Addressing Neandertal evolutionary genetics at three different resolution levels: admixture with modern humans, demography and social structure

Federico A. Sánchez Quinto

TESI DOCTORAL UPF / 2014

DIRECTOR DE LA TESI

Dr. Carles Lalueza-Fox

Institut de Biología Evolutiva (CSIC-UPF)

A mi familia

Gracias por estar tan cerca, estando tan lejos

Acknowledgments

Después de casi cinco años de vivir en Barcelona, recuerdo que al llegar vi el momento de terminar mi doctorado muy lejano en el tiempo, tan lejano que pensé que nunca llegaría. Escribir esta tesis, ha tenido muchos momentos muy diferentes. Algunos los he disfrutado, otros no tanto, y de muchos he aprendido. En lugar de ser un logro individual, el trabajo científico que he desempeñado como parte de la tesis, ha sido un esfuerzo de grupo y por lo tanto hay mucha gente a quien agradecer. Sin lugar a dudas he disfrutado enormemente de mi tiempo en Barcelona, mi trabajo de investigación en el IBE y los diversos proyectos científicos en los que he estado involucrado ya sea para mi tesis u otras colaboraciones del grupo. Creo que he crecido mucho tanto académicamente como personalmente durante estos años, y sin lugar a dudas, muy buena parte de esto se lo debo a todas las personas con las que he interactuado. Estoy seguro de que sin su valioso apoyo, mucho de esto no habría sido posible. Durante las próximas líneas me gustaría tomarme un momento para dejar un breve reconocimiento a todos aquellos que de una manera u otra han contribuido que yo pueda estar escribiendo estas líneas en estos momentos.

En primer lugar, también me gustaría expresar mi más sincero agradecimiento a mi director de tesis, Carles Lalueza Fox, quien ha jugado un papel crucial para el desarrollo del contenido científico de ésta tesis y en mi formación como investigador. Habrá tiempo después de decir más palabras al respecto, pero concretamente quiero agradecerte por todo el apoyo científico y personal durante éstos últimos 4 años. Creo que ha sido mucho lo que he llegado a aprender con tu ejemplo, y te estoy muy agradecido por preocuparte de que sacará el máximo provecho posible de cada aspecto de mi experiencia académica. Gracias por la siempre muy buena disposición y excelente humor que me brindaste; los comentarios irónicos y humor negro fueron un muy buen plus! Sin embargo, sobre todas las cosas, gracias por esa gran trabajo de equipo que realizamos durante los últimos años, se le extrañará bastante.
Los siguientes agradecimientos van hacia todos los miembros (antiguos y presentes) del grupo de paleogenómica. Particularmente me gustaría agradecer a Óscar e Iñigo. Gracias Oscar por la amistad y en especial por el tiempo que le dedicaste a revisar la discusión de ésta tesis. A mi amigo Iñigo me gustaría no sólo agradecerle todos esos momentos en los que me echaste una mano y me sacaste de un apuro, sino también por todo el compañerismo y excelente trabajo de trabajo equipo que se creó cuando colaboramos.

Mis siguientes pensamientos van hacia todos los integrantes de BioEvo, particularmente con los que he interactuado directamente. Su compañerismo y buen humor hicieron que mi estancia en el PRBB fuera aún más placentera. Sin lugar a dudas, una de las razones por las cuales uno podría trabajar en éste sitio es por la gente con la que uno interactúa a diario. Particularmente me gustaría reconocer el trabajo del personal que se asegura de que todo se lleva a cabo como debe ser. Sin de su trabajo, la mayor parte del nuestro, sería imposible de llevar acabo. Un agradecimiento especial a Txema, Angel, Jordi, Mónica y Judith. Un agradecimiento también particularmente especial va para *Dipòsit de les Aigües*!, que fue mi lugar de retiro para escribir buena parte de ésta tesis. Hubiera sido imposible concentrarse para escribir en otro lugar que no fuera ese. Siguiendo con los agradecimientos, también me gustaría dar las gracias especialmente a tres personas del IBE que estuvieron a cargo de la gestión burocrática y el papeleo del financiamiento que me fue otorgado por parte del estado español. Muchas gracias a: Anna Pérez, Emiliano González y Rita Arias. Particularmente me gustaría agradecer la comprensión y la paciencia de Rita sobre mi forma de abordar la burocracia. Sin su ayuda, mis estancias de investigación fuera de España y muchos otros trámites, nunca hubiera podido llevarse a cabo.

Speaking about research stayings, I want to specially thank two research groups that hosted me during a couple of research internships carried out during my PhD. First I would like to thank Prof Tom Gilbert and his group for hosting me during a staying regarding for a collaboration project that we had in common; and which ended with a beautiful publication as the magazine cover of a journal! Amazing work! I would like to thank particularly thank Hannes Schroeder for carrying out the experimental part of the project, and in the process teaching cool details about the aDNA experimental work. My most sincere thanks go to Tom, who closely supervised my visit, make sure I had everything that I needed and kindly invited me to ISBA as well as many other research events while I was there.

The next research staying that I carried out was developed at the Max Planck institute. First of all, I would like to thank Prof Svante Pääbo for agreeing to my visit to the institute and supporting the projects that I carried out while I was there. I would like to thank Sergi Castellano and Aida Andrés for all

their scientific mentoring and personal support. Particularly Sergi, who was my official supervisor and gave me both good professional and personal advice. Also I will like to thank the crowd at Leipzig who made my staying more enjoyable both within and outside the MPI. Special thanks go to Fred, Madel, Martin K, Jesse, JuanRa, Futbol people, Giovanna, Tobias, Ana Lucero & Co. En especial muchas gracias a ti Gio y tu banda; gracias por hacerme sentir cómo en casa.

Por otra parte, me gustaría agradecer a todas las personas que han contribuido directamente a mi bienestar personal durante mi estancia en Barcelona.
Entre ellos, me gustaría empezar dando las gracias a todos los que han sido habitantes de Aragó 396 # 3-3; mi cuartel general mientras estuve gran parte de mi tiempo en Barcelona. Así que muchas gracias a todos desde tiempos lejanos cuando vivía con Pablo y Gitza hasta cuando compartí piso con Clumps, Sebastian, Lili y Jorge. Muchas gracias por todos aquellos buenos momentos y por entender mis difíciles horarios de trabajo.

Siguiente línea de agradecimientos personales, mis agradecimientos también van para todos los ex alumnos de la LCG y amigos con los que conviví, ya fuera en eventos organizados por nuestra asociación o que tuve el gusto de visitar; particularmente gracias a la banda que organizo los eventos. Muchas gracias a María Gutiérrez, Sandra, Marce, Fernando (esos viajes a Viena fueron lo máximo), Ana Gutiérrez, Drulto y María Ávila. Varios agradecimientos personales van para Mari (Miss Mari como le gusta hacerse llamar). La sobriedad de éste texto no me permite entrar en detalles, pero muchas gracias por todo el hospedaje, apoyo, asociación científica y amistad que me has ofrecido a lo largo de estos años. Muchas felicidades todo el éxito laboral y personal que estas teniendo ahora mismo, y ojala algún día se nos haga ese instituto científico en Puerto Escondido.
Muchos agradecimientos también van para mi familia en Barcelona: la banda del Tupper club. Muchas gracias por todos esos magníficos almuerzos juntos. Cada día, fue una experiencia increíble y única, me dio mucho gusto haber formado parte de éste grupo políticamente incorrecto, anarquista y *bully*. Fue un honor haber sido parte de éste experimento y gracias por su apoyo cuando tuve que ausentarme de él por un tiempo. Sin lugar a dudas una de las cosas que no me gustaron de haber escrito ésta tesis, fue alejarme del tupper club. Ojalá haya logrado mi cometido al mostrarles un poco de la comida de México y haya despertado su interés por los muchos placeres de nuestra gastronomía. Es difícil vivir sin el tupper, y tomándome la libertad de citar a uno de nuestros miembros, sin lugar a dudas “El taperclub es lo mejor que me

ha pasado en el doctorado...de lejos!". Luego habrá tiempo de dar agradecimientos más individualizados y emotivos pero por mientras tanto, muchísimas gracias a miembros presentes y ausentes como: Aléx, Alicia, Diego, Elena, Johannes, Juan, Katharina, Ilaria, Lara, Lisa, Marc, Marco, María, Miruna, Nino, Pierre y Vero; su entrañable amistad ha sido un factor fundamental para que yo disfrutará aún más mi estancia en Barcelona.

Por otro lado, muchas gracias también por toda la amistad a Neus, Javi, Ignasi, Nuria, Marina y la banda de mi despacho. Con algunos tuve un trato más cercano que con otros, pero que en cualquier caso fue muy grato haberlos conocido, molan mogollón!

En este momento también tengo que agradecer muy especialmente todo el compañerismo y apoyo de algunas personas del tupper club. A María "coleguilla", muchísimas gracias por todo el lio legal/burocrático que te aventaste hace ya unos ayeres para ayudarme a tramitar mi NIE a distancia. Nunca más! Muchas gracias por esa muestra de apoyo y por muchas más, que me has demostrado a lo largo de los años. A Marc muchísimas gracias por todo el apoyo y ayuda que tuviste conmigo al principio, y en especial para el artículo del Current Biology. Tú participación fue muy importante para que pudiera sacar ese proyecto adelante. Pierre, tu apoyo también ha sido fundamental. No sólo a echarme la mano con algún trámite cuando lo necesité, pero también tu opinión científica, asistencia técnica y "canciones del día" me ayudaron a que esté escribiendo estas líneas en éste momento. Diego, carnal. ¿Qué decir?, tu apoyo incondicional y sabio consejo me han llevado muy lejos. En concreto muchas gracias por tomarte el tiempo de leer parte de ésta tesis y hacerme muy buenos comentarios al respecto. Nunca dejes la ciencia, el mundo académico te necesita!

La siguiente sección está dedicada a una muy especial persona que conocí hasta el final de mi estancia en Barcelona, pero que ha transformado mi vida en muchos aspectos. Anna, muchísimas gracias por todo tu apoyo incondicional que me ofreciste particularmente durante varios momentos difíciles en los últimos meses de escribir ésta tesis. Tu cariño y palabras de aliento han sido fundamentales para que yo pudiera haber terminado este manuscrito. Las siguientes líneas son tuyas.

Ένα μεγάλο ευχαριστο, στο αγαπημένο μου ηφαίστειο Ελληνικών συναισθημάτων. Ευχαριστώ για όλη την υποστήριξη κατά τη διάρκεια συγγραφής αυτής της διπλωματικής εργασίας. Χωρίς αυτή δεν θα ήταν

δυνατό να γράψω αυτές προτάσεις τώρα. Σ' αγαπάω πολύ και περιμένω με ανυπομονησία να δω τι επιφύλασει το μέλλον για μας

Por último, agradecimientos muy muy especiales van para mi familia. Muchísimas gracias por siempre haberme apoyado, mandado souvenirs a distancia y que me han tenido en su pensamiento. Qué decir del apoyo incondicional de 4 personas fundamentales para mí: mis padres y hermanos. Güitos y Pepis, muchas gracias por su apoyo, entendimiento cuando he tenido que ausentarme y cariño incondicional, los tengo muy presentes. Finalmente, los responsables directos de que este por conseguir éste mérito académico son mis padres. Infinitas gracias por de una manera directa o indirecta inculcarme éste gusto por la ciencia, por todo su apoyo moral e inclusive económico para que yo pudiera estudiar lo que yo deseaba y alcanzar mis metas profesionales. Sin lugar a dudas éste logro es en gran parte de ustedes también. No ha sido fácil pasar 5 años a distancia, pero su cariño y apoyo incondicional me han hecho mucho más fácil el camino. Las palabras no me alcanzan para agradecerles por todo lo que me han dado, sólo queda mi admiración e inmenso cariño por ustedes.

Federico Sánchez Quinto a 17 de Septiembre de 2014. Barcelona, España

This research was supported by a PhD fellowship ”Formación de Personal Investigador” BES-2010-030405 and part of the Proyecto Nacional BFU2009-06974 funded by El Ministerio de Competividad y Economía de España.

Abstract

Almost 20 years of Neandertal paleogenetics studies have significantly increased our knowledge about their evolutionary history. The analysis of DNA recovered from Neandertal remains to date, suggest that although they were a distinct hominin population to modern humans, a certain degree of gene flow occurred between the two of them. Furthermore, recent evidence suggests that archaic introgressed material could have been biologically relevant for modern humans to adapt to new environments. Moreover, insights from a wide geographic and temporally different sampling of Neandertal mitochondrial sequences and from a high-coverage genome, suggest that Neandertals probably had a low effective population, which was possibly decreasing towards the end of their evolutionary time. This thesis focus to address the evolutionary genetic history of Neandertals at three different levels of resolution from: analyzing further aspects of their relatedness to modern humans, better characterizing their population history and identify the genetic basis for some of their distinctive morphological features, to describing their genetic structure within a social group. Insights from these three lines of research intend to reconstruct key aspects of their population history and its implications towards their eventual demise.

Resumen

Casi veinte años de estudios de paleogenética Neandertal han incrementado significativamente nuestro conocimiento sobre su historia evolutiva. El análisis de secuencias genéticas recuperadas a partir de fósiles Neandertales, sugiere que a pesar de que éstos era un grupo de homínidos diferentes a los humanos modernos, cierta grado de introgresión genética ocurrió de Neandertales hacia humanos modernos. Más aún, estudios recientes sugieren que el material genético introducido a éstos pudo haber sido relevante biológicamente para adaptarse a nuevos ambientes. Por otro lado, inferencias a partir de datos genéticos mitocondriales provenientes de muestras de diferentes zonas geográficas y origen temporal, a la par con la secuencia de un genoma completo de alta calidad sugieren que los Neandertales tenían un tamaño efectivo de población reducido y que probablemente estaba disminuyendo hacia el final de su tiempo. La tesis aquí presentada, se enfoca a abordar la historia evolutiva Neandertal a tres niveles de resolución diferentes, analizando datos genéticos provenientes de fósiles. Primero, se analizan otros posibles eventos de introgresión genética con humanos modernos, no descritos hasta la fecha. Posteriormente, se caracteriza a detalle su demografía e identifica cambios específicos para su linaje evolutivo que podrían estar relacionados con las bases genéticas de algunos de sus rasgos morfológicos más distintivos. Finalmente, se describe la estructura genética y dinámica de un grupo social Neandertal. Las perspectivas de estas tres líneas de investigación pretenden no sólo reconstruir aspectos claves de su historia evolutiva, sino también entender las consecuencias que ésta pudo haber tenido con su eventual extinción.

