M = 3.0	$U = 3821.5$ $p << 0.001$ $z = -4.24$ $r = -0.36$	0.762
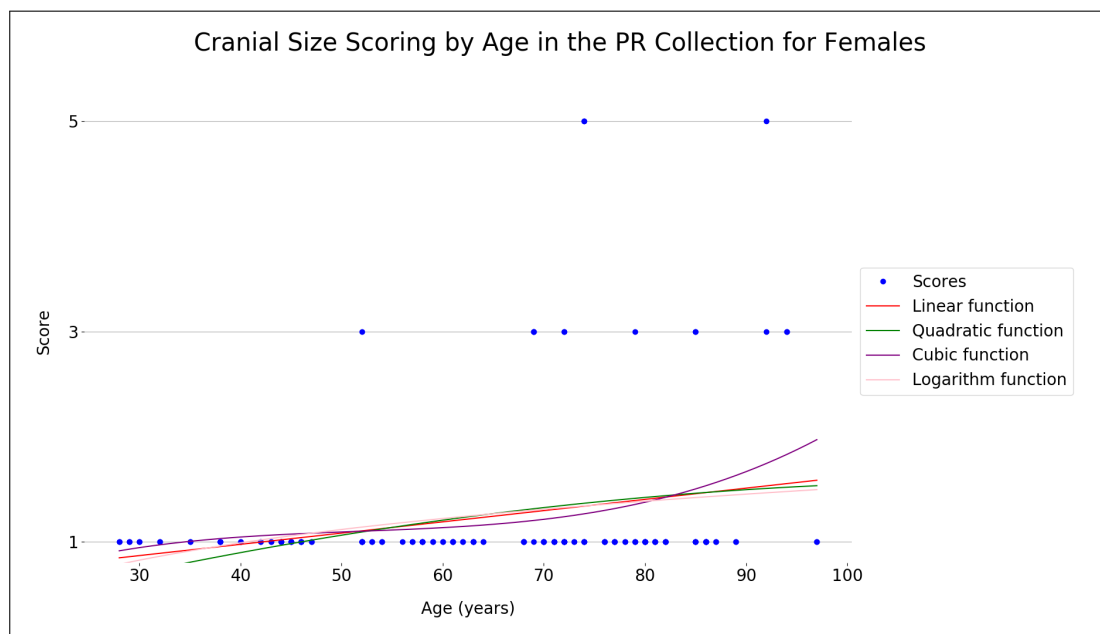
Figure F.79: A scatterplot of age vs. cranial size scoring for females in the PR collection, with four fitting functions.

Figure F.80: A scatterplot of age vs. cranial size scoring for "Black" females in the PR collection, with four fitting functions.

Figure F.81: A scatterplot of age vs. cranial size scoring for "White" females in the PR collection, with four fitting functions.

Figure F.82: A scatterplot of age vs. cranial size scoring for males in the PR collection, with four fitting functions.

Figure F.83: A scatterplot of age vs. cranial size scoring for "Black" males in the PR collection, with four fitting functions.
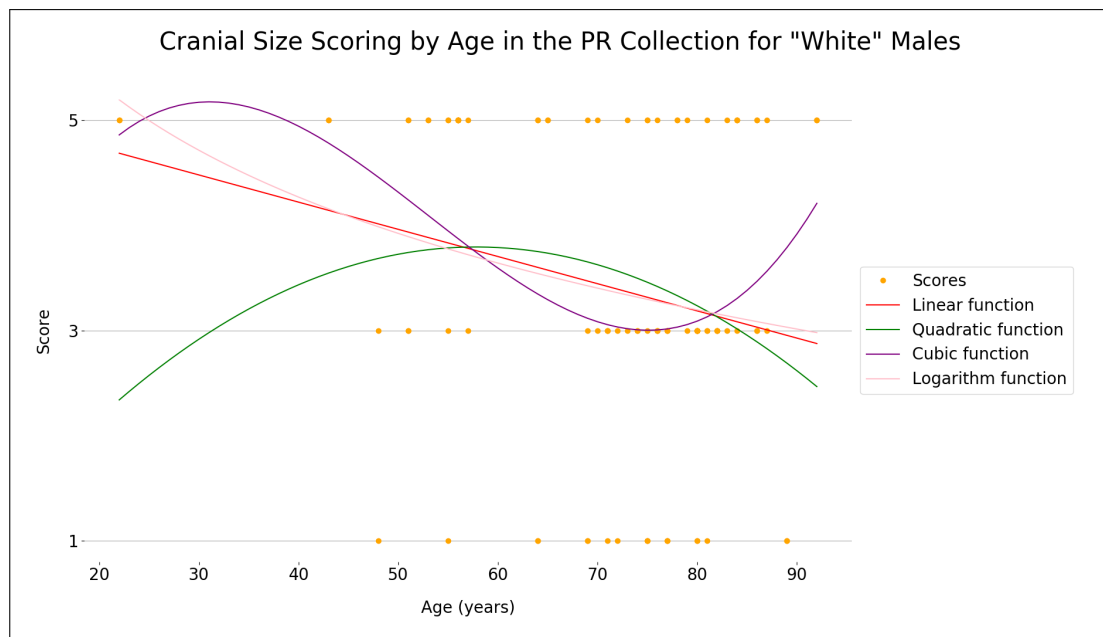
Figure F.84: A scatterplot of age vs. cranial size scoring for "White" males in the PR collection, with four fitting functions.

APPENDIX G: C2C Distances for Crania Pairs

# Generated Using DAVID 4
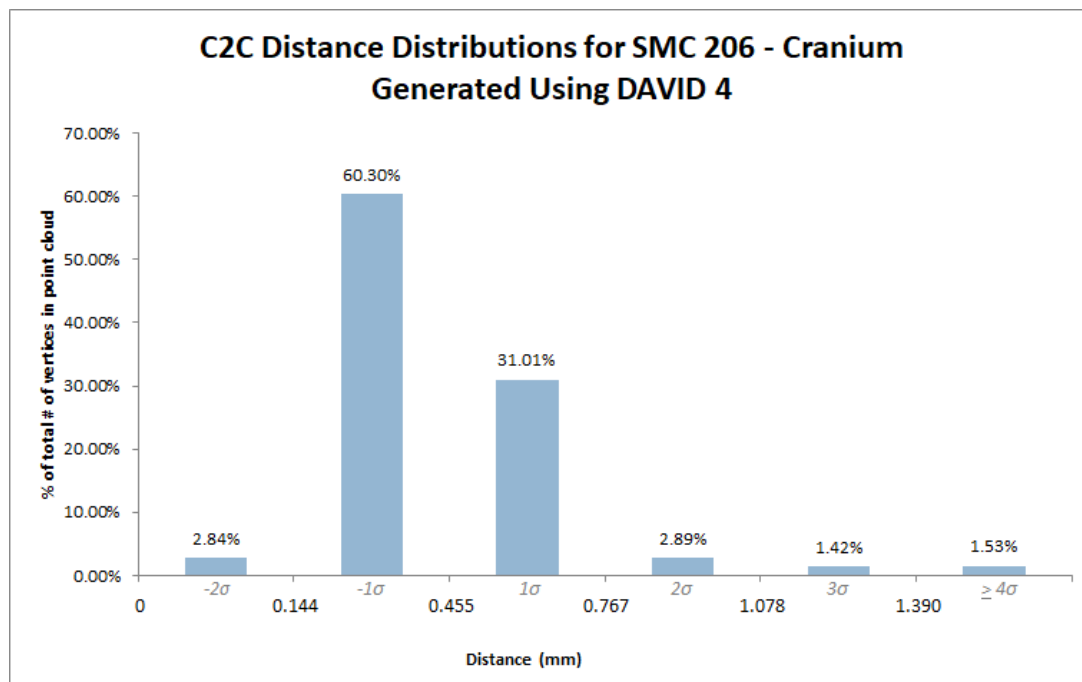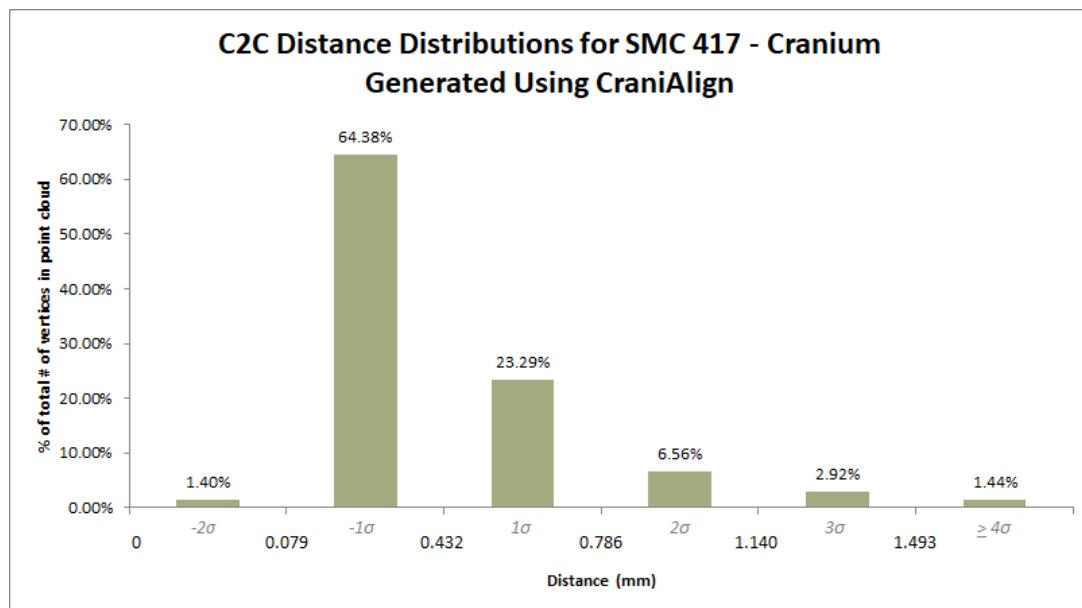
Figure G.1
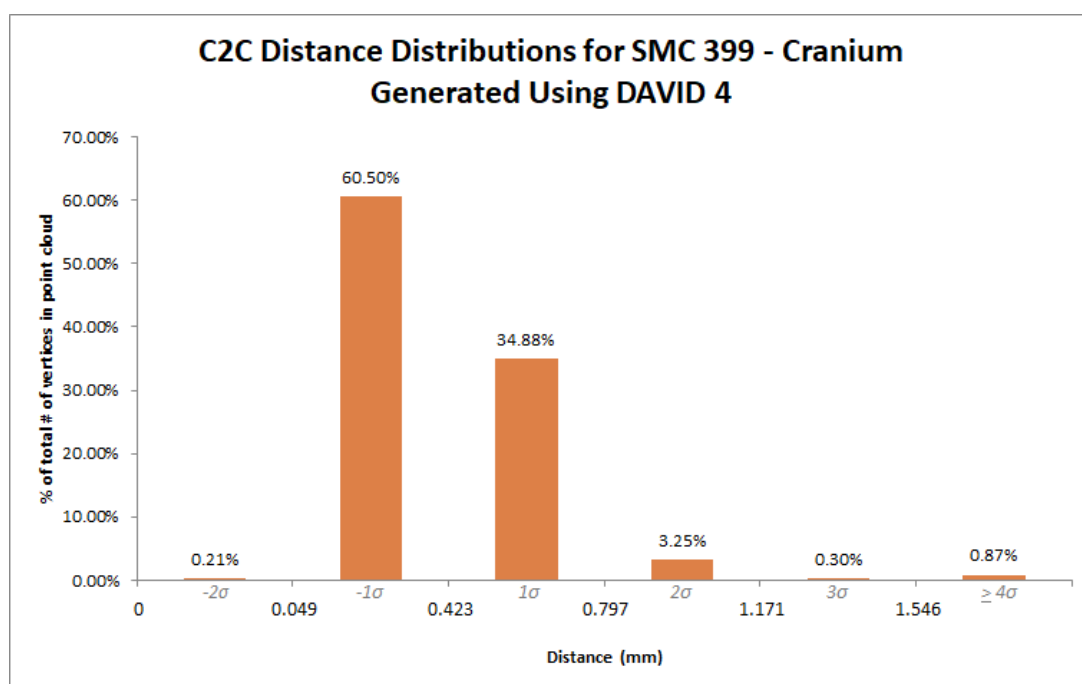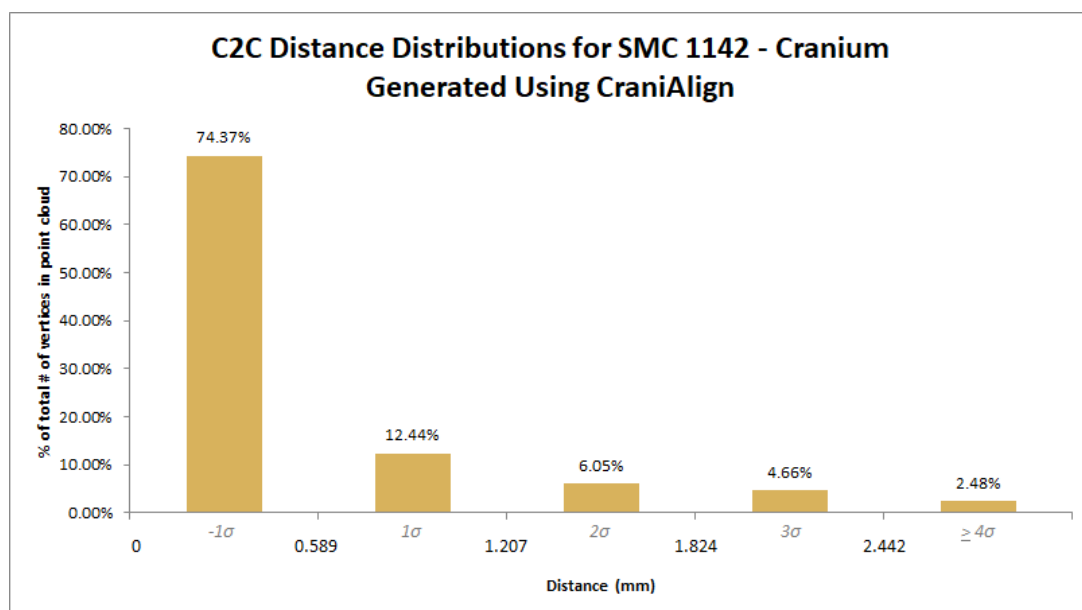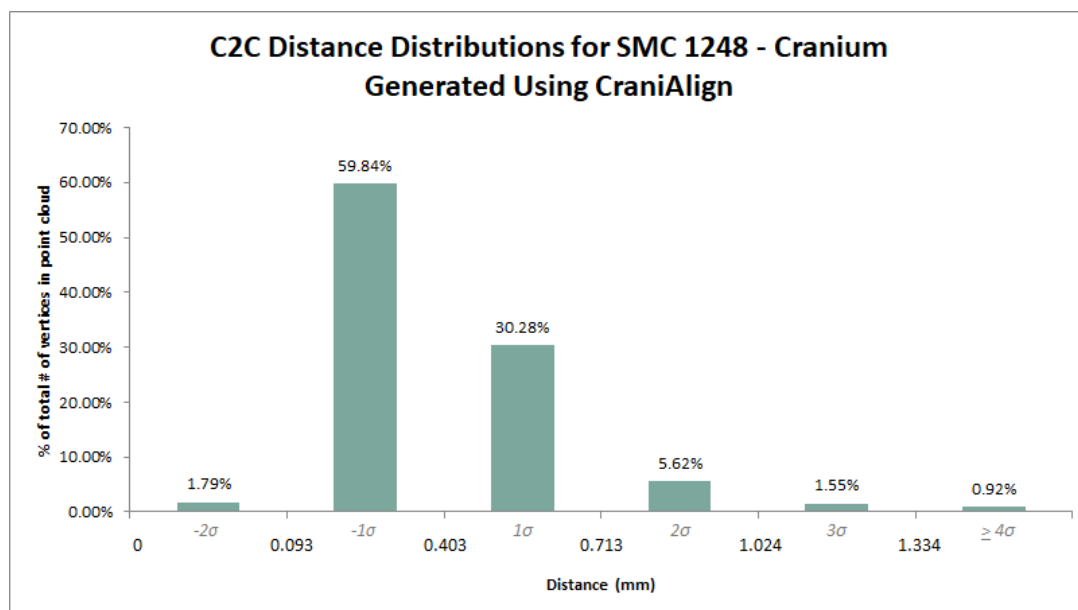

Figure G.2
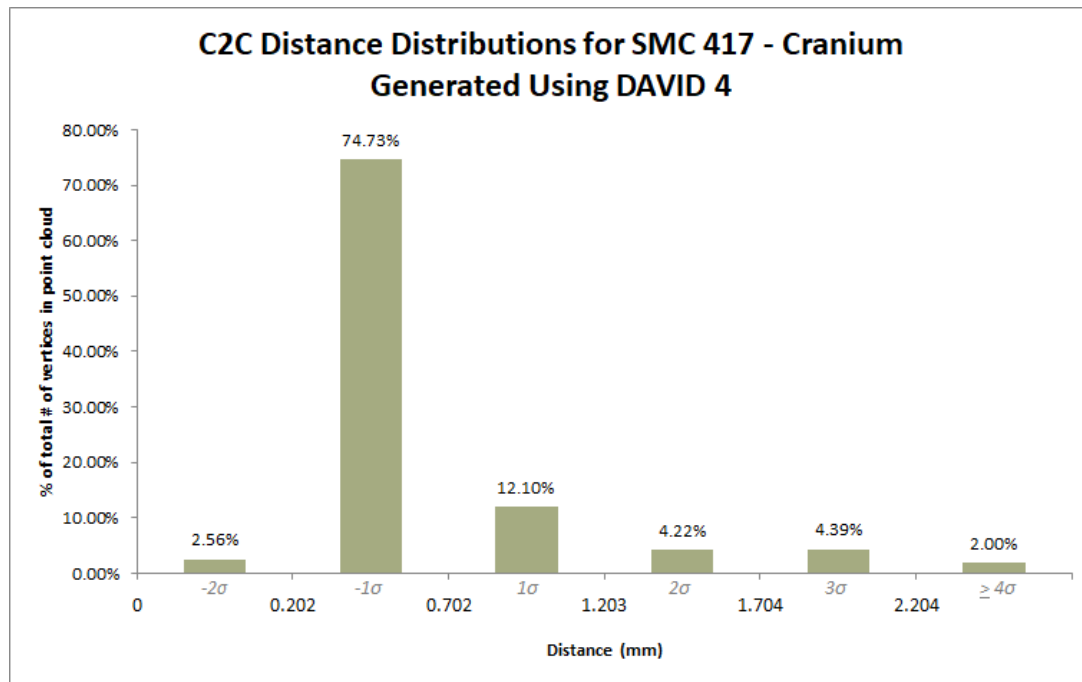
Figure G.3
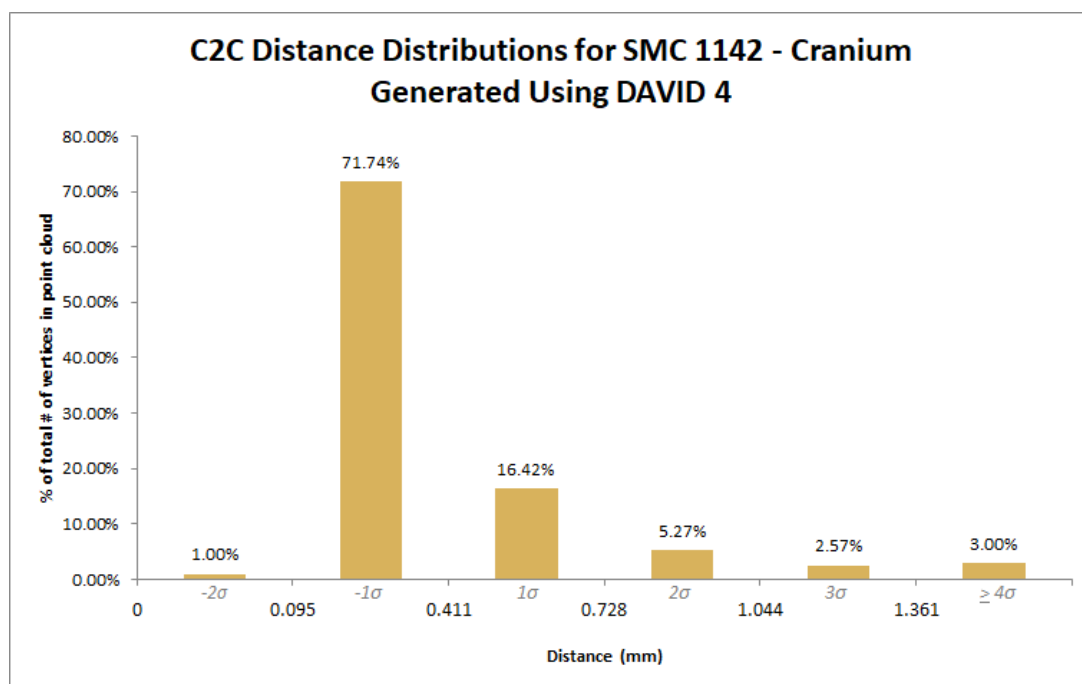
Figure G.4
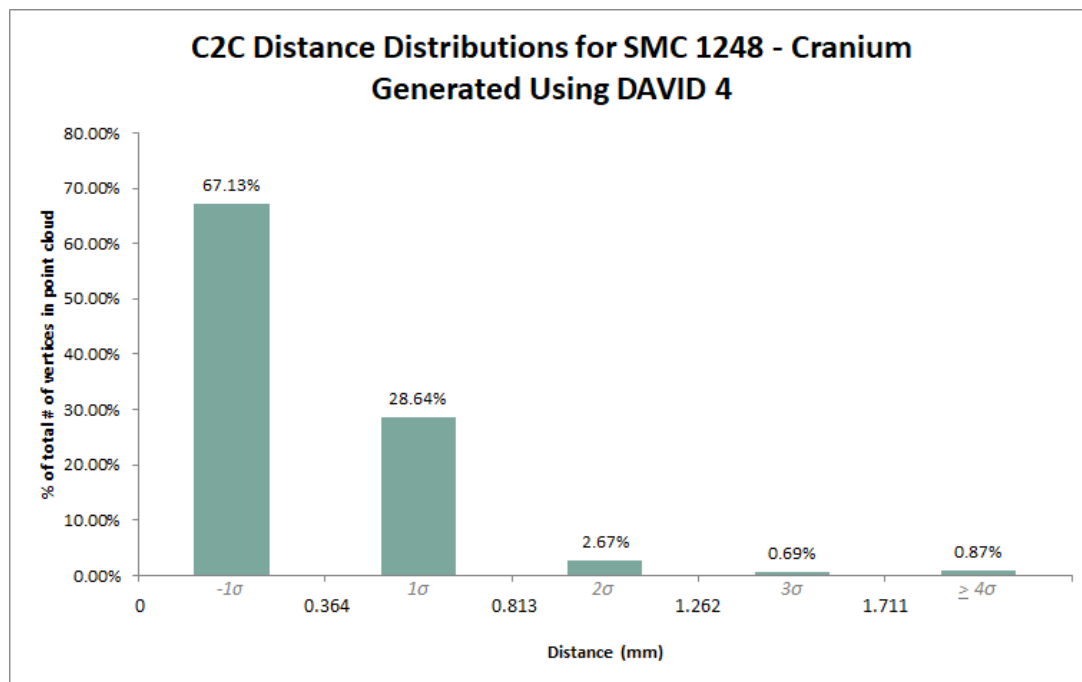
Figure G.5