Resum

Gairebé vint anys d'estudis de paleogenètica Neandertal han incrementat significativament el nostre coneixement sobre la seva història evolutiva. L'anàlisi de seqüències genètiques recuperades de fòssils neandertals fins a l'actualitat, suggereix que, tot i que aquests eren un grup d'homínids diferents als humans moderns, es va produir un cert grau d'introgressió genètica de Neandertals cap a humans moderns. Encara més, estudis recents suggereixen que el material genètic introduït a aquests podria haver estat biològicament rellevant per adaptar-se a nous ambients. D'altra banda, inferències a partir de dades genètiques mitocondrials provinents de mostres de diferents zones geogràfiques i diferents origen temporals, junt amb la seqüència d'un genoma complet d'alta qualitat, suggereixen que els Neandertals tenien una mida efectiva de població reduïda i que probablement s'anava disminuint cap al final del seu temps. La tesi aquí presentada, s'enfoca en abordar la història evolutiva Neandertal a tres nivells de resolució diferents, mitjançant l'anàlisi de dades genètiques provinents de fòssils. Primer, s'analitzen altres possibles esdeveniments d'introgressió genètica amb els humans moderns, no descrits fins ara. Posteriorment, es caracteritza amb detall la seva demografia i s'identifiquen canvis específics en el seu llinatge evolutiu que podrien estar relacionats amb les bases genètiques d'alguns dels seus trets morfològics més distintius. Finalment, es descriu l'estructura genètica i dinàmica d'un grup social Neandertal. Les perspectives d'aquestes tres línies d'investigació pretenen no només reconstruir aspectes claus de la seva història evolutiva, sinó també entendre les conseqüències que aquesta podria haver tingut amb la seva eventual extinción.

Traduït per Marc Pybus

Preface

Discovered in the 19th century, Neandertals – an extinct population of hominins named after *Neander valley* (Germany) where their fossils were initially recovered – have had a paramount role in the understanding of our evolutionary history and to identify our uniqueness as a species.

It was the discovery of Neandertals what ultimately caused, coupled to the publication of "The origin of species" and "The descent of man" by Charles R. Darwin, not only a radical change to the perception of our origins, but the emergence of paleoanthropology -the study of human evolution from morphological and archeological features found in fossil remains-.

After almost 150 years of addressing human origins with such a scope in addition to insights from genetic data from present-day human and great apes, have contributed to elaborate a first integrative perspective regarding the beginning of our species. Data showed that while our evolutionary lineage shared a common ancestor with chimpanzees around 6 million years ago and our species originated within Africa ~200,000 years ago, out of a phylogeny of hominins that existed after the split with the ancestor with chimpanzees, Neandertals were our closest known relatives. Such relatedness scenario, allowed addressing the long-lasting debates about our evolutionary success. Analysis based on genetic comparisons to our closest living relatives (great apes and monkeys), uncovered genetic changes not shared by other primates. However, given the long evolutionary distance to them, most of those

genetic differences would not be specific to our species. An alternative approached based on the origin and evolution of Neandertal morphological features and the stone tools that they produce (similar to those of other hominins while different from modern humans) provided an approach to characterize more precisely traits and features unique to our lineage. However, the scarcity and fragmentary nature of their fossil remains and the possible occurrence of cultural transmission between hominin groups complicated the exclusive use of morphological and archeological data, to accurately reconstruct the Neandertal evolutionary history and elucidate the basis of our evolutionary success. Curiously, it was again data from Neandertals genetic data, obtained by breakthrough technological innovations to extract DNA from fossil remains, what finally contributed to clearly define their relatedness to modern humans and identify the genetic basis that might be responsible for our uniqueness as a species.

This thesis is superimposed on the conceptual revolution to extract genetic data from ancient remains in order for to better characterize human origins and evolution. While genetic changes unique to our lineage provide a ground basis to address the nature of our evolutionary success, an accurate reconstruction of the Neandertal evolutionary history (as well as that of other hominins), will continue to shred light about our singularity as a species by addressing, why after living over such a long evolutionary time other hominins got extinct whereas modern humans prevailed.

Table of contents

Abbreviations

A: Adenine
aDNA: Ancient DNA
AMH: Anatomically modern human
Bp: Base pair
BWT: Burrows–Wheeler transform
C: Cytosine
COB: Capture on beads
CNV: Copy Number Variants
DNA: Deoxyribonucleic acid
G: Guanines.
LD: Linkage disequilibrium
Kb: Kilobases
Ne: Effective population size of a population
NGS: Next Generation Sequencing
NT: Nucleotides
MRCA: Most common recent ancestor
MYA: Millions years ago
PCA: Principal component analysis
PCR: Polymerase chain reaction
PE: Pair-end
PEC: Primer extension capture
PSMC: Pairwise sequentially Markovian coalescent
SAM: Sequence Alignment/Map
SE: Single end
SNV: Single Nucleotide Variants
SV: Structural variants
SNP: Single nucleotide polymorphism
T: Thymine
TGS: Third generation sequencing methods
WISC: Whole genome capture method

List of figures

List of Tables

I INTRODUCTION

1.1 The Neandertals in the context of primate and hominin evolution

Neandertals are an extinct population of hominins that once lived across Eurasia from several hundreds of thousand years ago until around ~40,000 years ago.

Neandertals have always fascinated scientists and the public alike, mainly due to their wide, heavily muscled bodies and rough appearance. However, because of that distinctive look, unfortunately they have also been portrayed by some sectors of society as mere brutes, troglodytes or regarded as evolutionary irrelevant. However, Neandertals are essential to the understanding of human evolution.
The first Neandertal type specimen (**Figure 1**) was discovered in 1856 near Düsseldorf (Germany) by lime quarry workers at a cave called "Feldhofer Grotte" in the Neander Valley. Thinking that the fossils belong to a cave bear, the remains were brought to a local teacher and amateur naturalist, Johann Karl Fulrott, who recognized them as having rather a "humanoid" origin but at the same time being different from

Figure 1: Skull-cap of Feldhofer 1 Neandertal.
[Adapted from http://www.talkorigins.org/faqs/homs/feldhofer.jpg]

contemporary humans. Seeking a second opinion, he showed the remains to Hermann Schaffhausen, a professor of Anatomy at the University of Bonn. After a careful assessment, in 1857, they announced the discovery stating that in their view those bones represented the remnants of a previously unknown type of human race, different from the human skeletal morphology known at the time (Tattersall 1996). It took quite some time for this anthropological discovery to become accepted, as it contradicted the main religious notion of the origin of man and Charles Darwin's work had not yet been published. For instance, early proposals suggested that Neandertal skeletons simply represented the pathological remains of modern humans; for example, Rudolf Virchow asserted that the Neandertal skeleton belong to a rachitic individual (Virchow 1872), while J. Barnard Davis maintained that the Neandertal skull presented an extreme case of synostosis -fusion of two or more bones (Davis 1867).
Although the idea that humans were morphologically and anatomically more similar to great apes than to other species, had been obvious to people for some time, the notion that all species shared a common ancestor and that some of the biological differences between species could be explained by means of natural selection, were not consider until after Charles Darwin publication of "On the Origin of Species" in 1859. As Darwin´s ideas started to permeate into the scientific community, other scientists like Thomas Henry Huxley began to incorporate them to explain the origin of humans. Moreover, as more humanoid fossil remains with archaic features continued to appear, in addition to a change of perspective of human origins, paleoanthropology was eventually established. Finally later on the 20 century, as more fossils with consistent Neandertal appearance accumulated, in addition to a scientific multidisciplinary approach to the study of human evolution, alternative pathological explanations became untenable (Stringer and Gamble 1994; Trinkaus and Shipman 1992) and eventually they were recognized as an extinct hominin form.

Moreover, after several decades of morphological and anatomical comparison of modern humans with great apes and monkeys, it is commonly believed that great apes are our closest living relatives. However, which specific great ape is the closest to us, was for a long time contentious. Genetic approaches allowed us to investigate evolutionary relationships between humans and great apes in much greater detail. Thus, by studying the deoxyribonucleic acid (DNA) complete sequences of both human and the other great apes genomes, we have confirm that humans are more closely related to apes (chimpanzees, bonobos, gorillas and Orangutans) than to other organisms. Moreover, present-day humans are ~98.8% similar to chimpanzees -at the DNA sequence level- and genetic differences between species suggest that the split from the common ancestor of the two species occurred about 6-7 million years ago (MYA), while the human-chimpanzee-gorilla speciation happened about 8-10 MYA, and the human-chimpanzee-gorilla-orangutan 12-14 MYA (**Figure 2**) (Rogers and Gibbs 2014).

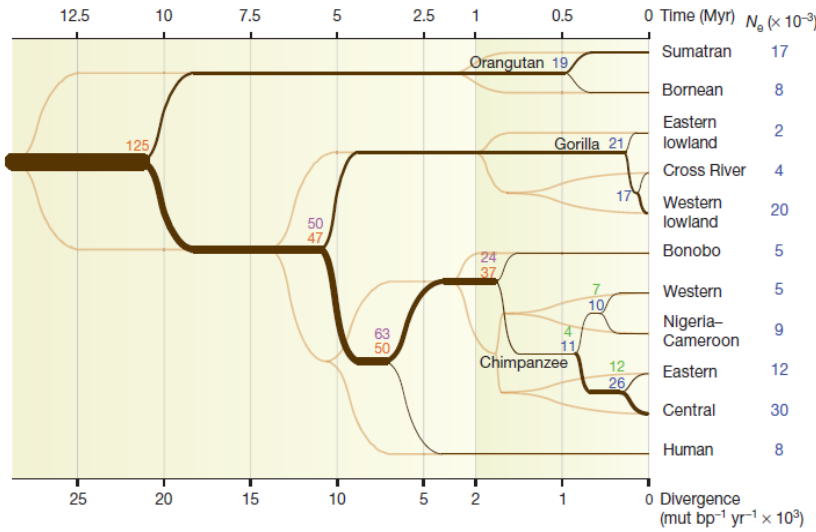


Figure 2: Population splits and effective population sizes (Ne) during great ape evolution. Split times (dark brown) and divergence times (light brown) are plotted as a function of divergence on the bottom and time on top. Time is estimated using a single mutation rate of 1 x 10^{-9} mut bp^{-1} $year^{-1}$. The ancestral and current effective population sizes are also estimated using this mutation rate. The x axis is rescaled for divergences larger than 2 x 10^{-3} to provide more resolution in recent splits. The terminal Ne corresponds to the effective population size after the last split event. [Adapted from (Prado-Martinez et al. 2013)].

On the other hand, paleoanthropological findings during the last ~150 years have revealed that we were not alone in our evolutionary lineage after the split with chimpanzees, and thus many other species of hominins –a term reserved for all kinds of human species and their immediate ancestors - existed but became extinct.
For instance, examples of earliest hominin fossils from that time till about 4.2 MYA are largely represented by *Sahelanthropus tchadensis, Orrorin* and *Ardipithecus* (Brunet et al. 2002; Senut et al. 2001; White et al. 2014). Furthermore, fossils dated after about 4.2 MYA and before the appearance of the Homo, are usually associated to the genus of Australopithecus. Various *Australopithecus* species (*A. afarensis, A. anamensis and A. africanus*) have been found in multiple sites in east and southern Africa, dating from around 4 MYA to 1.8 MYA; being *A. anamensis* generally considered the oldest one. The most well-known fossil of *Australopithecus* is the partial skeleton "Lucy", dated to 3.2 MYA, as well as the Laetoli footprints 21, dated to 3.5 MYA, both of these belong to the species *Australopithecus afarensis*. A common feature of all hominins is their unequivocal display of bipedal locomotion. Moreover, it is still under debate, which species or fossils of *Australopithecus* represent the ancestor of our own *Homo* genus, but *afarensis* and *africanus* have been considered as good candidates *africanus* (Jobling et al. 2014) (see **Figure 3**).

Furthermore, the origin of the ***Homo*** genus is a contentious matter. While for many years, *H. habilis* was considered the earliest member of the genus (L. Leakey, Tobias, and Napier 1964), recent publications refer to it as *A. Habilis* instead (Wood 1996; Wood B & Collard M 1999). However, the origin of the genus continuous to be highly debated issue. While some consider *H. ergaster* is represented by mainly African individuals and reserves *H. erectus* for those found outside Africa, an alternative analysis would include all these early *Homo* specimens as a single widespread and variable species, *H.*

erectus; being the latter classification the most popular view in recent times (Jobling et al. 2014).

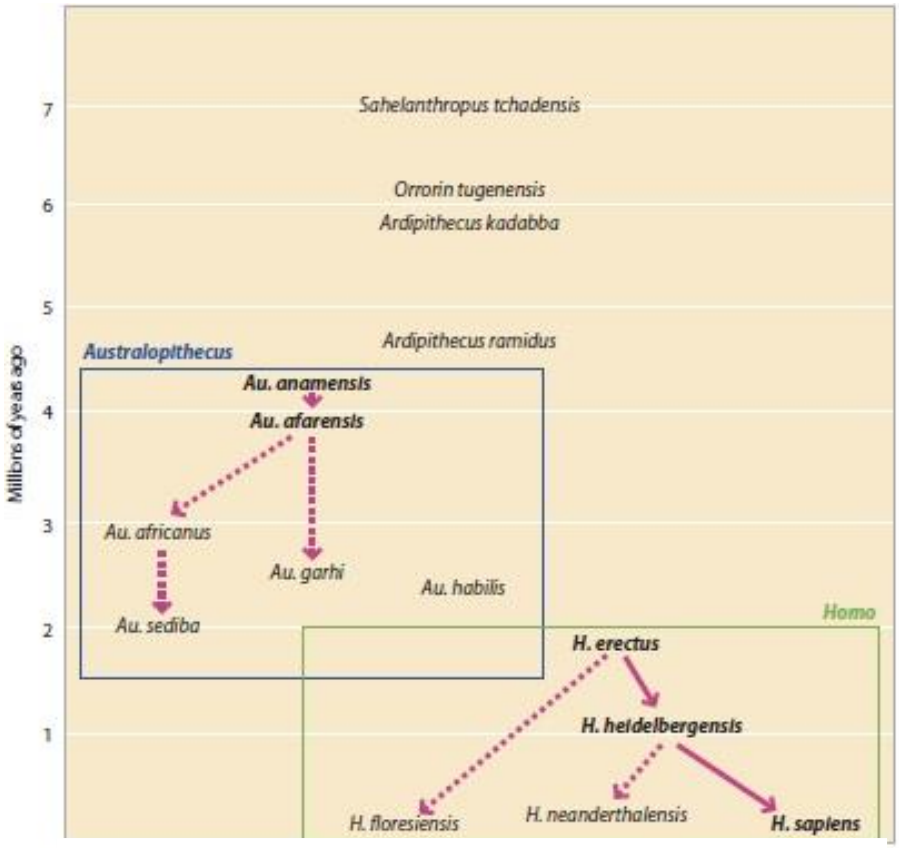


Figure 3: Relationships of fossil hominin species, indicating plausible human ancestors
[Adapted from (Jobling et al. 2014)].