APPENDIX H: Probability Distribution Function Modelling for DAVID 4 C2C Distributions

List of continuous random variables used:

- Alpha
- Beta prime
- Burr (Type III)
- Burr (Type XII)
- Chi
- Chi-squared
- Erlang
- Exponential
- Exponential power
- Exponentiated Weibull
- F-distribution
- Fatigue Life (Birnbaum-Saunders)
- Fisk (log-logistic)
- Folded Cauchy
- Folded Norm
- Fréchet right (variation of Weibull minimum)

- Gamma
- Generalized exponential
- Generalized gamma
- Generalized half-logistic
- Generalized Pareto
- Gilbrat
- Gompertz
- Half-Cauchy
- Half generalized normal
- Half-logistic
- Half-normal
- Inverse-gamma
- Inverse Gaussian
- Inverse Weibull
- Three-parameter kappa
- General Kolmogorov-Smirnov one-sided test
- Kolmogorov-Smirnov two-sided test

- Lévy
- Log-Laplace
- Log-normal
- Lomax
- Maxwell
- Mielke's Beta-Kappa
- Nakagami
- Non-central F-distribution
- Non-central chi-squared
- Power log-normal
- Rayleigh
- Reciprocal inverse Gaussian
- Rice (Ricean)
- Truncated normal
- Wald (variation of inverse Gaussian)
- Weibull minimum

To calculate the bin sizes for the histograms required for modelling the data, the Freedman-Diaconis rule (1981) was followed:

$$Binsize = 2(IQR(x))/(\sqrt[3]{n})$$

where IQR = interquartile ratio; x = the distribution of a given sample; and n = the number of

observations, i.e. the number of data points in the given distribution.

Table H.1: The parameters of the probability distribution functions (PDF's) that best approximated each sample's C2C distance distribution for DAVID 4 samples.

Sample #	PDF	PDF Parameters
SMC206	Log-Laplace	(3.293, -0.028, 0.438)
SMC399	Mielke's Beta-Kappa	(2.739, 4.253, 0.002, 0.435)
SMC417	Folded Cauchy	(5.454, 0.003, 0.101)
SMC1142	Folded Cauchy	(3.621, 0.005, 0.080)
SMC1248	Fisk (log-logistic)	(2.904, 0.002, 0.284)
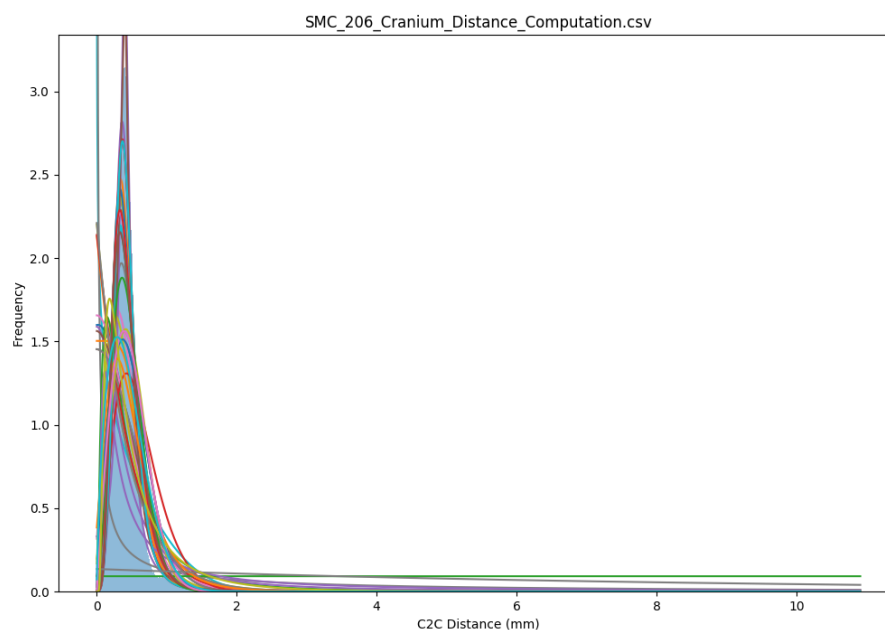
Figure H.1: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC206, with all the fitted PDF's.
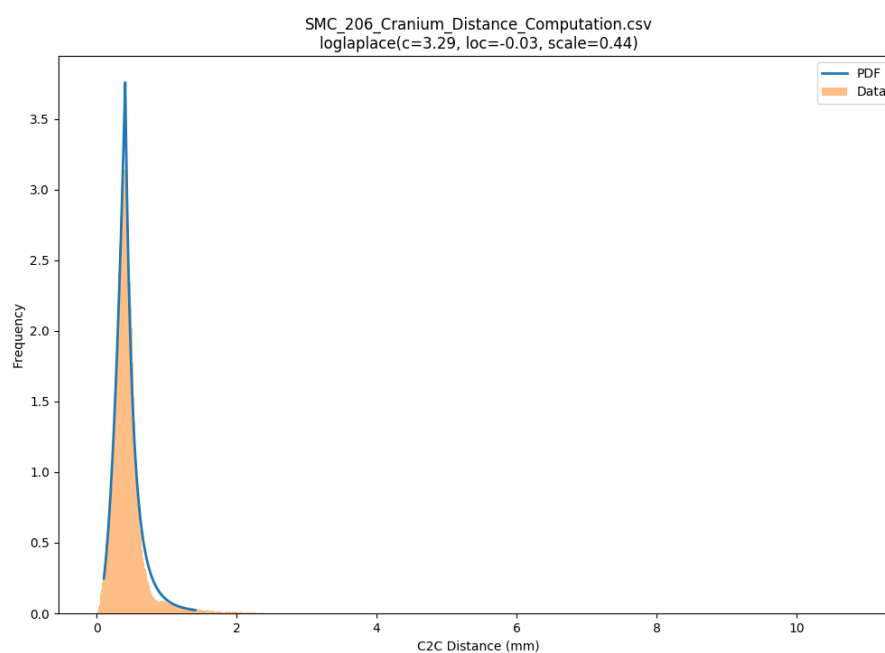
Figure H.2: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC206, with the best fitted PDF.
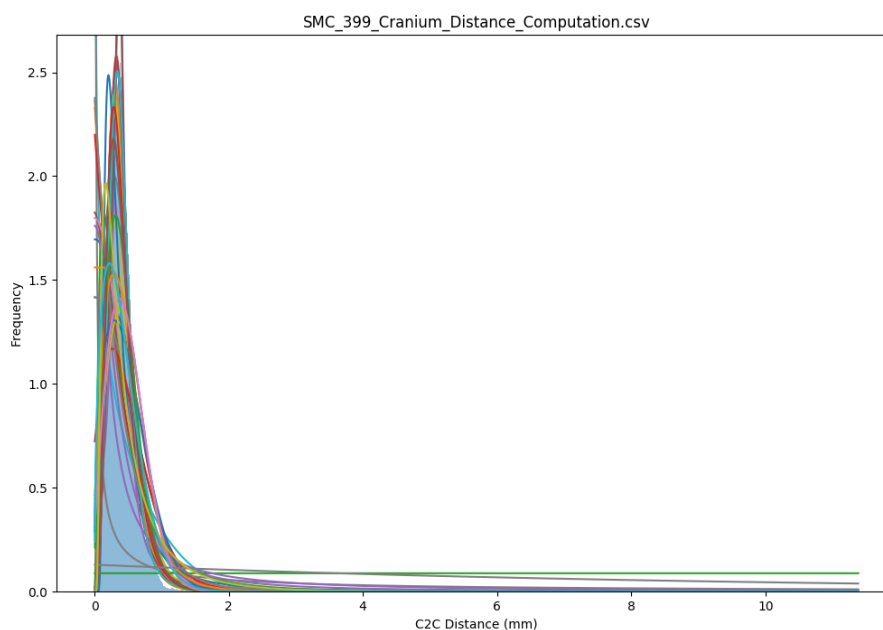
Figure H.3: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC399, with all the fitted PDF's.
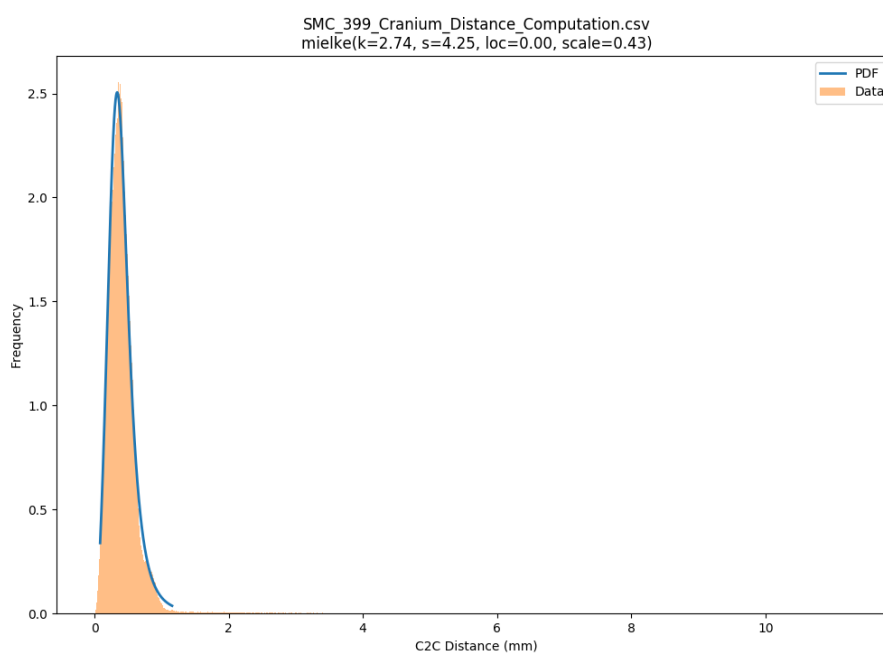
Figure H.4: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC206, with the best fitted PDF.
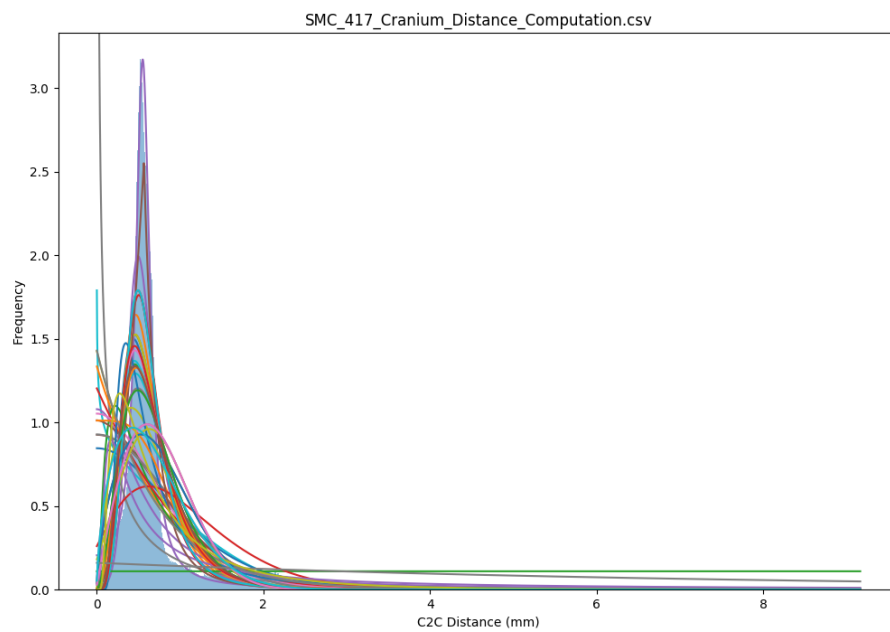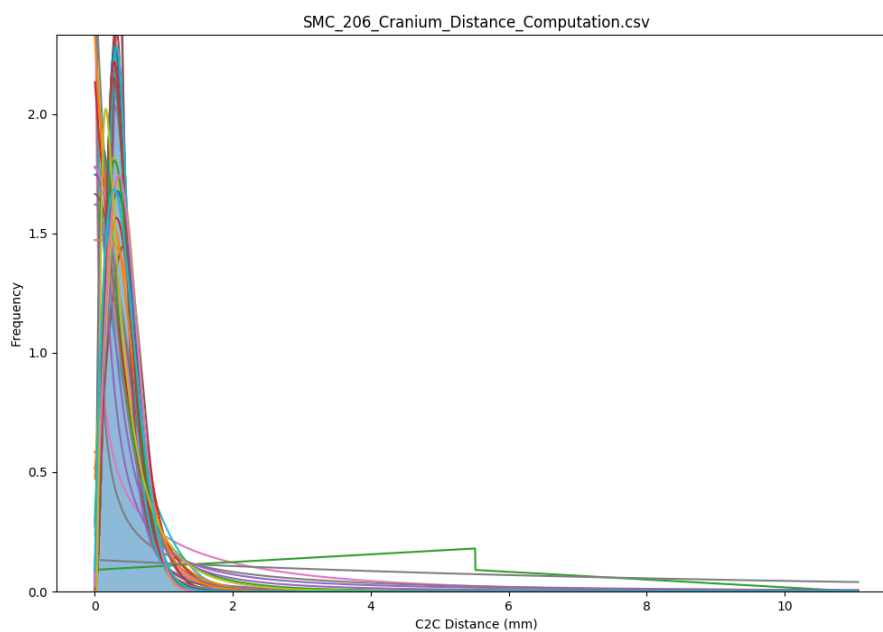
Figure H.5: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC417, with all the fitted PDF's.
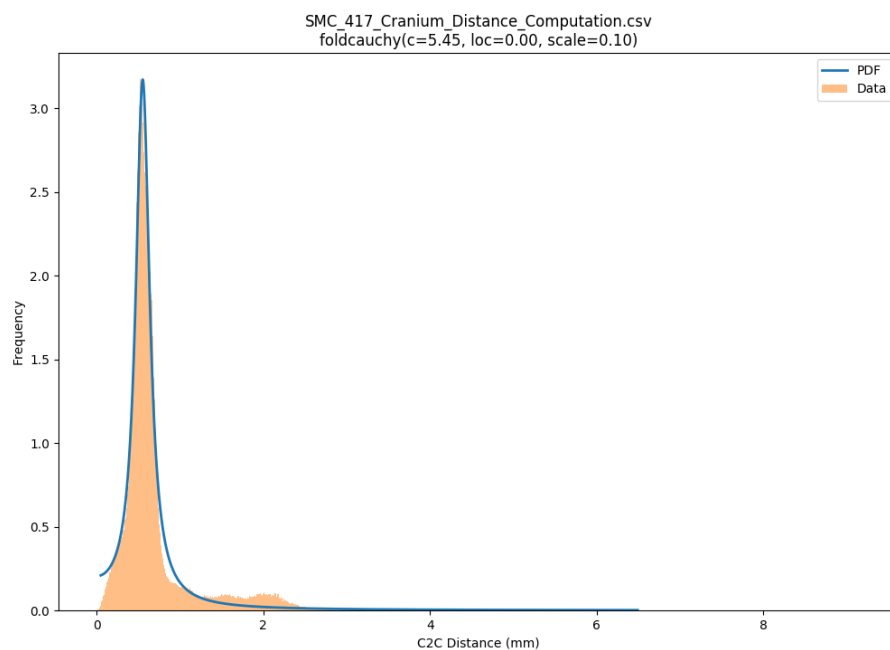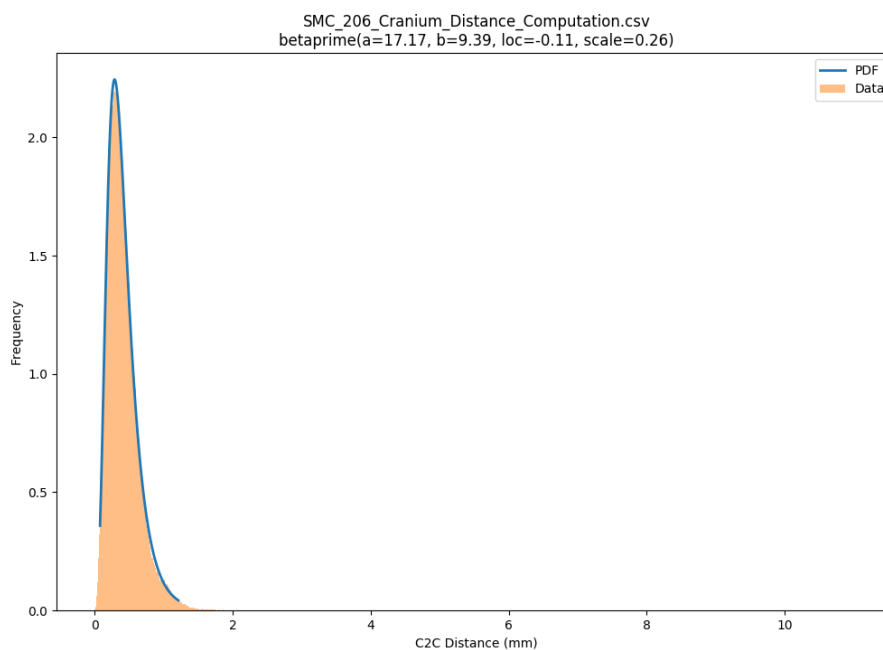

Figure H.6: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC417, with the best fitted PDF.
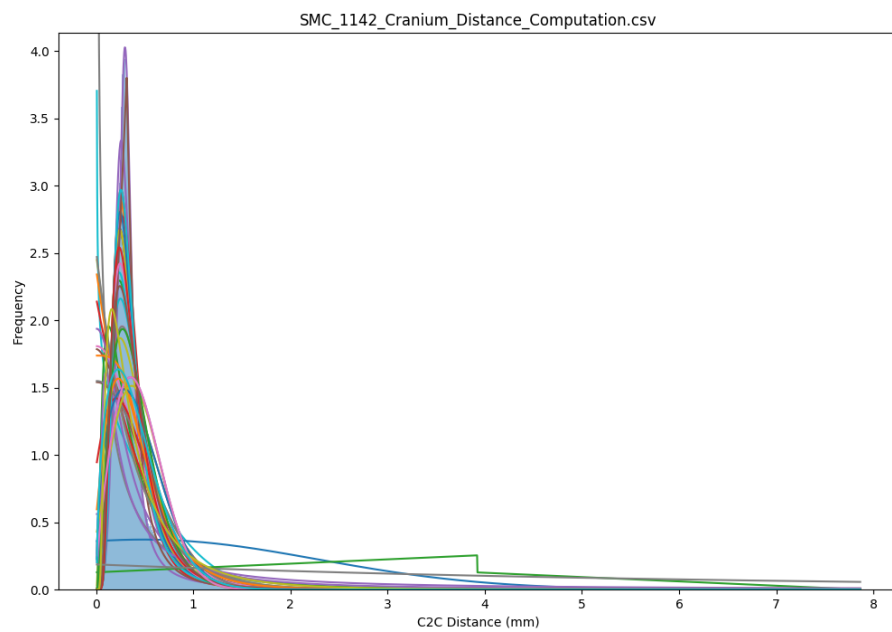
Figure H.7: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC1142, with all the fitted PDF's.
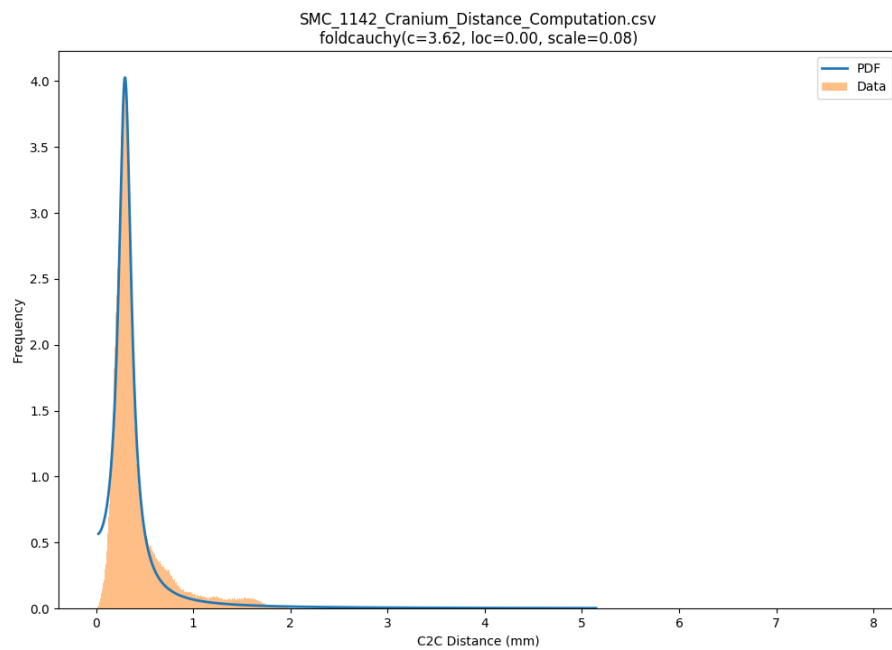
Figure H.8: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC1142, with the best fitted PDF.
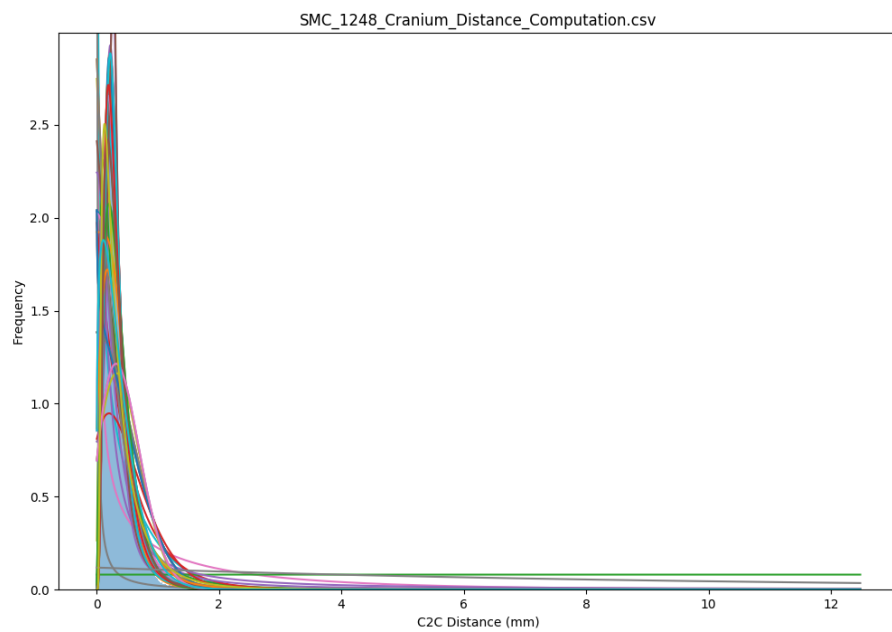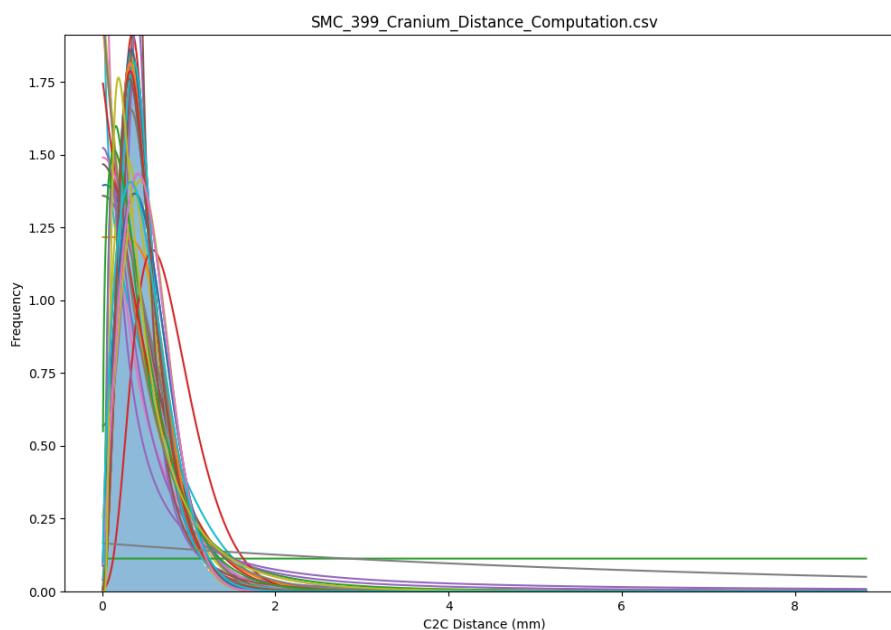
Figure H.9: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC1248, with all the fitted PDF's.
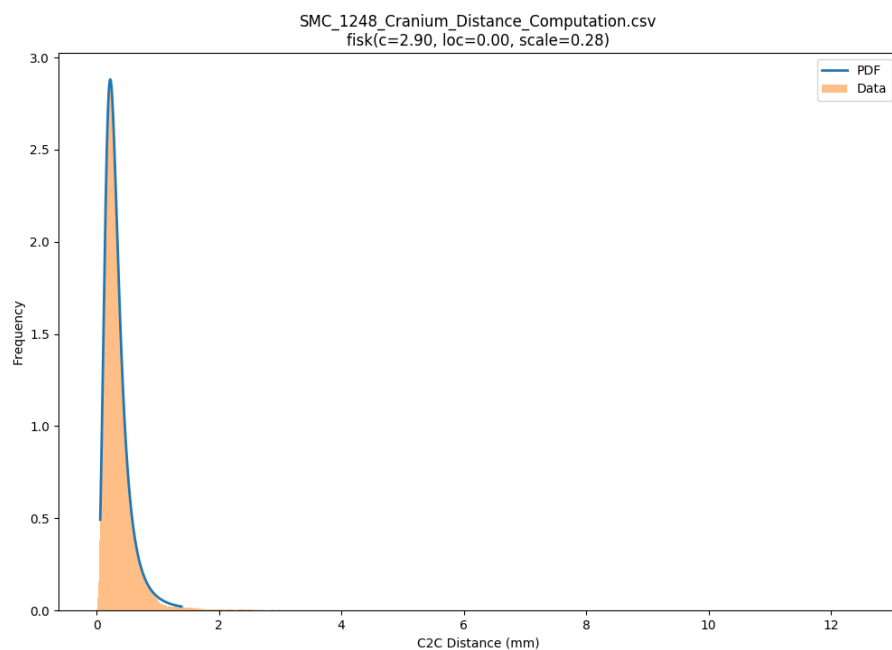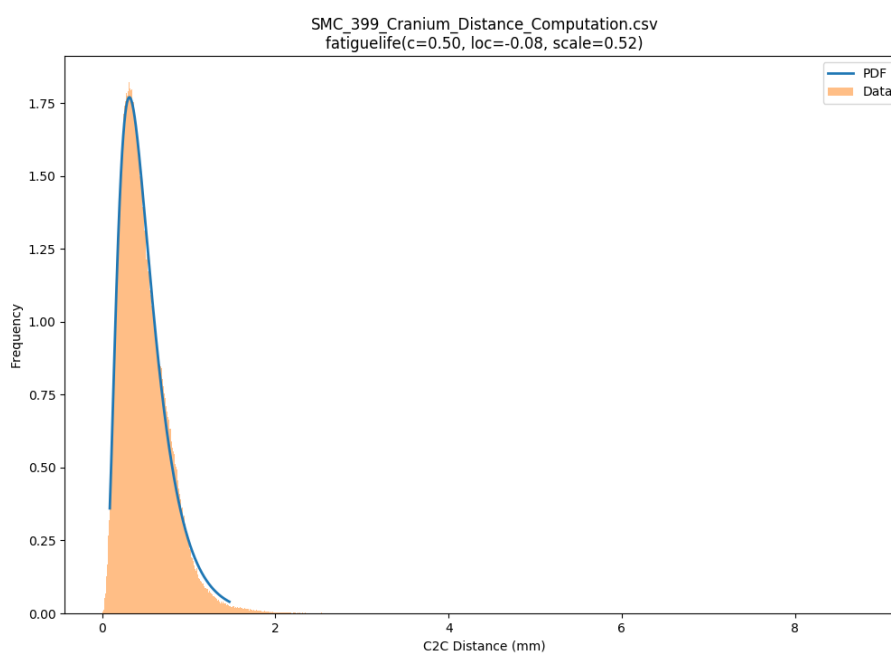