The first indisputably specimens of H. *erectus* date between 1.8 and 1.9 MYA (Pickering et al. 2011) and, like all earlier hominins, are found in Africa, further suggesting an African origin for our genus (W. A. & Leakey 1993). Moreover, recent findings suggest hominin remains found at Dmanisi (Georgia) and dated to ~1.8 MYA, are those of *H. erectus*; making it the earliest hominin to be found outside Africa (Lordkipanidze et al. 2007; Swisher III et al. 1994; Ferring et al. 2011).

Other *H. erectus* found outside African and which confirm their range throughout Asia are the well-known fossils of the "Java Man" (Indonesia) and "Peking Man" (China). Furthermore, it has been suggested that the large body size of *H. erectus* provided tolerance to heat stress and dehydration, and coupled with its particular improved stone toolkits, may have allowed this species to live in a wide range of environments and thus expand out of Africa rapidly (Wood 1996). Although these hypothesis need to be properly addressed, it is compelling to think that some sort of adaptive process must have occurred to these first hominins migrating out of Africa, so they could eventually settled down in environments notably different from the sub-Sahara habitat where they are thought to have first evolved.

Moreover, in 2004 *H. floresiensis*, a tiny 1-m-tall hominin species and named after its place of discovery, was found in the island of Flores in Indonesia (Brown et al. 2004). Even though, it has been suggested that *H. floresiensis* might represent modern humans suffering from microcephaly, the finding of the remains from multiple individuals and archaeological deposits spanning the period 17,000–74,000 years ago, including the fairly complete 380 cc cranium and skeleton of the type specimen (LB1), has led most paleontologists to accept them as distinct hominins (Brown et al. 2004); probably descendants of *H. erectus* surviving on an isolated island poor in resources since ~1 MYA and resulting in selection for a dwarf hominin version of *H. erectus*.

Later *Homo* from Africa and Europe are less robust and have larger brains (~1200 cc instead of ~900 cc on average) than early *H. erectus* and are often designated *H. heidelbergensis*, the type specimen based on a ~609,000 Jaw from Heidelberg (Germany) **(Figure 3)** (Wagner et al. 2010). Specimens with similar morphological characteristics include the massive Bodo cranium (Ethiopia, ~600,000 years ago), the tibia (lower leg bone) from Boxgrove (England, ~500,000 years ago), and the Petralona 1 cranium (Greece, age is uncertain but with estimates

between 200,000 and 700,000 years ago). In addition, many would also place the 1.2–0.80 MYA specimens from Gran Dolina and Sima del Elefante, Spain, designated *H. antecessor* by their discoverers (Bermudez de Castro 1997) within *H. heidelbergensis*, although a consensus is yet not achieved. According to this view, *H. heidelbergensis* would have been a widespread and somewhat variable species, perhaps originating from *erectus* in Africa some time prior to 1 MYA and giving rise to more recent *Homo* species, including anatomically modern humans and Neandertals (J. Hublin 2009).

However, other researchers, however, prefer to call the post ~1 MYA African specimens *H. rhodesiensis* after the ~300–125-KY-old Kabwe (or Broken Hill 1) skull from Zambia. In South Africa, other potential specimens assigned to *H. rhodesiensis* include the Saldanha Man skullcap from Elandsfontein dated between 1.1 and 0.6 MYA and the Cave of Hearths material which perhaps dates to between 800,000 and 400,000 years ago (Herries 2011). In this more complex scenario, which will not be further discussed here, *H. heidelbergensis* is a European species giving rise to Neandertals, while *H. rhodesiensis* an African species being the ancestor of *H. sapiens*(J. Hublin 2009).

Furthermore, hominin remains with derived morphological features like Neandertals (previously refereed as *H. neandertalensis)* are commonly found in the fossil record from the Middle to the Late Pleistocene, from the Iberian Peninsula in Spain up to the Altai Mountains in Siberia. While a broader analysis and discussion about their origin and evolutionary history is review in the next sections, it is interesting to point out that yet another group of archaic hominins - named Denisovans (named after Denisova cave, a cavern within the Altai Mounts in South Siberia (Russia) where they were discovered) - and shown be closely related to Neandertals, also cohabitated the earth during Middle-Late Pleistocene (see **sections 2.5-2.7**).

Finally, the origin of modern humans has probably been the most contentious issue in the field over the last 30 years, and obviously there is an important distinction between morphology and behavior when regarding this issue. Anatomically modern humans (AMH), also known as modern humans or simply *Homo sapiens* (which are us) differ from earlier hominins ("archaic humans" or "archaic *H. sapiens*") mainly on a couple of cranial features, such as extent of the globular shape of the skull and degree of retraction of the face. Using these two features as system of comparison allows a clear distinction between AMH and archaic humans, with zero overlap, but has the disadvantage that relatively complete specimens are needed and to analyze fragmentary specimens it is necessary to use less reliable criteria.

The oldest fossil classified as AMH have been found in Omo-Kibish (Ethiopia) and dated to ~ 195,000 years ago (McDougall, Francis, and Fleagle 2005). Moreover, other well-known fragmentary specimens are known from Klasies River Mouth in South Africa at 90,000–120,000 years ago, and two sites from the Qafzeh cave in Israel, dated to between 90,000 and 130,000 years ago.
Furthermore, although previously contentious, the African-origin of modern humans and its eventual dispersion to the rest of the world by a major out-of-Africa migration now it is widely accepted. Albeit the initial out-of-Africa routes remain under debate, according to genetic and archeological evidence, human expansions reached Australo-Melanesia by ~60,000, Europe by ~45,000, East Asia by ~35,000, the American continent by ~15-20,000 and eventually the Pacific edge some ~3,000 (Henn, Cavalli-Sforza, and Feldman 2012).

Thus, while our genus has a long evolutionary history, the emergence of humans that are morphologically identical to present-day people is quite recent. Interestingly, within the last 50,000 of our history, we have become the most dominant primate species on the planet. We have developed an incomparable capacity for culture that has allowed

us to grow in number, to extend our range to almost all regions of the planet, and to impact for better and for worse the lives of many other animals, plants, and ecosystems. Hence, there has been an historic interest in understanding our singularity as a species and to define *"what makes us humans"*.

Precisely this is where the evolutionary importance of Neandertals resides. First, the fact that anthropological evidence points out that they were our closest known relative makes them ideal candidates to identify those traits that might have originated within our own evolutionary lineage, and can help us to characterize *what define us as a species*. Moreover, reconstructing key aspects of their demographic and evolutionary history could give us insights about of why, even though they seem to have lived such a long time, they eventually became extinct.
A review of Neandertal morphology, biology and behavior is necessary in order to set up a framework under which genetic data from Neandertals can be discuss to better understand their evolutionary history.

1.2 Neandertals morphological description and geographical range

1.2.1 The origin of Neandertals and their morphological features

The first appearance in the fossil record of morphological traits associate to Neandertals remains contentious. Nonetheless, the most accepted model to try to explain the Neandertal origin is the ''Accretion Model'' (Rosas et al. 2006; J. Hublin 2009), which accounts for the progressive appearance of Neandertal morphology beginning around 450,000-430,000 probably from *H. heidelbergensis*.

According to this hypothesis, the hominin lineage eventually leading to Neandertals became isolated in Europe due to the severe climatic conditions of the Pleistocene. The geographic range of Middle Pleistocene European hominins would have been restricted and isolated during glacial maxima by ice sheets covering northwestern Europe and by their associated permafrost zones, but then extended again during interglacial periods.

It was in these conditions of isolation and expansions, that Neandertal morphology is thought to have appeared gradually, maybe through natural selection as an adaptation to cold climate conditions (Weaver 2009), but also most likely through genetic drift, (J. Hublin 1998).

Moreover, recent findings from Arsuaga et al 2014 sums additional evidence supporting such scenario. Briefly, the authors describe how fossils belonging to 27 individuals from the archeological site of "Sima de los Huesos" in Atapuerca, Spain (dated ca 430,000 years) - previously classified as *H. heidelbergensis* -, presented an almost full set of Neandertal derived traits in their face and teeth, whereas the braincase still retained "primitive" attributes. Interesting, one does not find Neandertal braincase shapes in Europe before ~200,000 years ago (Arsuaga et al. 2014; J.-J. Hublin 2014). Therefore, these observations suggest that Neandertal features did not evolve as a block but rather were fixed at different rates and paces in different parts of the anatomy (J.-J. Hublin 2014). In addition, the time period of the Sima de los Huesos population coincides with the onset of a long interglacial period that followed a severe glacial episode, which would have permitted hominins with such morphological features to disperse again into higher latitudes.

Furthermore, after this time, extreme climatic conditions prevailed roughly every 100,000 years within the Europe, occasioning that human occupations remained scarce and discontinuous in northern latitudes like modern-day Germany and the British Isles; most likely

involving periodic demographic crashes during subsequent climatic cycles.

Moreover, it is interesting to note, that given the fragmentary nature of most hominin fossil remains and the morphological variability of several cranial traits, for some skeletal features Neandertal and modern human ranges of variation seem to overlap. However, when multiple features are considered altogether, most specialists agree that Neandertal and modern human crania can be distinguished morphologically from one another (Weaver, Roseman, and Stringer 2007; Weaver 2009).

Therefore, morphologically a classical Neandertal specimen can be defined by the co-occurrence of certain traits like: a longer and lower skull with a large endocranial capacity (ranging from 1,245 to 1,900 cc and averaging about 1,520), a continuous strong brow ridge, minimal chin development, a suprainiac fossa throughout ontogeny, a wide pelvis with a long thin superior pubic ramus, short distal limbs, wide thoracic chest capacity, a relatively short stature, mid-face projection (prognathous), and from the analysis of some complete post-cranial remains, it can be inferred that they had a reduced spinal lordotic curve (hypolordosis) (Hawks 2012). On the other hand, the morphologic features of anatomically modern humans show more globular brain overhanging with an endocranial capacity range between 1200-1850 cc and a more retracted face. A more complete description of morphological features between Neandertals and modern humans is found in **Figure 4**.

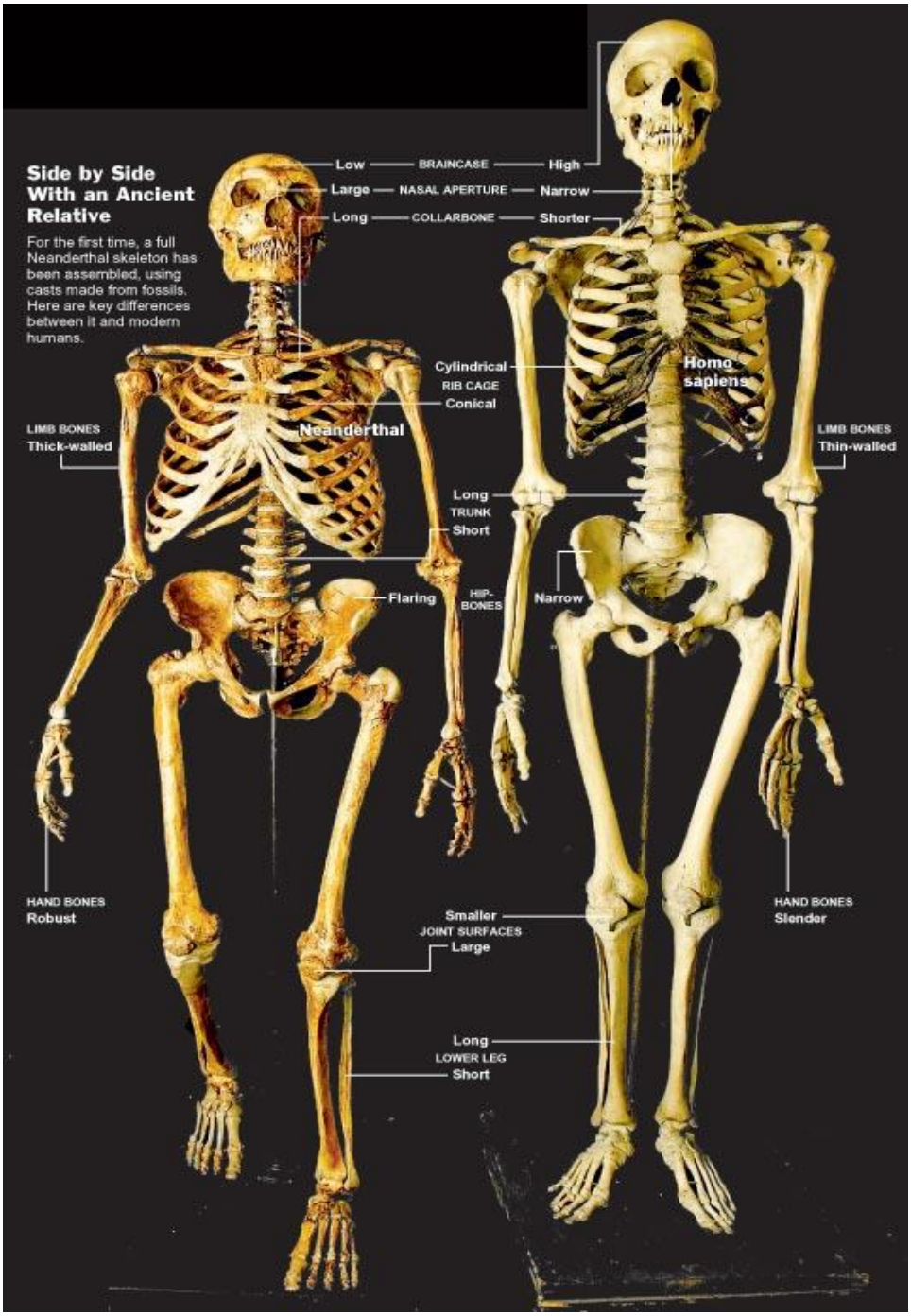


Figure 4: Morphological differences between Neandertals and Modern humans.
[Adapted from diogeneschilds.wordpress.com]

Moreover, not only their origin is controversial but also the evolution of their morphological traits has been subject of great debate. For instance, (Howell FC 1957) considered that given their cranial and post-cranial – non-cranial skeletal remains morphologic variation, there seemed to have been three main morphological varieties throughout their evolutionary history:

1) Early Neandertals. According to (Howell FC 1957), the postcranial skeleton of this early Neandertals samples seemed to be "more anatomically modern" than that of later Neandertals. This group included the Neandertals from Krapina (Croatia), Saccopastore (Italy), and Ehringsdorf (Germany), representing times from last interglacial (now dated at approximately 120,000 years ago) or earlier.

2) Classic Neandertals. This group of Neandertals is defined by the co-occurrence of several morphology traits (as mentioned before) that clearly differentiates morphologically modern humans from them. This group includes most of the well-known remains from Europe, as old as 100,000 years and as recent as 40,000 years mainly from Europe.

3) Southwest Asian Neandertals. This group included the entire known fossil record of the Levant, including sites like Skhul, Tabun, Zuttiyeh, and Qafzeh and Shanidar. It is interesting to note that given the morphological variability of Southwest Asian Neandertals, some specimens like the Skhul cranium seemed to be more similar to modern humans than to classic Neandertals; suggesting a "sapiensization" of Southwest Neandertals specially seen in late specimens of this geographical area. Some have even argued that some of this specimens like Skhul and Tabun represent a separate population anatomically intermediate between classic Neandertals and modern humans (McCown and Keith 1939; Hawks 2012). As further reviewed in insight from this morphological variety are at least interesting in light of what is currently known about genetic admixture with modern humans (Green et al. 2010).

On the other hand, (Rosas et al. 2006) consider that there are four broad stages of Neandertal evolution:

- Stage 1 includes ‘‘early pre Neandertals”, the Middle Pleistocene archaic specimens. These hominins are considered to show incipient Neandertal features mainly in the facial region.
- Stage 2 specimens termed ‘‘pre-Neandertals’’. They are thought to exhibit Neandertal morphology more clearly, showing Neandertal features also in the occipital area.
- Stage 3 ‘‘early Neandertal’’ specimens show most Neandertal traits in the posterior cranium and some also in the temporal region.
- Stage 4 comprises the ‘‘classic Neandertals’’, showing fully expressed Neandertal morphology.

Thus, as a summary, it seems that while the oldest specimens exhibited Neandertal distinctive traits only to a slight degree (and some of them more modern-human like features), traits typically associated to Neandertals became more accentuated over time up to the end of Neandertals existence (Hawks 2012). Notwithstanding, a word a caution should be emitted regarding the different Neandertal morphological varieties, given that although there seemed to be clear differences between early and late members of the Neandertal lineage, some opinions have question the veracity and unity of European and Asian varieties of this hominin group (Weaver 2009; Hawks 2012). It will be of interest to address how Neandertals from these different “varieties” relate to each other genetically, and to which extent they all fall within the genetic variation of Neandertals or if maybe they do represent real Neandertal-modern human hybrids.

Furthermore, either by archeological or morphological analysis of their remains, there is evidence that Neandertals lived across Eurasia, from the extreme edges of the Altai Mountains to the far west of the Iberian

Peninsula, where the most recent Neandertal archeological dates to circa 40,000 years ago or even later, are thought to have lived. Moreover, Neandertal remains have also been found in sites as south as the Tabul and Skhul caves in the Levant; therefore their geographical range also extended throughout all West Asia and Middle East regions (**Figure 5**). Additionally, there is controversial evidence from Mousterian lithic industry, which suggests that they could have also potentially lived as north as the North pole circle (Slimak et al. 2014).
The fact that their presence has been recorded across all Eurasia for ~400,000 years overlapping with several major environmental changes, in addition to that similar body proportions have been observed in other cold adapted mammals enabling preservation of body heat in cold environments, have motivated some researchers to hypothesized that maybe the Neandertal morphology reflected a 'hyper-arctic' adaptation (Holliday 1997). For instance, it has been suggested that the Neandertals' midface projection, including inflated cheeck bones and a broad nose, could have been adaptations to cold climate. However, this view has recently been challenged, as the large nose sinuses characteristic of Neandertals were found to be atypical for cold-adapted mammals (Rae, Koppe, and Stringer 2011).

Figure 5: Geographical range of Neandertals.

[Adapted from http://en.wikipedia.org/wiki/Neandertal].

On the other hand, the gradual appearance of derived Neandertal traits, their probable demographic changes due to harsh environmental condition fluctuations and their putative rapid differentiation from their most recent common ancestor, hints towards a rapid morphological distinction driven probably by drift (Arsuaga et al. 2014). For instance, while the skulls of a modern human (Cro-Magnon 1, far left) and a Neandertal (La Ferrassie 1) show a morphological divergence between the two hominin species 500,000 to 600,000 years after their ancestral populations split, in contrast, after about 2 million years of separation, common chimpanzees and bonobos still display very similar morphologies (likely due to strong stabilizing selection). Therefore, like polar bears and brown bears (Liu et al. 2014), the extent of morphological divergence of Neandertals from the shared common ancestor with modern humans, seems to have occurred over a very short period of time.

Regardless of the true explanation, the fact that across their geographical range and gradually through time Neandertals exhibited such distinctive morphological features implies that at least some of these traits should have a genetic basis.

Obtaining DNA and addressing the patterns of genetic variation from contemporaneous Neandertal samples, as well from fossils from different geographic and temporal locations, will be paramount in order to understand the evolutionary force(s) that have shaped their distinctive phenotypical appearance.

1.2.2 Neandertal Technology, cognitive abilities and behavior

For most of their history, Neandertals produced a Middle Paleolithic technology known as Mousterian industry, named after the lithic material – stones with different inorganic composition use to construct tools- found at the site of Le Moustier in the Dordogne (France), where they were initially described. Typical Mousterian industries (**Figure 6**) consisted in the use of certain flaking techniques from locally available lithic material for the production of flakes that could be converted to a

wide range of shapes, including various kinds of side scrapers, retouched points, denticulates, notches, and sometimes small hand axes (P Mellars 1996; Shea, J.S., Brooks 2000).
However, it is of importance to note that this technology was not restricted to Neandertals, as similar stone tools were likely also produced by early modern humans outside of continental Europe like in in the Caucasus, the Near East (where it is associated with both Neandertals and early modern humans) and North Africa (where there is no fossil record that Neandertals ever existed) (Stringer and Gamble

Figure 6: Typical Mousterian industries from Beuzeville, Eure (France).
[Adapted from http://en.wikipedia.org/wiki/Mousterian]

1994; RG 1999; Lalueza-Fox and Gilbert 2011).

Thus, the recent finding of Mousterian tools near the Arctic Circle (Slimak et al. 2014) does not automatically point out a Neandertal presence in the region. Mousterian technology remained largely unchanged for most of the Neandertal period. However, shortly after the arrival of anatomically modern humans in Europe about ca 45,000 years ago and before their final extinction (Stringer and Gamble 1994), some Neandertals began to produce so-called transitional industries

which showed Upper as well as some Middle Paleolithic characteristics. These industries include among others the Chàtelperronian in France, the Szeletian in Central Europe, and the Uluzzian in Italy and have been traditionally interpreted as a sign of the acculturation of the last surviving Neandertal populations by anatomically modern humans (Stringer and Gamble 1994). On the other hand, the perception that Neandertals were merely copying Upper Paleolithic tools may be too simplistic, and may instead have been the consequence of an independent development by Neandertals (Zilhão et al. 2006). Alternatively, some of these industries may have been produced by anatomically modern humans and previously incorrectly associated to Neandertals (Hoffecker 2009), as shown by a recent analysis which indicates that some teeth found in Uluzzian levels stem in fact from anatomically modern humans (Benazzi et al. 2011; Lalueza-Fox and Gilbert 2011).

1.2.3 Cognitive behavior and symbolism

Another contentious and interest subject about Neandertals concerns the cognitive abilities. Although once thought to be fundamentally inferior to that of anatomically modern humans (including a lack of language), recent findings suggest that well before the arrival of modern humans, Neandertals have used ornaments to decorate their bodies: for instances with pigment-stained marine shells at Cueva de los Aviones (Spain) (**Figure 7**) (Zilhão et al. 2010).

Figure 7: A perforated upper half-valve of Pecten *maximus* from Cueva de los Aviones. The internal, naturally red side (Left) and the external, whitish side that was painted with an orange colorant made of goethite and hematite (Right) [Adapted from (Zilhão et al. 2010)].

This behavior implies some sort of modern symbolic comportment and support the view that they could have had comparable cognitive abilities to anatomically modern humans (Zilhão et al. 2010; Peresani et al. 2011). However, the attribution of some of these personal ornaments to Neandertals and the claim of complex symbolic behavior among them, has been challenged by those who argue that for most of their evolutionary history such evidence is lacking and that they only seemed to appear when anatomically modern humans start migrating into Europe (Paul Mellars 2010). Moreover, it has been shown that Neandertals and anatomically modern humans may have had different phases of brain development after birth, a characteristic that potentially could underlie cognitive differences between both groups (Gunz et al. 2010; Lalueza-Fox and Gilbert 2011)

Moreover, meanwhile anthropological and archeological findings are important to better understand the environment, morphology, and, to some extent, behavior of Neandertals, genetic data is indispensable in

order to elucidate complete inferences about their evolutionary history. For example, while the number of archeological sites and the amount lithic technology and type could be used as a proxy to estimate their population sizes, accurate dating of fossil remains has proved to be an issue before (Higham 2011). In addition, a lack of archeological remains cannot be immediately translated into an absence of hominin inhabitation. Finally, cultural transmission and geographical and temporal overlapping with modern human groups complicates a specific identification, based on hominin lithic industry.

In the same direction, but from an anthropological characterization of the samples, even though remains from different time periods can be used to study the evolution of their morphological features, the scarcity and fragmentary nature of their fossil remains, as well as the unknown heritability of morphological traits, prevents us from using exclusively morphological data to make accurate inferences from their evolutionary history. Thus, without genetic data, understanding the nature of their relatedness to modern humans and other archaic hominins, elucidating the defining evolutionary forces, as well as, the genetic bases behind their particular morphological appearance; in addition, to further insights into their demography history and the causes of their eventual demise, could never be addressed.

1.3 How to use genetic data for evolutionary inferences?

1.3.1 Genetic data as a key factor to understand evolution

The certitude that we can use genetic data as record of past evolutive or demographic events relies on this fact: all organisms share a common ancestor with each other in the past, their morphology and biology have

a genetic basis, and their genetic information is inherited from the genomes of its progenitors, generation after generation.

Moreover, in order for an organism to evolve i.e. into having a certain phenotypical appearance or resistance to a given pathogen, two necessary and separate steps need to occur. First, genetic variation needs to be generated largely by mutation and recombination; however, admixture (with close related populations/species) can add new genetic variation to a population (see below). The second necessary step is the action of adaptive and demographic processes which would shape the genetic makeup of the population and cause certain genetic variants to change their frequencies or become fixed.

Therefore, by comparing the genetic data from two species or from individuals within a population, one can address different questions from estimating when did the genetic basis of a certain feature aroused, to unravelling if a population has recently experience any adaptive events. It is the shared evolutionary heritage and close relatedness between to organisms, what allows making accurate inferences about their common and private evolutionary histories.

However, before one can address how the genomes of organisms can serve as record to make inferences about their evolutionary and demographic histories, one has to understand the nature of the genetic material, how it is inherited, the sources of genetic variation and which evolutionary forces might cause that while some changes will become common to all individuals others will be deleted from the population.

1.3.2 Biological sources of genetic variation and its functional consequences

The genetic material or genome of an organism is encoded in its deoxyribonucleic acid (DNA), which is usually distributed in to smaller

units called chromosomes and which ultimately is constituted mainly from four monomeric subunits called, *nucleotides* (NT), which at the same time vary depending on its chemical structure mainly into: adenines (A), guanines (G), cytosines (C) or thymines (T).

Therefore, it is the sequence of these four different nucleotides throughout the genome, what carries the necessary information in order to create an organism and dictate what it's morphological, physiological, and biological features will be.

Even though this thesis is about Neandertal, given that modern humans are its most closely related hominin and that it's impossible to recover a complete Neandertal genome (including repetitive sequences and centromeres), a brief description about the genetic material from modern humans and its sources of genetic variation is in order.

For instance, the human genome consists of about 3 billion nucleotides, and in every generation, several of these nucleotides are affected by mutations in the male and female germ-line (during meiosis) so that subsequent generations will inherit slightly different versions of the ancestral genomes. However, also different bits of DNA ranging from few to many thousands of bases (Kb) can get lost (deletions), be added (insertions) or copied several times, and even turned around (inversions) throughout this same biological process, creating even more genetic variation to be inherited (Andrés Moreno-Estrada 2009).

Recent modern human DNA sequencing projects (Lander et al. 2001; The International HapMap Consortium 2003) have provided some insights into the genetic variation of healthy individuals and have largely proposed to catalogued variants according to their corresponding number of base pairs (Bp), mainly into three main categories (Figure 1.3):

1. structural variants (SVs) encompassing from few kilobases (> 10Kb) to few megabase (> 1 million bps), including large deletions and insertions, inversions, macrosatellites and Copy Number Variants (CNVs), that change the number of copies of a gene or any piece of DNA.
2. Structural variants encompassing few hundreds of base pairs such as medium sized insertions and deletions and minisatellites (repeats of 10-100 bps).
3. Variants of few bps such as small insertions and deletions, microsatellites (repeats of 2-6 bps) and single base pair substitutions, called Single Nucleotide Variants (SNVs).

From all types of genetic variation, SNVs are the most common and the most studied ones, and since in the present thesis most analysis have been performed using them, a deeper description is in order.
Unofficially and incorrectly SNVs are sometimes also referrer as Single Nucleotide Polymorphism (SNP). However, this term is reserved for SNVs for which the frequency of an observed *allele* (type of nucleotide) has a certain arbitrary minimum (typically 1%) and thus is *polymorphic* in the population. For instance, some DNA molecules in a population may have an **A** at a particular site, whereas other individuals within the same population may have a **C** at the same site. Most SNPs are biallelic, that is two alleles or forms for the polymorphism exist; the most frequent is usually referred to as "major allele", while the other as "minor allele". A further definition worth mentioning is *genotype*, which is the specific combination of two alleles (one coming from the father and the other for the mother) for each genetic locus.

Furthermore, even though most SNVs or SNPs occurring in the DNA have no apparent known function, some variants arise in coding regions of genes - altering amino acid sequences of proteins - or in regulatory regions that affect gene expression, and thus might be biological important. Coding SNPs that result in amino acid replacements are

known as nonsynonymous SNPs, while those that are also present in coding regions but do not result in an amino acid replacement are known as synonymous SNPs. Moreover, even though the genetic contribution to phenotypes is usually not easily established, nonsynonymous SNPs as well as functional SNPs in regulatory regions are commonly believed to have a strong direct effect on phenotype.
As mentioned before, SNPs represent by far the majority of the genetic variants, they are so common that estimates initially pointed that all present-day humans are ~99.6-99.8% identical at the nucleotide sequence level (Przeworski, Hudson, and Di Rienzo 2000; Reich et al. 2002). Therefore, the genetic variation contained in that remaining 0.2–0.4% in addition to the expected differences in structural variation and CNVs, account for our individual uniqueness at the DNA level.