Figure H.10: A histogram of the C2C distributions from comparing DAVID 4 alignments for SMC1248, with the best fitted PDF.

APPENDIX I: C2C Distances for Crania Pairs Generated Using CraniAlign
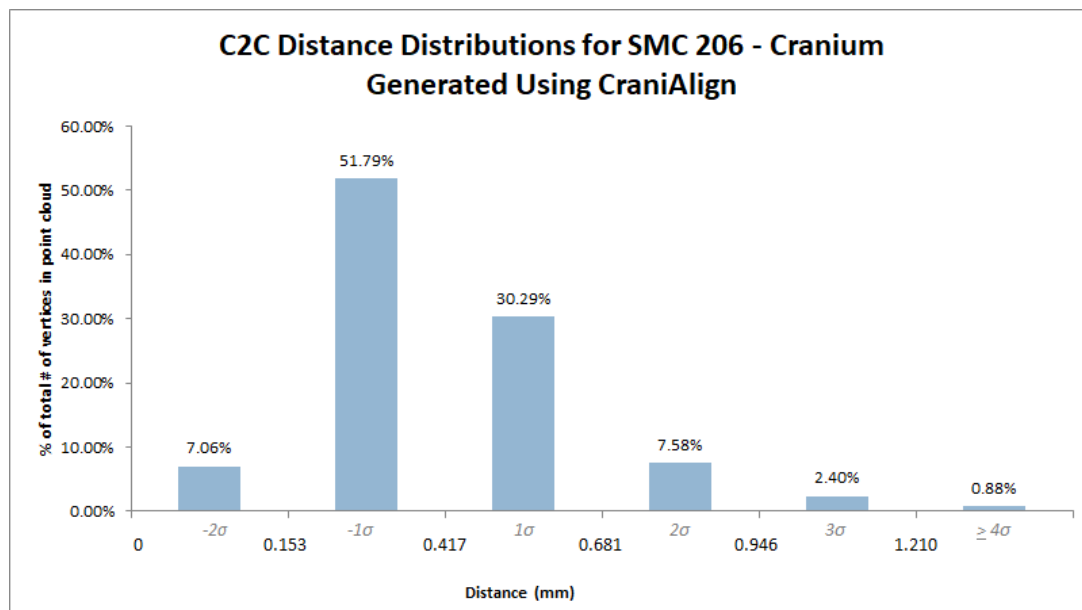
Figure I.1
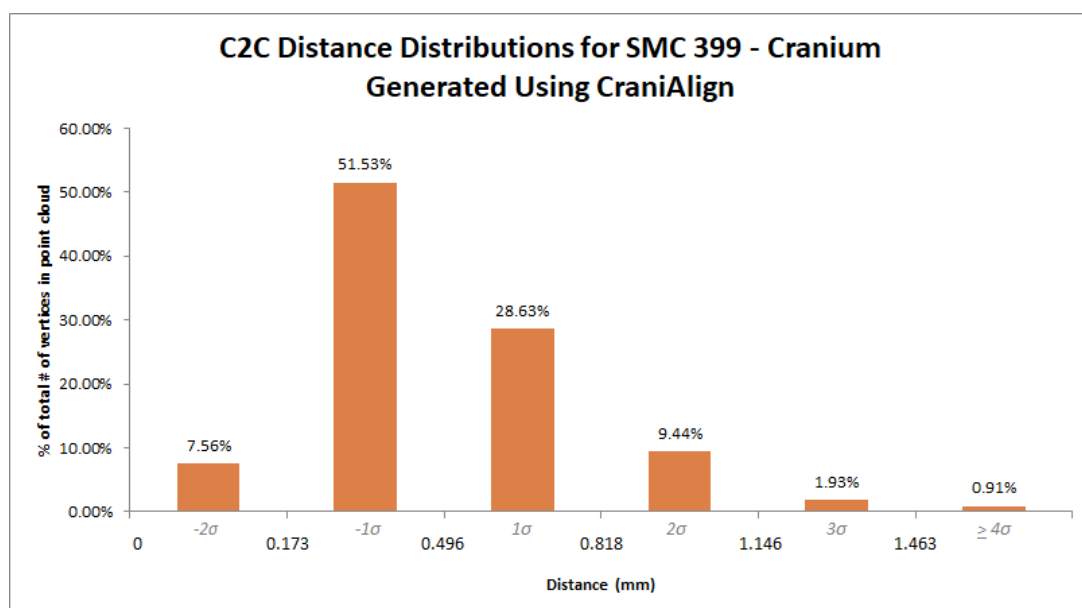
Figure I.2