Estimates of human genetic diversity state that approximately 1 of each 1000 nucleotides will be different between two non-related individuals randomly selected (Reich et al. 2002). These genetic differences exist at the same time between modern human populations. Contrary to the what would be expected based on the phenotypical variability observed between modern human populations, several studies (Excoffier, Smouse, and Quattro 1992; Barbujani et al. 1997; Romualdi et al. 2002; Jorde et al. 2000) have point out that when the human individuals are hierarchically clustered by populations and continents, approximately 80% of the variance of the model is explained because of differences between individuals of the same population rather than their geographical origin. Moreover, only 5% to 10% is explained because of differences between populations of the same continent and a 10% to 15% of the variance is explained by genetic differences between continents.

Moreover, given the known phenotypical differences between present-day humans, it is quite interesting that we are so genetically alike; a fact that is most likely related to the recent origin of our evolutionary

lineage. During the last almost 100 years, scientists have been devoted to understand which could be the explanation for the patterns of genetic variation observed in modern human populations and to which extent adaptive and demographic inferences can be interpret from them. In order for evolution to occur there must be a change in allele frequencies in time. If this is not the case, a population is said to be in Hardy-Weinberg equilibrium, which means that the allele frequencies in one generation can predict the genotype proportions in the next generation, as a result of the absence of processes shaping tis diversity. In nature, however, populations are under the operation of one or more evolutionary forces, mainly: mutation, natural selection, genetic drift and migration; which are capable of shaping genetic variation. A review of each of these evolutive forces is of use in this case, in order to understand how they modify the distribution of genetic diversity and to disentangle the effect of one from the other, so demographic and adaptive events can accurately be reconstructed for the evolutionary history of any organism, by the means of genetic data.

1.3.3 Mutation and recombination

Mutation is defined as a structural change in the DNA molecule. It is the ultimate source of genetic variation and thus, without it evolution could not take place (Crow, 1997). Mutations range from single base changes to small insertions, duplications, and deletions up to chromosomal changes, such as translocations and the formation of polyploids. Base substitutions, or point mutations, occur when there is a change from one base pair to another at a single position in the DNA sequence. Changes from one purine (A or G) to another or one pyrimidine (C or T) to another are referred to as *transitions*, whereas changes from a purine to a pyrimidine or, vice versa, are called *transversions*. Moreover, the molecular mechanisms by which they are generate, go from chemical changes (for instance, cytosine deamination), physical mechanisms (for instance, breaking of the

double helix for generating an insertion) to enzymatic actions (for instance, slippage of the DNA polymerase).
Mutations can be generated in whatever cell type that form an organism, but only those that appear in the germ line could have the opportunity of passing on to the next generation. Mutations that occur in the rest cell types (they are called somatic mutations), although could be associated to particular phenotypes or disease will not have direct evolutive consequences since they will not pass to the next generation. Each type of mutation has different probabilities of occurring, depending on the type of mutation, if there is a mechanism to correct it or the context of the genome where it happens. For instance, while for the nuclear genome a mutation rate of $\sim 1.2 \times 10^{-8}$ bp^{-1} $generation^{-1}$ is usually utilized (although the subject is still contentious an a rate of $\sim 0.5 \times 10^{-8}$ bp^{-1} has been recently proposed), the non-coding region of the mitochondrial genome the mutation rate is $\sim 1.6 - 2 \times 10^{-7}$ bp^{-1} per year.

Additionally, given that modern humans reproduce sexually, new allele created by mutation will be reshuffled by recombination, creating even more genetic variation on which selection and other evolutionary forces could act. During this process of recombination, which occurs during meiosis, some alleles will sometimes be carried more frequently together on the same chunk of chromosomes than others, creating stretches of genetic variants inherit together, are called *haplotypes*. Thus, two loci are in linkage disequilibrium (LD) if there is a specific combination of their alleles that are observed on the same haplotype more often than expected at random in function of the distance separating both loci. Knowing the haplotypes provides valuable information about ancestry and inheritance to perform evolutionary studies.

1.3.4 Genetic Drift

Once mutations arise, different outcomes are possible. They can be lost, maintained, or become fixed (when all chromosomes carry the same allele). The expected fate of mutations in natural populations is partly determined by randomness. Because each generation represents a finite sample from the previous one, chance alone can change allele frequency between generations solely through the stochastic process of sampling. This evolutionary process is known as *random genetic drift*. The magnitude of the genetic drift is inversely proportional to the number of chromosomes that pass to the next generation. This latter concept is related to the *effective population size of a population* **(Ne)**, which is the total number of individuals with which one could reconstruct all the genetic diversity of a population. Therefore, alleles with the same initial frequency (e.g. 0.5) will become either fixed or lost much more rapidly in small populations (i.e. Ne = 25) than in larger ones (i.e. Ne = 2,500), where they will persist over generations with more or less subtle fluctuations in frequency depending on population size (Andrés Moreno-Estrada 2009).

Additionally, genetic drift affects the whole variability in our genome and not a locus in specific. Moreover, there are two demographic processes where genetic drift has an important role in shaping the genetic diversity; a reduction of genetic diversity due to detrimental of population sizes in the past (see **Figure 8**). There processes are bottlenecks and founder events. **Bottlenecks**, occurs when there is a drastic reduction in the number of individuals from an original population due to environmental, geographical or maybe even social conditions. This demographic event has two main consequences: first, the descendants of the individuals which survive the population reduction could carry allelic frequencies that are not representative of the original population, and second, there are big chances that the allelic frequencies could fluctuate until the population grows back again in size.

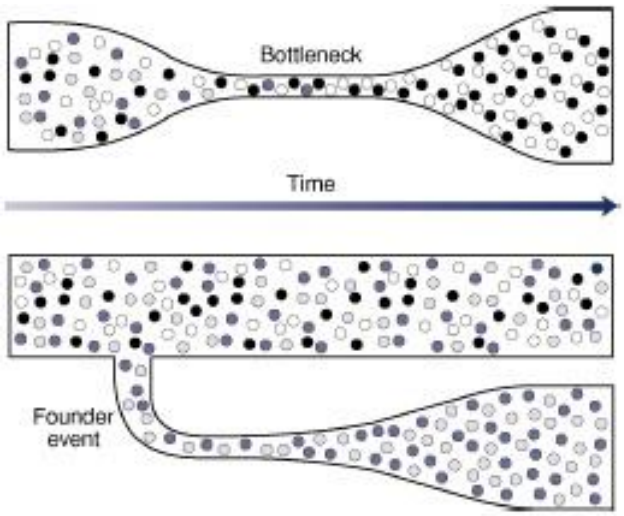


Figure 8: Bottlenecks and founder events.
Circles of different colors represent different alleles. Both events result in a loss of allelic diversity due to reduced population sizes in the past [Adapted from (Jobling, Hurles, and Tyler-Smith 2004)].

Consequently, common alleles in the original population could have a small frequency in the new population or even could disappear. On the contrary, rare alleles in the original population could increase their frequency in the new population, even if they have deleterious effects in the individuals that carry them. As we will see in the next section, this observation is due to the fact that purifying selection (the evolutionary force that removes deleterious variants) is more efficient at removing damaging variants in populations with large Ne, since (and assuming that there is random mating) individuals that do not carry those variants will have a better fitness and will pass on their genes, than individuals in small populations where the probability that an individual contributes genetically to the next generation its more based on change. Finally, **Founder events** are produced when a small set of individuals from a bigger population is separated towards the colonization of new regions; this process has roughly the same demographic consequences than a bottleneck.

1.3.5 Migration

Most great eukaryotic organism populations are spread over large areas and sometimes there divided into smaller subpopulations within which individuals tend to mate as they share the same local habitat. If such a separation continues for a given amount of time it will create a *population structure*, where there is a genetic differentiation among the subpopulations of a population and thus there are allele frequencies differences between them. However, subpopulations rarely stay completely isolated. Therefore, if there is movement of individuals from one place to another (i.e., migration), and they leave descendants, *gene flow* is said to occur.

A continuous gene flow between populations will tend to homogenize the amount of genetic diversity present in both populations, resulting in a global population or meta-population. Thus, the fingerprint of the migration will be only observed when the allelic frequencies differ between both populations, and the homogenization process has not yet finished. In the presence of random genetic drift, the homogenization by gene flow is reduced.

Another important effect of gene flow is that it can create linkage disequilibrium by the mixing of populations. In a mixed population, there will be an excess of haplotypes considered as characteristic of the various parental populations, and thus there will be an increase of LD. Moreover, the spatial distribution of modern humans, as for well for the other kinds hominins that once roamed the earth, has been historically structured in subpopulations with different barriers to movement and with varying densities and environmental conditions. Such demographic pattern stirs up interesting questions about the capacity of populations to adapt to local environments or how do favorable alleles spread over populated areas. Finally, recent studies have revealed that genetic introgression events occurred between archaic humans and anatomically modern humans in the past (Green et al. 2010). Such

findings have showed us not only how gene flow between two populations - previously conceived as different species - modifies the genetic variation of a population, but also how the sequences genetically introgressed into modern humans could have served as an extra source of variation towards adaptive events.

1.3.6 Natural selection

Natural selection is the only process that leads to adaptation. The way it shapes diversity, is actually by its manipulation. When inherited variants cause the organisms to differ in their ability to survive and reproduce (i.e., fitness), there is a generational change in which those individuals with the fittest variants will tend to leave more offspring and, consequently, those variants will tend to increase in frequency. In this way, a population can become progressively better adapted to a given environment.

Models of selection compare the relative fitness of a genotype with that of other genotypes. Mutations that reduce the fitness of the carrier are subject to *negative selection*, also known as *purifying selection;* since they tend to remove variants from the population (see **Figure 9**). Alternatively, mutations that increase fitness undergo *positive selection*, and therefore tend to increase rapidly in frequency due to a mutation that gives them some advantage in the adaptation to the environment. Because both processes shift the overall makeup of the population, either by favoring one allele over another, or acting against unfavorable deleterious mutations, they are usually associated with a model of *directional selection.*

Alternative models consider the interaction between alleles to determine the impact of mutations on the fitness of the genotypes. Thus, scenarios like *overdominance*, *underdominance*, and *codominance*, refer to when the advantage is conferred to the heterozygote, the two homozygotes or one homozygote, respectively. When selection acts on quantitative traits, overdominance can be understood as selection that favors intermediate trait values; this acts to reduce variation and thus is called *stabilizing selection*. As opposite to this, underdominance would be related to *diversifying selection*, which favors extreme phenotypes (i.e. traits in the tails of the distribution) and

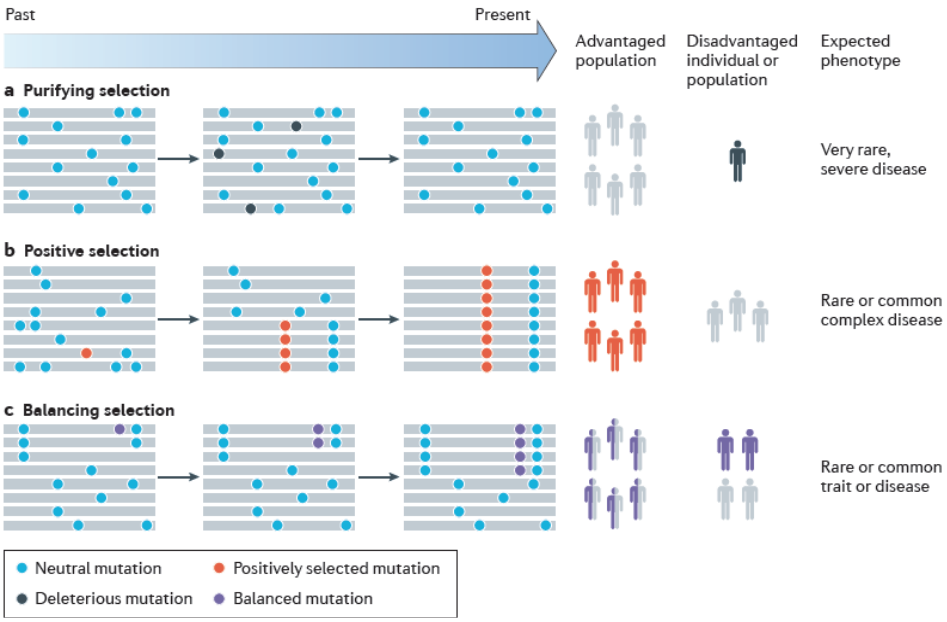


Figure 9: The different modes of natural selection.
[Adapted from (Quintana-Murci and Clark 2013)]

consequently tends to increase variation.

In particular, overdominance creates a balanced polymorphism, although this is not the only mechanism by which balanced polymorphisms can be generated. Alternatively, *Frequency-dependent selection* whereby the frequency of a genotype determines its fitness (i.e. rare alleles are favored), fulfil the same action. Both models then

belong to what is referred to as *balancing selection*, which maintains polymorphism in the population (see **Figure 9**).

The fate of selected genes depends to a great extent on the genomic context. When mutations are undergoing selective fixation, they tend to drag flanking variation with them through a process of *genetic hitchhiking* (Smith and Haigh 1974). The hitchhiking effect occurs for a simple reason: if two alleles at different loci are in linkage disequilibrium, then directional selection for one of them will cause either of them to increase in frequency or be eliminated. This results in a reduction in the genetic variation for a region surrounding the selected target. This phenomenon, known as *selective sweep* when positive selection is taking place or as *background selection* when negative selection is occurring, leaves several characteristic molecular signatures whose recognition is essential for detecting selective events in the genome.

In order to contextualize selection even more, two additional considerations are worth to mention. First, selection acts on the phenotype, not on the genotype, and a phenotype is determined by many genes that interact with each other as well as with numerous environmental factors. Second, that selection does not act alone; there is a complex interplay among the different evolutionary forces, so while mutation, recombination and migration tend to increase genetic diversity, random genetic drift decreases it, and selection can do either of them.

1.3.7 How to make evolutionary inferences from genetic data?

Once we have learned that genetic diversity data contains information about the sizes and movements of past populations, or on adaptive events to changing environments, we can start to analyze the patters of genetic diversity at both population and species level in order to elucidate which of the distinctive evolutionary forces (or a combination of them) could have caused the observed patterns.

Moreover, there are two main approaches to address genetic diversity from a population: one is to give a description of the distribution of genetic diversity and the other is to make inferences about how or when the genetic variation was generated. While the former methods allow comparisons between species or between populations within a species, the latter require explicit or implicit models of the evolutionary processes, in order to reconstruct the evolutionary past of an organism.

Furthermore, descriptions of present diversity almost inevitably lead to discussions of how it might have arisen. It is worth noting that there is often no simple and unique answer to how could a particular pattern of genetic diversity arose, and thus a combination of several analytical approaches is advisable.

In addition, paleogenetic and genomic studies now provide a way in which one can calibrate or improve such inference models in accordance to the genetic data extracted from fossils. Finally, not regarding what type of methods is employed, it is worth noting that a single locus inevitably contains less information on our evolutionary past than many loci do, no matter how informative that individual locus is (i.e. the mitochondrial DNA). A single locus gives a single account of the evolutionary process, and given that the history of any single may be biased (whether by drift or by selection), genome-wide analysis are advised.

1.3.7.1 Summarizing genetic variation

Summary statistics that resume the amount of variation do not encapsulate all information present in the data, but allow comparisons between populations and between loci. Heterozygosity is the most commonly used approach to measure genetic diversity. It can be calculated in several ways, for instance, from the number of differences between two sequences within a certain genomic regions, or by Nei´s

genetic diversity estimate which measures the probability that two alleles drawn at random from the population will be different from each other.

Furthermore, genetic diversity can also be summarized by estimating the *Number of segregating sites* or the *Nucleotide diversity* π, which describes the probability that two copies of the same nucleotide drawn at random from a set of sequences would be different from one another.

Moreover, θ (known as the neutral parameter estimator or population diversity estimator) denotes the amount of variation expected at each nucleotide site under neutral evolution. Thus, if selection is absent, and if each mutation is thought to occur at all possible sites (and, of course, if the other assumptions are met), the expected values of π and θ should be equal.

In addition, there are several different methods for estimating θ from sequence data. These methods use different parameters derived from the observed diversity, including: the number of alleles, the number of segregating sites (S), the number of **singletons** (η), the observed homozygosity (F) and the mean number of **pairwise** differences (π)

In an ideal neutrally evolving population these different estimators of θ will have the same value. If the different estimators of θ give significantly different values for a single locus, we can infer that the population departs from the neutral model, for example because of a different demography or because selection is acting. Such comparisons form the basis of several methods used for detecting selection, which measure how different the various estimates for θ actually are. For example, utilizing the *Tajima´s D statistical test,* one can compare the amount segregating sites vs the nucleotide diversity in the sample to decide which scenario it's the most likely to have occurred. For instance, while an excess of segregating sites and low nucleotide diversity can be indicative of a positive selection event or a recent

population growth, the inverse will suggest instead, that balancing selection could be taking placed or the population was that the population is structured.

Finally, comparisons of the statistics estimating θ for a single locus and the rest of the genome can help distinguish between selection occurring at the locus in question, and demographic events affecting diversity across the genome.

Notwithstanding, the difference rate (mismatch distribution) between pairs of sequences can also be used to represent genetic diversity. In this case the distribution of the number of such differences between each allele and every other allele summarizes the discernible genetic diversity, and it is the shape of the distribution what is indicative of population history, in particular being influenced by episodes of population expansion. For instance, a smooth, bell-shaped mismatch distribution indicates a period of rapid population growth from a single haplotype, whereas a ragged, multimodal distribution indicates a different situation, i.e. a population whose size has been constant over a long period

As mention before, all these methods (as well as others) are estimators to summarize genetic variation and vary depending on the hypothesis behind the most likely scenario to have created such patterns of diversity.

Moreover, genetic data can also be employed to measure genetic distance, which allow us to compare the relatedness of populations or individuals; the greater the evolutionary distance between them, the greater the numerical value of the statistic.

There are several ways to measure genetic distances between populations, which are largely used depending on the type of data and different expectations about the underlying evolutionary processes.

Two commonly used classical measures of genetic distance are *F*st and Nei´s standard genetic distance, *D*. While Fst (also known as fixation indices) measures the proportion of the total variance in allele frequencies that occurs between subpopulations and it ranges from 0 to 1, *D* varies between 0 and infinity, and relates to the probability of drawing two identical alleles from the two different populations (which is equal to the probability of drawing identical alleles from the same population.

We can also generate a matrix of distances between all sampled individuals for all for all genotypes where each number in the matrix represents the degree of similarity between two individuals. Later on this covariance matrix can be represented visually by a principal component analysis (PCA), which will allow us to infer the evolutionary relationship of individuals or populations. Cluster analysis can also be utilized to this purpose generating a matrix where each individual is given a value or membership coefficient for every cluster, which reflects the probability of that individual belonging to that cluster; after that they relatedness can be display graphically as well. STRUCTURE and more recently ADMIXTURE are softwares that have been widely used for that purpose (Jobling et al. 2014).

Furthermore, the distance between individual genomes can be calculated by counting the number of differences between the two genomes directly, or by counting the number of differences when compared with a reference genome. This can then be corrected using a specific model of sequence evolution. An alternative is the statistic Patterson's *D*, which was devised as a test statistic to analyze the relationship of ancient genomes to modern human genomes, but can be used for any four-way genome comparison. The format of the statistic is *D* (ref1, ref2, test, outgroup) where we want to compare the relative closeness of the test genome to the ref1 and ref2 genomes, given a known outgroup genome. If *D* is zero, then the derived alleles in the

test sequence, for example in the Neandertal genome, match alleles in the two human reference genomes equally often. If *D* is positive then the derived alleles in the Neandertal match alleles in ref2 more often than ref1, and if *D* is negative then the derived alleles in the Neandertal match alleles in ref1 more often the ref2. Moreover, given that the phylogeny between the individuals is assumed to be correct, a scenario where D is significantly different from zero, could be explained as a genetic flow occurring between the ancestors of the test (i.e. Neandertals) and one of the reference individuals (r1 or r2, depending on the sign of D). This approach is also known as the ABBA/BABA approach, referring to the ancestral (A) or derived (B) alleles of the four genomes (**Figure 10**).

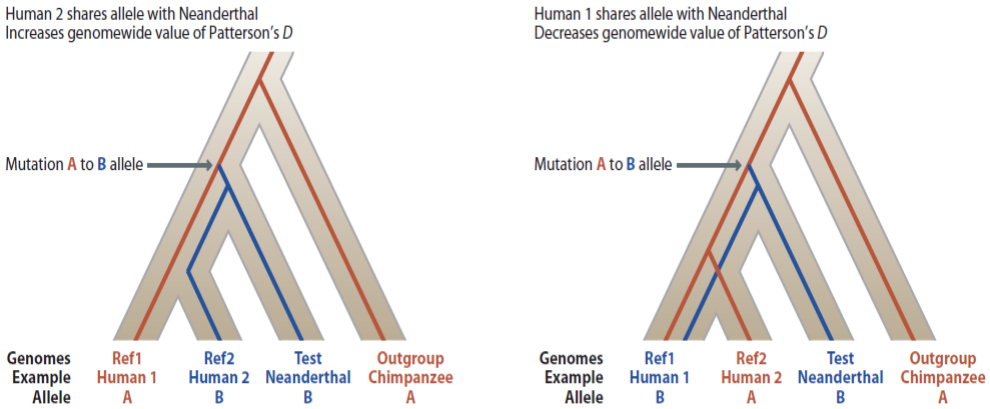


Figure 10: Measuring the distance between individual genomes by the ABBA/BABA approach [Adapted from (Jobling et al. 2014)**]**

However, as mentioned in **section 2.5**, population splits are believed not to occurred punctually, and thus if past complex demographic scenarios between the ancestors tested individuals occurred (i.e. ancient population structure), the assumptions on the historic relatedness between individuals might not be correct (Eriksson and Manica 2012).

Additionally, under the same assumption that a known population phylogeny is correct, if instead of comparing allelic states one uses allele frequencies, one could test if the allele frequency data reconstructs the documented history between the populations or rather it suggest the occurrence of past admixture events. This statistical demographic approach is known as *f4* test; one of other *Fst*-based statistics, designed to inferred the relatedness between populations (Patterson et al. 2012). Briefly, the *f4* test is based on an unrooted population topology describing the relationship between four populations A, B, X and Y, such as (A, B), (X, Y). If the population topology is correct and populations have remained largely isolated after their initial split, then any allele frequency differences between each pair of lineages must have arisen solely after their separation, and thus the overall difference in allele frequencies ((A-B)*(X-Y))) between each pair of populations across all loci should be zero. However, if there is an asymmetry in allele frequency covariance that is inconsistent with the proposed topology, the most likely explanation is that gene flow occurred and the f4 statistic ((A-B)*(X-Y))) will be significantly different from zero. Moreover, several research groups have used the f4 test (and several derivations from it), in order to assess if given a population phylogeny, differences in alleles frequencies between present-day human populations, can be explained by past archaic genetic introgression events (Durand et al. 2011; Reich et al. 2011; Patterson et al. 2012).

1.3.7.2 Evolutionary inference methods for genetic data

As mentioned before, evolutionary inference methods require an explicit or implicit model of the evolutionary processes in order to account how or when the genetic variation was generated. While this section only intent to provide superficial review and mention important remarks about such methods, Chapter 4 from (Jobling et al. 2014) is recommend for an extensive revision.

Phylogenies are a basic approach to inform us about the evolutionary relationship of different species, especially such close species like modern humans and Neandertals. Sometimes the trees can be informative about both the actual ancestral relationships of these entities, as well as, the model by which diversity arose i.e. population growth or expansion vs structured populations (see **Figure 11**).

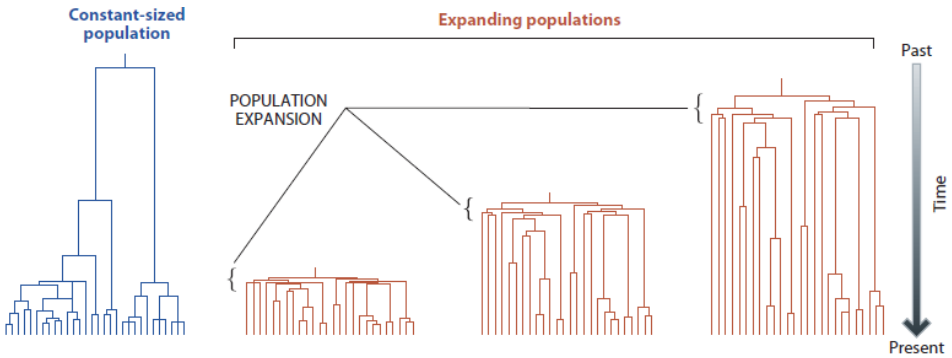


Figure 11: Genealogies for a constant sized population and three populations that have undergone population expansions at different times. [Adapted from (Jobling et al. 2014)]

Moreover, Coalescent analysis is another way of modeling trees, but of individual alleles in a population. However, the intention of coalescent analysis is very different from that of phylogenetic analysis, because they are not seeking to make a representative tree of the history of a DNA sequence, but rather to use a distribution of many thousands of possible DNA sequence trees to infer various parameters concerning the population, such as effective population size or the time to the most recent common ancestor (TMRCA) (Jobling et al. 2014). A coalescent approach is very powerful in making these inferences from within-species data; therefore, it forms a crucial analytical method in human evolutionary genetics. Although not the subject of this thesis, further details can be found elsewhere (Wakeley 2008).

Furthermore, another way of inferring past demographic events, is by analyzing the selective constraint by which a population is evolving, and further hypothesize about past process or events which could have

affected such constrains. For instance, the null case of the test will be that if the effective size of a population has been large enough in the past, then purifying selection will have acted to remove deleterious variants from the population. Therefore, one could assess the ratio of the occurrence of functional changes to that of more *"neutral"* ones and analyze if the observed fraction is more concordant with populations evolving under a neutral or lower selective constrain. Hence, if the proportion of nonsynonymous to synonyms changes is high, then this could be indicative that a population experienced demographic events like a bottleneck that changed its Ne and therefore modify the efficiency of purifying selection to remove damaging variants (**see Chapter 2**). Recent studies have used this conceptual approach to assess the demographic consequences of some present-day human populations having experienced a different population history over others (Lohmueller et al. 2008).

Finally, it is important to note that even though genetic data can help to reconstruct past adaptive and demographic events, analyzing data exclusively from present-day organisms can bias interpretations from their evolutionary history given that given that previous incidents could have been "erased" by later demographic or adaptive events. Therefore in order to make more complete and accurate deductions about the evolutionary histories of modern organisms, genetic data from both living individuals and well-preserved organic remains should be analyzed. Throughout this section I have summarized the importance of genetic data to reconstruct the evolutionary history of any organism, as well as the approximations and methods by which this can be attained. As mentioned before no method is good enough by itself to accurately reconstruct past evolutionary events. However, by combining evidence from different analyses which point to similar conclusions, better interpretations can be inferred. **Table 1** has a brief summary description on most of the methods previously describe, what sort of data is needed and the limitations from each approximation;

information which can be of help when choosing which the best approach to use.

Method	What it does	Type of data needed	Limitations
Calculation of θ	a measure of genetic diversity	DNA sequence data from the same region in different individuals	interpretation not intuitive
Principal component plots	shows genetic distance between individuals or populations	genetic variation data from many loci, typically from SNP chips	sometimes difficult to interpret results in terms of population history
STRUCTURE-like plots	assigns an individual to a group based on genetic diversity	genetic variation data from many loci, typically from SNP chips	requires prior estimation of the number of groups
Phylogenetic trees and networks: maximum likelihood (ML) and Bayesian	compares the relationship between DNA sequences	DNA sequence data from different species or individuals	Bayesian and ML methods have different limitations; see text and Box 6.4
Use of F_{ST}	a measure of population differentiation	genetic variation data from one or more loci	affected by both demography and selection
Parameter estimation using coalescent methods	provides an estimate of various population parameters	DNA sequence data from different individuals	dependent on assumptions made in the coalescent model
Extended haplotype tests	measures recent positive selection	SNP haplotype data from different individuals	only detects recent, strong, positive selection sweeps
Allele frequency spectrum tests	measures departures from neutrality	DNA sequence data from the same region in different individuals	affected by both demography and selection

Table 1: Descriptive and inferential methods for making evolutionary interpretations from genetic data. [Adapted from (Jobling et al. 2014)]

II BACKGROUND

2.1. Introduction to the field of ancient DNA

Although studying Neandertals evolution from a genetic perspective would have been unthinkable 40 years ago, nowadays we are able to analyzed DNA from fossil remains by means of the field of ancient DNA. The term ancient DNA (aDNA) refers to any DNA extracted from a dead organism or fossil remain, and can therefore be employed to consider all traces of ancient DNA subjected to autolysis or degradation (Rollo and Marota 1999). This discipline started thirty years ago, when Higuchi and collaborators (Higuchi et al. 1984) by means of molecular cloning, extracted and sequenced DNA fragments from a museum specimen of the quagga (an Equid from South Africa that became extinct in the nineteen century). This finding revolutionized the field of molecular biology by showing that it was possible to retrieve DNA from an organism that died several years in the past. Just a year later, Svante Pääbo (Pääbo 1985), retrieved DNA molecules from a 2,500 year old human Egyptian mummies using largely the same methodology. At this early stage, putative endogenous DNA was recovered exclusively by means of cloning and direct sequencing, an approach that proved to be methodologically difficult and extremely laborious (Pääbo 1985).