Figure I.3

Figure I.4

Figure I.5

APPENDIX J: Probability Distribution Function Modelling for CraniAlign C2C Distributions

List of continuous random variables used:

- Alpha
- Beta prime
- Burr (Type III)
- Burr (Type XII)
- Chi
- Chi-squared
- Erlang
- Exponential
- Exponential power
- Exponentiated Weibull
- F-distribution
- Fatigue Life (Birnbaum-Saunders)
- Fisk (log-logistic)
- Folded Cauchy
- Folded Norm
- Fréchet right (variation of Weibull minimum)

- Gamma
- Generalized exponential
- Generalized gamma
- Generalized half-logistic
- Generalized Pareto
- Gilbrat
- Gompertz
- Half-Cauchy
- Half generalized normal
- Half-logistic
- Half-normal
- Inverse-gamma
- Inverse Gaussian
- Inverse Weibull
- Three-parameter kappa
- General Kolmogorov-Smirnov one-sided test
- Kolmogorov-Smirnov two-sided test

- Lévy
- Log-Laplace
- Log-normal
- Lomax
- Maxwell
- Mielke's Beta-Kappa
- Nakagami
- Non-central F-distribution
- Non-central chi-squared
- Power log-normal
- Rayleigh
- Reciprocal inverse Gaussian
- Rice (Ricean)
- Truncated normal
- Wald (variation of inverse Gaussian)
- Weibull minimum

To calculate bin sizes, the Freedman-Diaconis rule (1981) was followed (see H).

Table J.1: The parameters of the probability distribution functions (PDF's) that best approximated each sample's C2C distance distribution for CraniAlign samples.

Sample #	PDF	PDF Parameters
SMC206	Beta prime	(17.173, 9.390, -0.110, 0.257)
SMC399	Fatigue life	(0.497, -0.085, 0.517)
SMC417	Mielke's Beta-Kappa	(3.824, 2.587, -0.019, 0.294)
SMC1142	Burr (Type XII)	(7.988, 0.217, -0.175, 0.362)
SMC1248	Burr (Type III)	(3.455, 0.803, 0.003, 0.378)

Figure J.1: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC206, with all the fitted PDF's.

Figure J.2: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC206, with the best fitted PDF.

Figure J.3: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC399, with all the fitted PDF's.
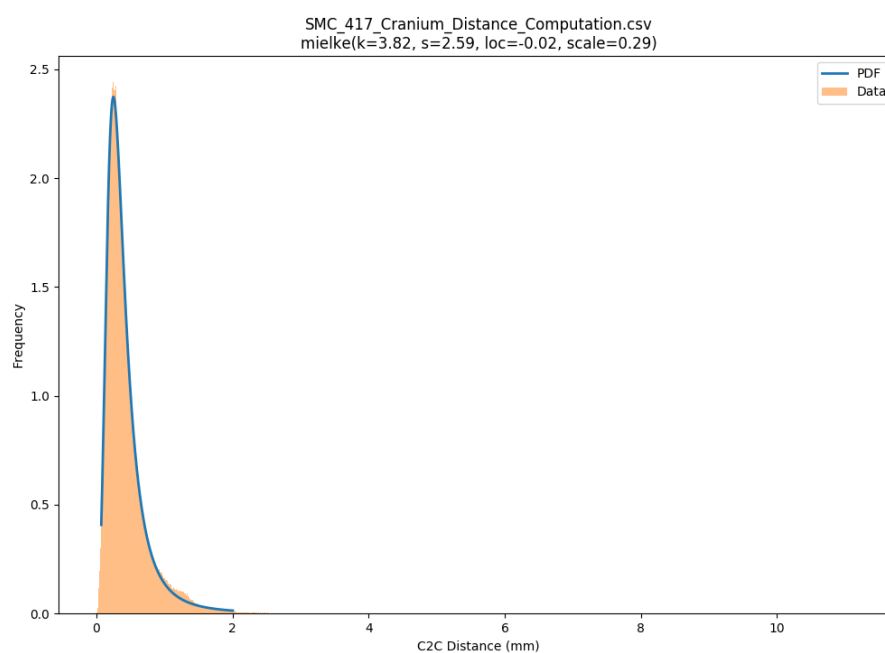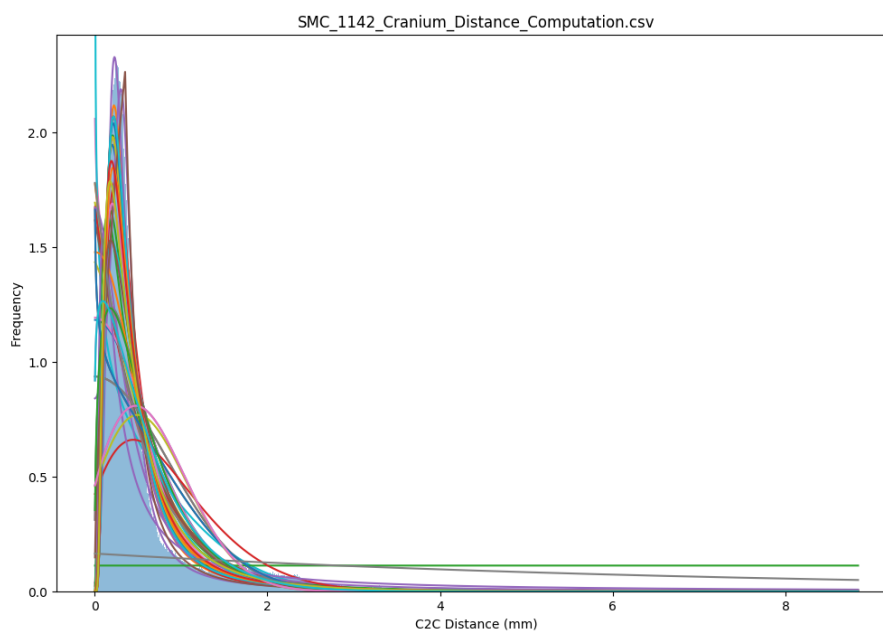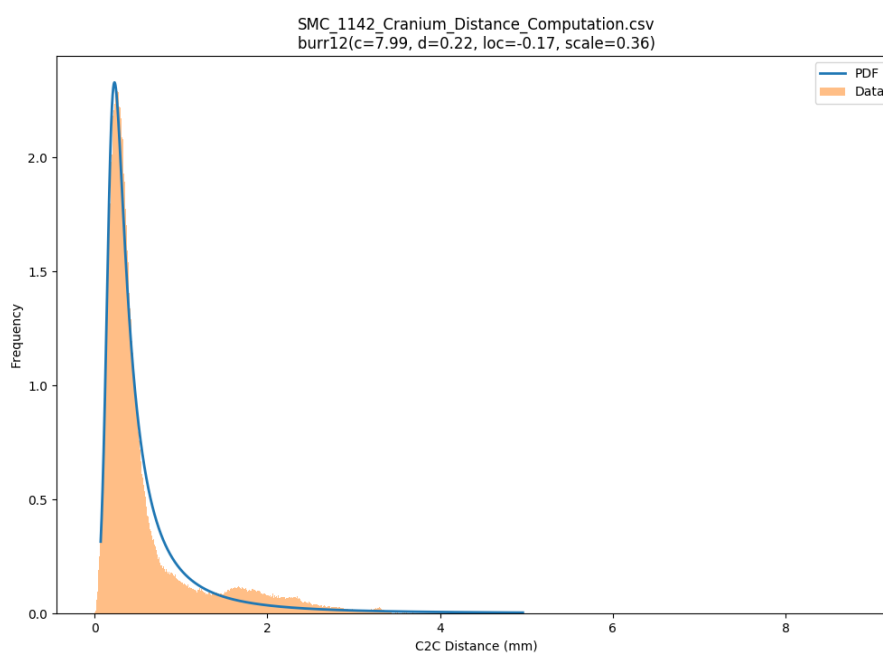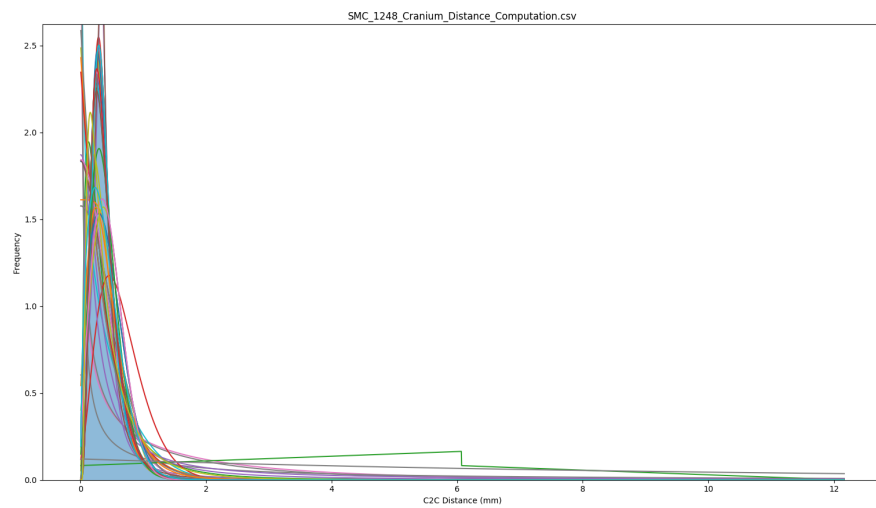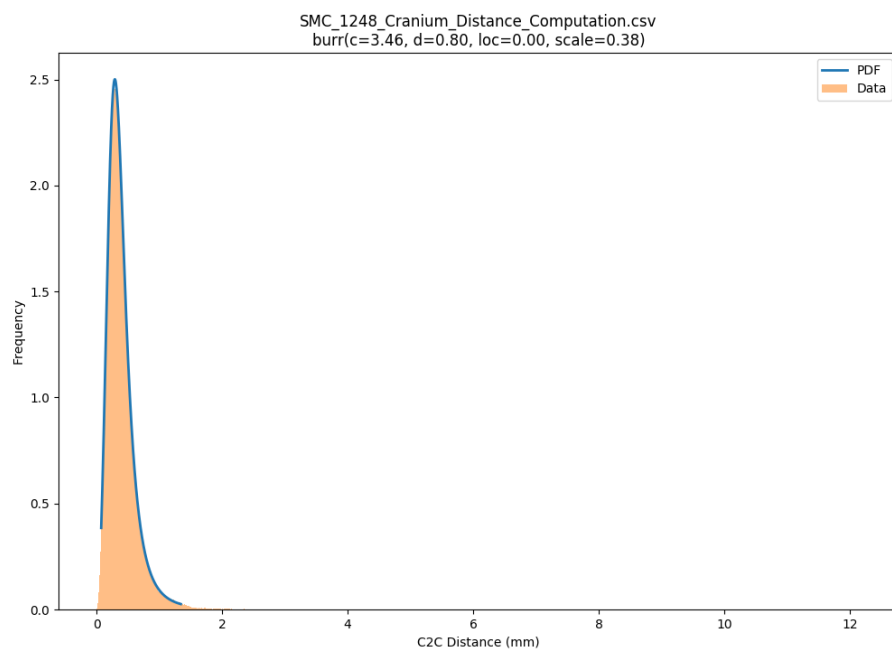

Figure J.4: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC399, with the best fitted PDF.
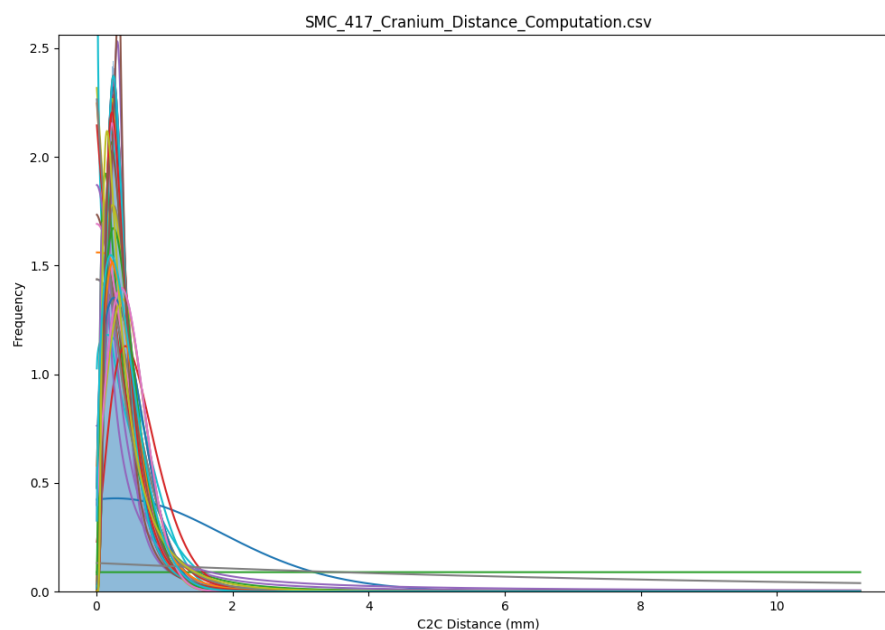
Figure J.5: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC417, with all the fitted PDF's.

Figure J.6: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC417, with the best fitted PDF.

Figure J.7: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC1142, with all the fitted PDF's.

Figure J.8: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC1248, with the best fitted PDF.

Figure J.9: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC1248, with all the fitted PDF's.

Figure J.10: A histogram of the C2C distributions from comparing CraniAlign alignments for SMC1248, with the best fitted PDF.

APPENDIX K: Validation Curves for the Sex Classification Model
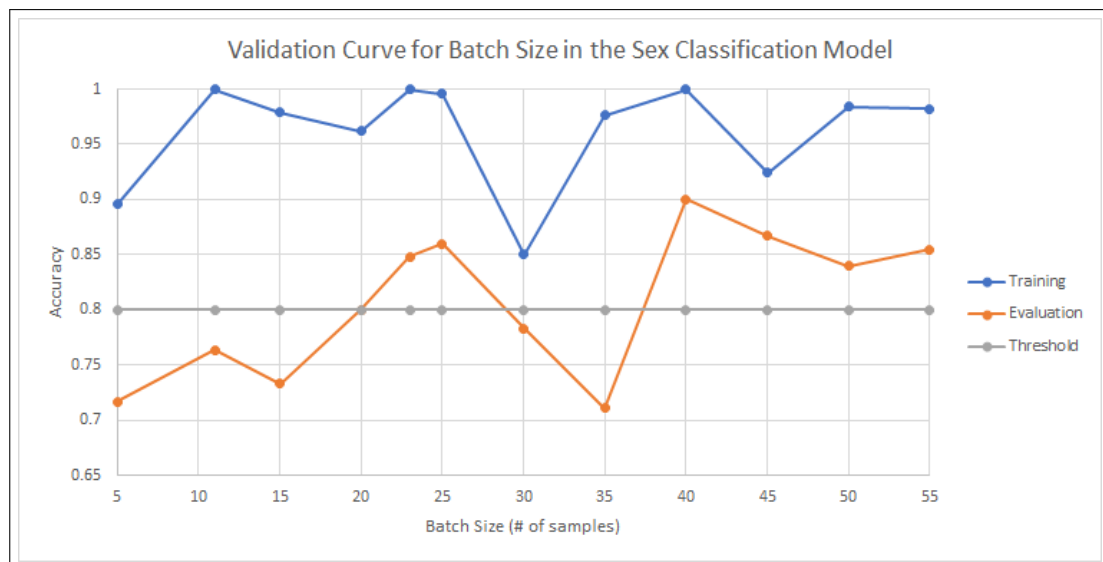
Figure K.1
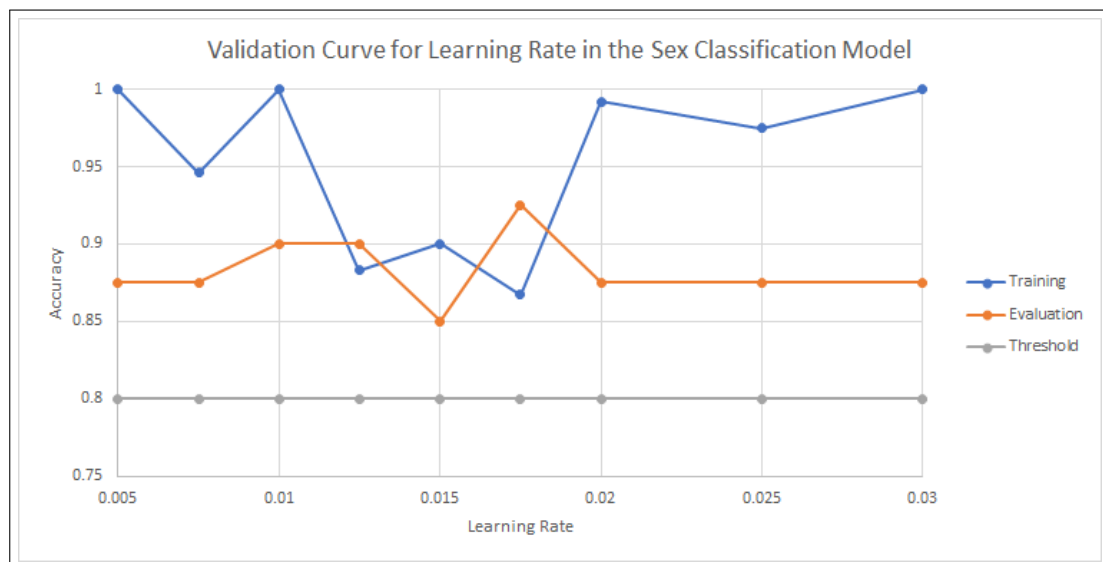
Figure K.2
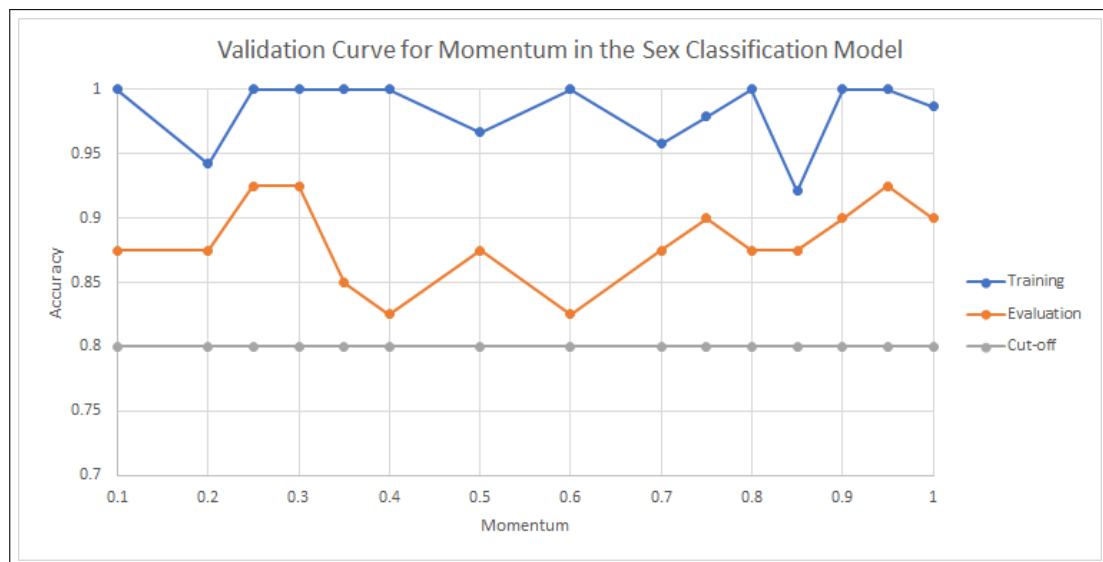
Figure K.3