The first major breakthrough of the field of ancient DNA occurred with the advent of Polymerase Chain Reaction (PCR) developed by Kary Mullis (Saiki et al. 1985; Mullis Kb 1987). In contrast to molecular cloning, PCR allowed the specific exponential amplification of little amounts of DNA of interest, avoiding the need to sequence DNA from other environmental sources (i.e. DNA from microorganisms) and therefore open the possibility to recover DNA from different sources. Consequently, soon after the discovery of the PCR, the list of publications related to aDNA increased notably. Given the high copy number of the mitochondrial DNA (mtDNA) than nuclear DNA per cell, traditionally it has been targeted more by aDNA studies than has

nuclear DNA. Indeed, by amplifying mtDNA sequences form several emblematic fossil specimens like: Mammoths (Hagelberg, Thomas, and Charles 1994; Hoss et al. 1996; Noro M, Masuda R, Dubrovo IA 1998; A Cooper et al. 1992), Moas (Cooper et al. 1992) or Cave bears (Hanni, Laudet, and Stehelin 1994), the phylogenetic relationships between extinct and extant individuals became known. Furthermore, extraordinary claims that DNA from millions of years old specimens could be successfully extracted and sequenced, for example stemming from the dinosaurs bones or insects preserved in amber were published, but eventually proven to be due to contamination (Pääbo and Wilson 1991; Lindahl 1993a; Gibbons 1994). As consequence of such studies, general considerations and protocols began to be undertaken in order to avoid contamination and produce reliable results (Pääbo 1989). For instance, (Alan Cooper and Poinar 2000) published a set of guidelines "which all aDNA studies must follow", in order to ensure the quality of aDNA data produced and reliability of the conclusions (see **Table 2 and section 3.2**)

Furthermore, it began to be understood that endogenous aDNA was usually fragmented and found in very low amounts (Lindahl 1993b). Early studies showed that after the dead of an organism, most of the endogenous DNA is rapidly degraded by the action of endonucleases, either "liberated" from previous isolated cellular sub compartments (upon death) or belonging to microorganisms that infect and digest the remains of the corpse. Over time, other spontaneous processes act to fragment and damage DNA (see below). While in the cells of a living person, DNA is continually being monitored and repaired, after death, the systems that accomplish this cease functioning and thus physicochemical damage to the DNA can go on unimpeded. Therefore, DNA that is recovered from bones or other tissues of long-dead remains is severely damaged by hydrolytic cleavage of the sugar–phosphate backbone, resulting in short DNA fragments, loss of bases (abasic sites), chemical modification of bases, and inter- or

intramolecular cross-linking of sugar–phosphate backbones (see **Figure 12**).

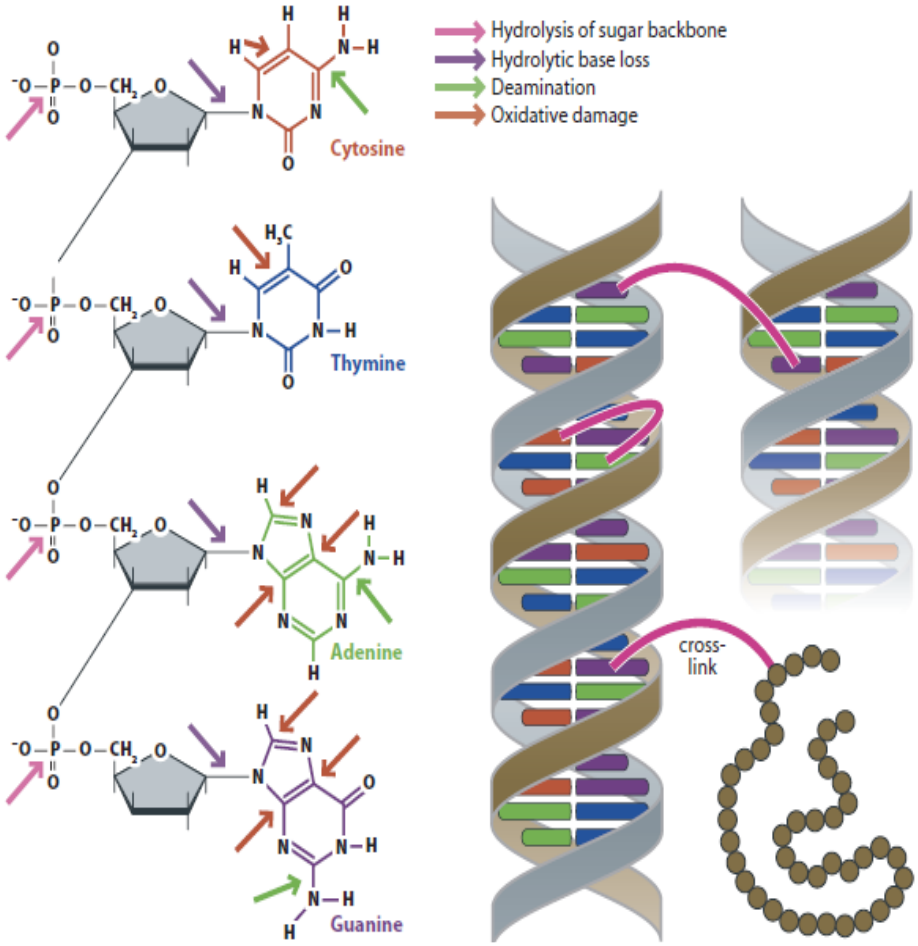


Figure 12: Postmortem DNA damage.
DNA is prone to spontaneous damage and degradation, including hydrolytic and oxidative damage, and cross-linking between or within helices, as well as to proteins. [Adapted from chapter 4 of Human Evolutionary Genetics edited by (Jobling et al. 2014)]

Early aDNA publications noted that among the chemical modifications reported from genetic sequences of ancient remains, deaminations, particularly cytosine to uracil conversions, were the most common

mutations (Hofreiter et al. 2001; Poinar 2002). However, other chemical nucleotide modifications (though much less frequent) like oxidized derivatives of cytosine or thymine called hydrations (important because polymerase amplification), are known to occur in ancient remains (Lindahl 1993b). Moreover, it seems that factors such as: low pH environments (Mitchell, Willerslev, and Hansen 2005; Lindahl 1993b), humidity, salt concentration and the temperature, appear to affect the rates at which these *post mortem* DNA modifications occur. For instance, it has been shown that low temperature can inhibit enzymatic activity and prevent the growth of microorganisms, as well as slowdown the chemical reactions that degrade DNA (Rollo and Marota 1999; Caramelli et al. 2003; Gilbert et al. 2007; Handt et al. 1996). Further publications into the matter have revealed that DNA is better preserved in cold, dry and alkaline environments, while its least conserved in hot, humid and acidic locations.

These early insights of the nature of aDNA, in addition to the implementation of anti-contamination measurements, eventually caused that aDNA eventually became a viable scientific discipline. In **section 3** the major considerations employed to guarantee the authenticity and reliability of the genetic sequences from ancient samples will be reviewed.

2.2 The first Neandertal sequences

The first Neandertal sequences were obtained by a team led by Svante Pääbo in 1997. By implementing the state-of-the-art extraction and anti-contamination protocols available at that time, Pääbo and his colleagues were able to amplify by PCR the mtDNA hyper variable region 1 (HVR1) from the Neandertal holotype specimen from Feldhofer cave in Germany. By comparing it against a panel of world-wide modern human and great mtDNA sequences, the data indicated that Neandertals were a sister group to anatomically modern humans,

providing no evidence of interbreeding between Neandertals and modern humans, at least to a level sufficient to result in Neandertal mtDNA introgression into the modern human mtDNA genetic pool (Krings et al. 1997). Moreover, during the fifteen years following that initial publication, other Neandertal sequences from different sites like Mezmaiskaya (Russia) and Vindija (Croatia) in 2000, Engis (Belgium), La Chappelle-aux-Saints (France), Les Rochers-de-Villeneuve (France) in 2004, El Sidrón (Spain) in 2005, 2006 and 2011, Monti Lessini (Italy) and Scladina (Belgium) in 2006, Teshik-Tash (Uzbekistan) and Okladnivok (Russia) in 2007 and Valdegoba (Spain) in 2012 were successfully amplified by the same approach (Ovchinniko et al. 2000; Krings et al. 2000; Serre et al. 2004; Bordes et al. 2005; Lalueza-Fox et al. 2005; Lalueza-Fox et al. 2006; Lalueza-Fox et al. 2011; Caramelli et al. 2006; Orlando et al. 2005; Krause, Orlando, et al. 2007; Dalén et al. 2012) (**Figure 13**).

A common observation of all these studies was that Neandertal mtDNA sequences were quite similar to each other - suggesting a general low diversity - and different to any reported human mtDNA, further corroborating the results from Krings et al 1997. Additionally, some studies began analyzing a possible phylogeographic structure; some of the basal sequences in the phylogenetic trees were from the easternmost Neandertals (located in Central Asia) or from the oldest ones samples (Valdegoba and Scladina) (Dalén et al. 2012). These data seemed to support an east-west genetic cline and also the existence of temporal bottlenecks that shaped the mtDNA diversity.

Recent western European Neandertals (roughly <50 kyr) constitute a tightly defined group with low mitochondrial genetic variation in comparison with both eastern and older (> 50 kyr) European Neandertals. Based on this data, Dalén *et al* 2012 suggested an initial divergence between Neandertal populations in Eastern and Western Europe approximately 55– 70 kyr ago, followed by an extinction of western Neandertals throughout most of their range and a subsequent

recolonization of the region either from the east or from a small refugium in the western part of their distribution, as the most likely explanation for their observations and simulations (Dalén et al. 2012).

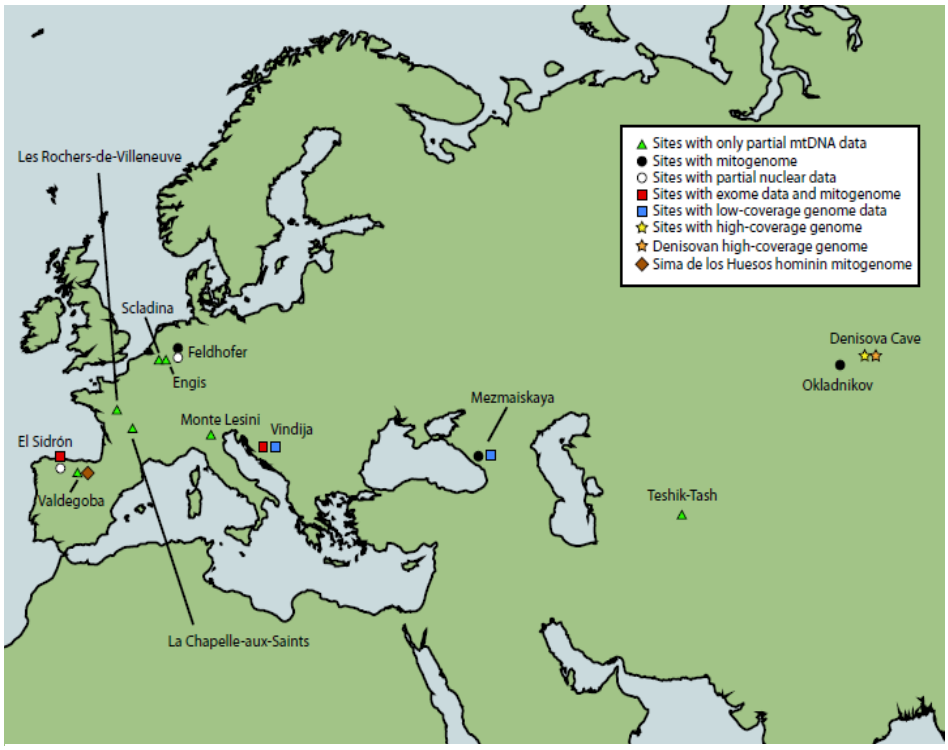


Figure 13: Geographic map showing the different Neandertal and Denisovan sites with different types of genetic data retrieved.
[Adapted from Sanchez-Quinto &Lalueza-Fox 2014, *accepted*]

2.3 Mitochondrial genomes and the advent of the new sequencing technologies

The second major breakthrough in the recovery of DNA from fossil remains occurred with the introduction of the Next Generation Sequencing (NGS) technologies (see **Section 3**) into the field of aDNA. After almost two decades of exclusively using conventional Sanger

Sequencing to recover genetic sequences, new technologies started to emerge. The high cost of the exiting technology, in addition to the interest in population-scale genome sequencing after the results of the human genome project led to the development of a diverse set of novel technologies that aimed to massively increase sequencing throughput and to reduce sequencing expenses (see **section 3.3.3-3.3.4**).

In 2005, (Margulies et al. 2005) developed a new sequencing DNA technology (sequencing by-synthesis) with which it was possible to sequence and assemblage millions bases in *reads* of ~100bp from independent starting molecules amplified into thousands of copies, in a four hour run. The aDNA community received NGS technologies with particular enthusiasm since these addressed two of the biggest limitation in the field, namely: the reduce proportion of endogenous compared to environmental DNA in ancient samples, and the fragmented nature of aDNA. Indisputably, the implementation of NGS allowed the field to move from a time and sample consuming conventional PCR-cloning of individual short DNA fragments, to genome-wide, cloning-free efforts, in which not only short fragments were usually sequenced, but far more contiguous regions sequences were retrieved.

The application of NGS to aDNA began almost immediately after the first sequencing platform (454) was introduce in the market, when Poinar et al (Poinar et al., 2006) sequenced 13 megabases of nuclear and mitochondrial; this study in addition to other sequencing projects like the low coverage Mammoth genome project (Green et al. 2006; Noonan et al. 2006; Poinar et al. 2006; Rasmussen et al. 2010), marked the dawn of *paleogenomics*.

Mapping analysis of such first NGS studies (see **section 3.6-3.11**) revealed that while the percentages of endogenous DNA were elevated in samples from permafrost environments (70-90%), the proportion of endogenous aDNA stemming from more template locations most commonly was as low as ~0.1-1 % (Green et al. 2010; Briggs et al. 2009).

On the other hand, the massive production of Neandertal sequencing data was not as straight forward, as previously conceived. Unfortunately all samples known at that time, originated from non-permafrost locations and initial NGS screenings showed that only 0.001% of DNA sequences determined from a typical well-preserved Neandertal specimen were derived from its mtDNA (Briggs et al. 2009). Therefore, given that a simple shotgun approach would require hundreds or thousands of high-throughput sequencing runs to recover a complete Neandertal mitochondrial sequence and that PCR was poorly suited to retrieve complete ancient mtDNA genomes (see **section 3.3.1**), a method that directly isolated specific DNA sequences from complex libraries of highly degraded DNA, was eventually developed (Briggs et al. 2009). As it will later be review in **section 3.5**, several capture enrichment techniques based on the hybridization to pre-define probes have been design to efficiently and specifically retrieve genomic regions of interest from samples containing marginally low endogenous amounts of DNA.