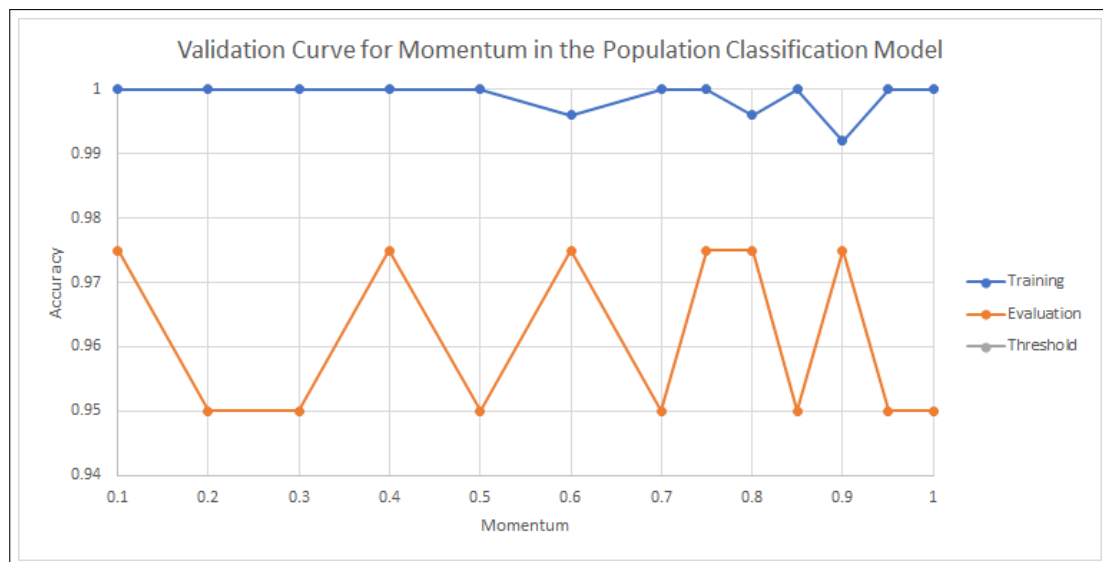# APPENDIX L: Validation Curves for the Population Classification Model
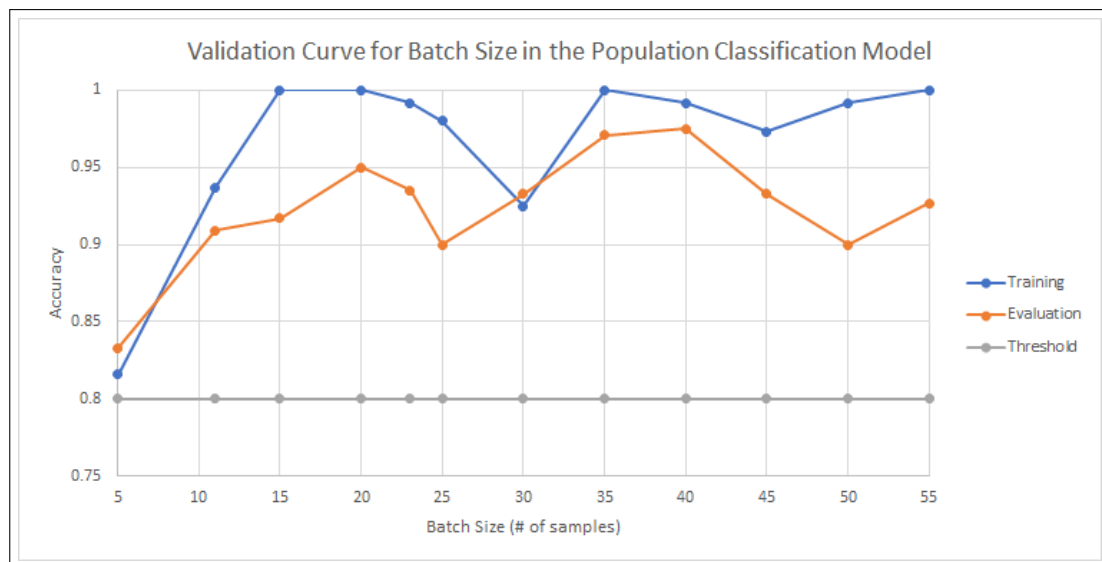
Figure L.1
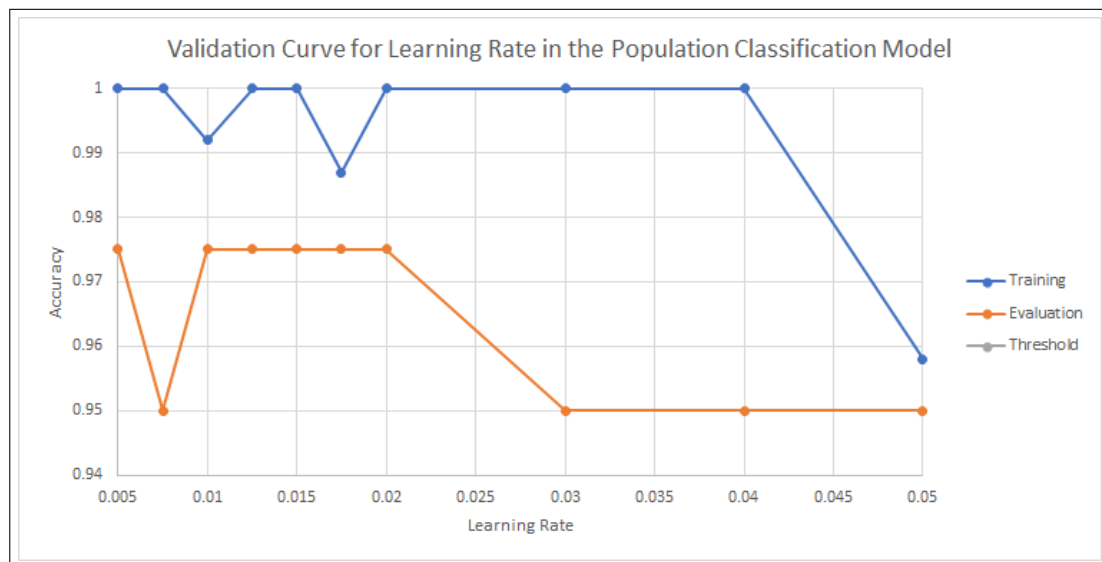
Figure L.2

Figure L.3

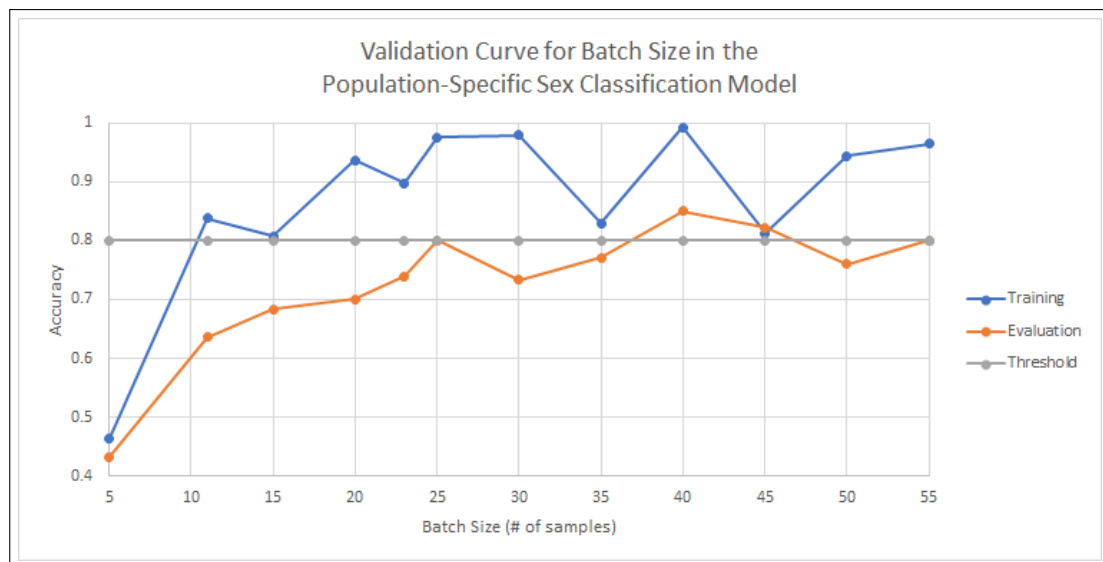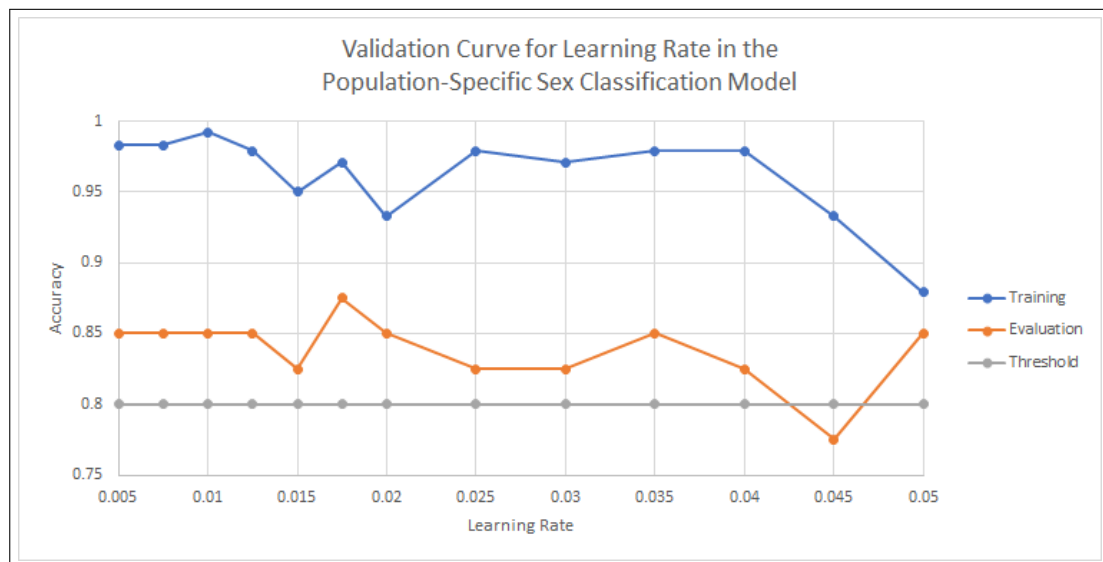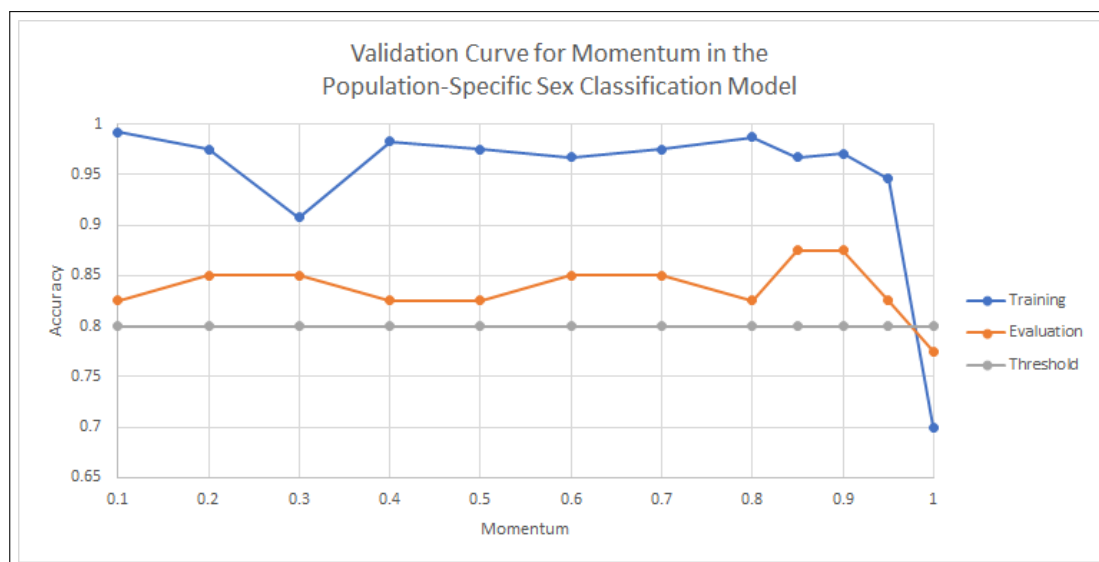# APPENDIX M: Validation Curves for the Population-Specific Sex Classification Model

Figure M.1

Figure M.2

Figure M.3