Finally, it eventually became possible to retrieve complete Neandertal mitochondrial genomes, first by shotgun sequencing of a very well preserved sample from Vindija cave (Green et al. 2008) and later with targeted hybridization-capture enrichment methods (Briggs et al. 2009). The whole mtDNA genome allowed a more precise estimate of the divergence time between recent human and Neandertal mtDNA lineage, which was reported to be ~660,000 years ago. Furthermore, another striking observation was that the ratio of non-synonymous to synonymous evolutionary rates was significantly higher on the Neandertal lineage, a result that would fit with Neandertals having a smaller effective population size, and thus evolving under lower selective constrain than modern humans (Green et al. 2008). By 2009, the analysis of six complete Neandertal mtDNA genomes indicated that the variation among Neandertals was approximately one-third of that estimated for modern humans worldwide, suggesting a female effective

population size of less than 3,500 individuals. This finding was surprising given that the Neandertal sequences stem from several distinct time points spanning thousands of years and thus it appears to be a conservative estimate with respect to sampling at a contemporaneous time period (see **Figure 14**). The most recent common ancestor (MRCA) of the Neandertal samples analyzed was estimated to have lived ~110,000 years ago, which is significantly less of what is estimated for modern human mtDNAs (Briggs et al. 2009).

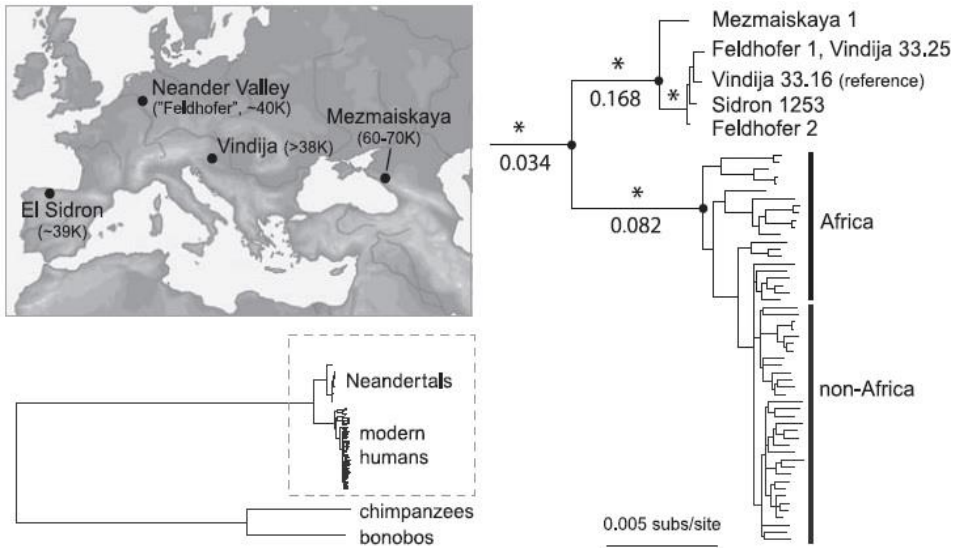


Figure 14: Phylogenetic ties of samples from where whole mtDNA have been recovered [Adapted from (Briggs et al. 2009)]

Furthermore, by incrementing the amount of genetic information per nucleotide position, these sequencing technologies allowed not only precise contamination estimates, but also a more complete description of the deamination events in aDNA fragments (see **section 3.10**). Such chemical modifications are also referred as *misincorporation patterns*. Thus, by generating of thousands of sequencing reads from ancient remains, the presence of a C to U deamination (which seem to increase

with time), ultimately became an hallmark of aDNA and its occurrence a way to authenticate DNA in a sample (Briggs et al. 2007; Sawyer et al. 2012). As later review in **section 3.12**, there are other matters of important details that need to be taken into account, in order to further authenticate aDNA data and ensure the reliability of downstream analysis.

2.4 The first Neandertal nuclear DNA sequences

As Neandertal mitochondrial diversity was being studied, attention also turned to nuclear loci. Although challenging, given the lower proportion of nuclear DNA compared to mtDNA, researchers were thrilled by the idea, as it provided the possibility of assess if emblematic functional and phenotypic modern human traits were shared by Neandertals.

Between 2007-2009, by amplifying small nuclear regions encompassing functional variants, researchers found that some Neandertals were probably red-haired and pale skinned (see **Figure 15**) (Lalueza-Fox et al. 2007), they had bitter taste perception ability (Lalueza-Fox et al. 2009) and presented the ABO blood type O (Lalueza-Fox et al. 2008). In addition, by having the same functional variants as modern humans in the FOXP2 - a gene that when mutated generates a speech and language impediment -, suggested that Neandertals were able to communicate with similar language capabilities to ours or at least they had the genetic potential to do so (Krause, Lalueza-Fox, et al. 2007). Notwithstanding, recent studies have found differences between most modern humans and Neandertals in a regulatory element near the FOXP2 gene that could have functional implications (Maricic et al. 2013).

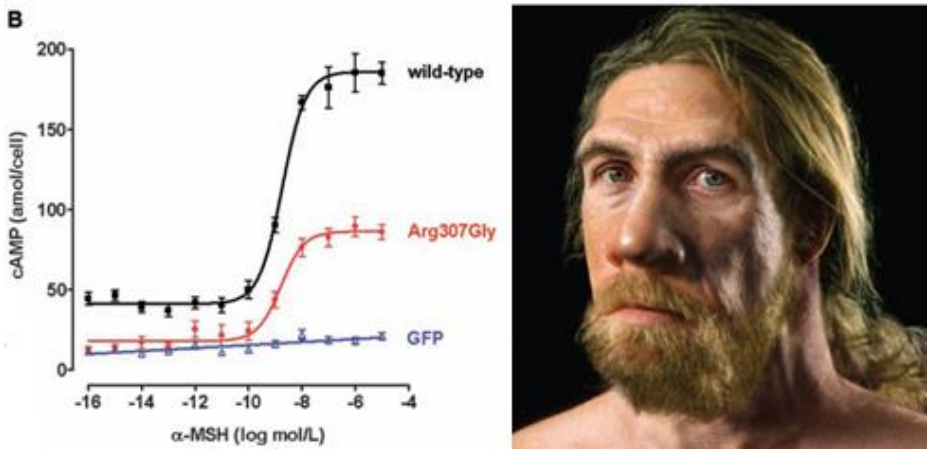


Figure 15: Phenotypical reconstruction of a Neandertal appearance.
A genetic variant found in El Sidrón sample individual, causing an amino acid change of arginine to glycine within the sequence of the MC1R gene, has been levels of the pigmentation pathway activation and thus he probably would have had a fair-skin and was red-haired. [Adapted from (Lalueza-Fox et al. 2007)]

While recovering short pieces of nuclear DNA became possible in well-preserved and uncontaminated specimens, the sequencing of a whole Neandertal genome proved to be a complicated endeavor due to the low amount of nuclear DNA sequences relative to environmental sequences and the limitations of the available technology. Two pioneer studies that intended to massively recover Neandertal nuclear DNA either by cloning and sequencing short fragments of DNA (Noonan et al. 2006) or by metagenomic sequencing (Green et al. 2006), managed to recover 65 Kb of nuclear DNA, and 1Mb of sequence respectively, and estimated coalescence times between modern humans and Neandertals to be roughly 700,000 and 500,000 years ago. Notwithstanding, it was subsequently demonstrated that at least ~20% of the data generated by the later study derived from modern human contaminants DNA (Wall and Kim 2007). As an outcome, more rigorous actions were adopted in the process of constructing the sequencing libraries as well as the use of bacterial endonucleases in order to eliminate as much potential environmental and human contamination as possible (Green et al. 2009; Green et al. 2010).

2.5 The Neandertal and Denisovan draft genomes

The year 2010 saw not only the publication of the long-expected Neandertal draft genome (Green et al. 2010) but also the discovery of the previously uncharacterized, Denisovans (Reich et al. 2010) at Denisova Cave. Currently only two teeth and a finger bone (the latter with extraordinary levels of DNA preservation, ~70% of endogenous DNA) have been attributed to the Denisovans. Both nuclear and mitochondrial DNA extracted from these remains suggests that, Denisovans were as genetically diverse as two living humans from different continents and more diverse than Neandertals throughout their geographical range; suggesting that their effective population size was relatively large (Pennisi 2010). Unfortunately, at that time there was only sufficient endogenous DNA available from the finger bone in order to sequence completely both the nuclear and the mitochondrial genome. Furthermore, by employing a user-defined hybridization-capture method, a high-coverage mtDNA genome was retrieved (Maricic, Whitten, and Pääbo 2010) and by comparing it to the sequences of modern humans, Neandertals and great apes, it was estimated that it diverged from the common ancestor of modern humans and Neandertals around one million years ago. Moreover, as both nuclear archaic genomes were sequenced, clearer phylogenetic relationships not depending on only one genetic locus, were established.

The most recent common ancestor of modern humans, Neandertals and Denisovans was found to have lived at least 800,000 years ago, while the Denisovan and Neandertal genomes were more closely related to each other - as sister species - and their divergence time was estimated to be 600,000 years ago (see **Figure 16**).

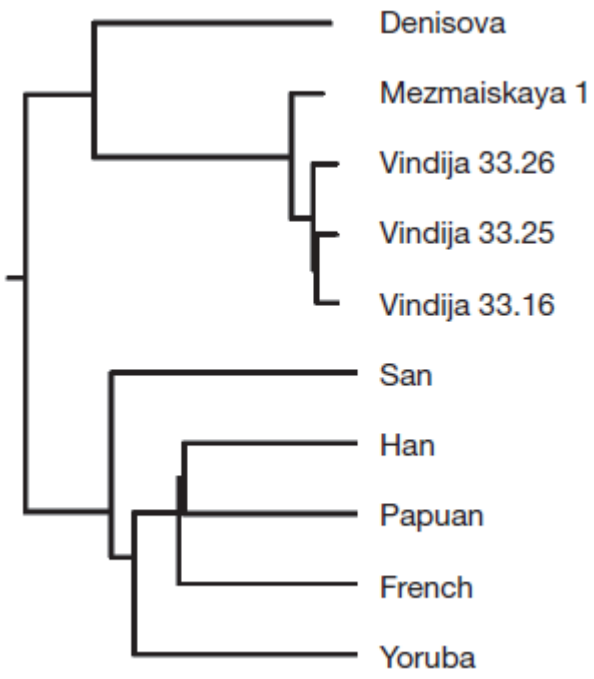


Figure 16: A neighbor-joining tree based on pairwise autosomal DNA sequence divergences for five ancient and five present-day hominins.
[Adapted from (Reich et al. 2010)]

In addition, it was found that all non-Africans shared 1–4% more derived alleles with Neandertals than with sub-Saharan Africans, whereas Melanesians and Australian aborigines also shared ~3.5% more of their DNA with the Denisovan individual than with the rest of modern human populations. The Neandertal signal was later also observed in African populations, which is likely the result of back-to-Africa migrations (Wang et al. 2013; Pickrell et al. 2014). These results were interpreted as evidence of Neandertals interbreeding with the ancestors of all non-Africans, and in addition the ancestors of Melanesians with a Denisovan-like population; moreover, marginal Denisovan admixture was also reported in continental Asian populations (Pontus Skoglund and Jakobsson 2011) and was later confirmed (Prüfer et al. 2014).

Notwithstanding, the proportions of admixture are likely overestimates if some degree of structure was present among ancient humans in Africa, as pointed out in (Green et al. 2010; Eriksson and Manica 2012; Lowery et al. 2013; Eriksson and Manica 2014). If this was the case, ancient hominin population structure within Africa followed by a series of founder events and reduced migration, would have caused Neandertals (as well as Denisovans) to seem more closely related to modern non-Africans than to present-day African populations (see **Figure 17**). However, under this scenario there are several major observations difficult to reconcile with such hypothesis.

For instance, given that Neandertals and Denisovans seem to be phylogenetically related as sister species, one will expect under this scenario that all non-African modern humans will share fairly the same amount of derived alleles with archaic humans. However, the observation that the two archaic humans have different relatedness stories to present-day humans populations, questions the extent of such argument. Moreover, the fact that the dimensions of Neandertal-like haplotypes found within contemporary modern humans have a predicted length more concordant with a late Pleistocene genetic admixture rather than with an ancient population structure (see below), also complicates the veracity of such scenario. Finally, as later mentioned, the fact that present-day humans are more closely related to a Neandertal sample from the Caucasus than to Neandertals from Europe or South Siberia, would be unexpected under an scenario in which the ancestors of Neandertals and modern humans emanated from the same ancestral population.

At the end, while aDNA data from Late Pleistocene modern humans close to the putative admixture events will be essential to favor one alternative other the other (based each model expectations), current lines of evidence show that although incomplete lineage sorting and not introgression could explain some genetic similarities between modern non-African humans and Neandertals, it will certainly cannot account for all of them.

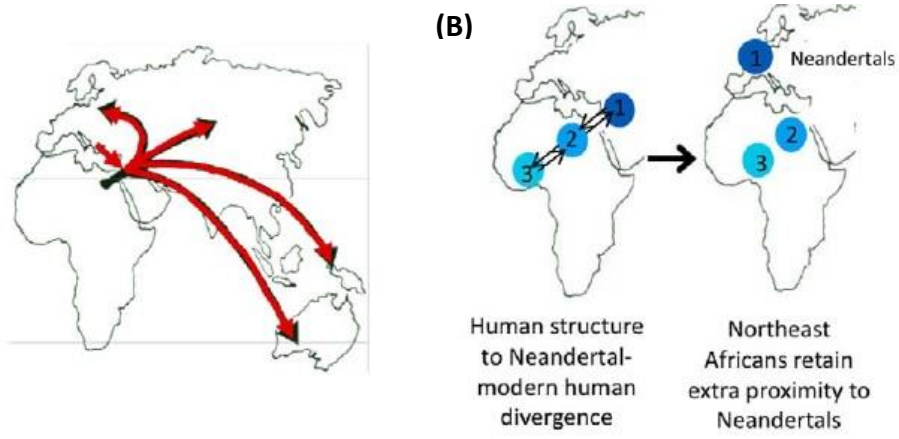


Figure 17: Schematic representation of two possible scenarios to account for the fact Neandertals seem to be more closely related genetically to non-African modern humans than to present-day African populations.
(A) A genetic introgression event(s) occurred between Neandertals and the ancestors of modern non-African humans, as the latter were migrating out-of-Africa.
(B) Ancient hominin population structure within Africa followed by a series of founders events and reduced migration, could have caused Neandertals to seem more closely related to modern non-Africans than to present-day African populations.
[Modified from (Sankararaman et al. 2012)]

On the other hand, another immediate use of a Neandertal draft was the so long-awaited possibility to defined genetic changes unique to our lineage. By comparing the Neandertal genome to the genomes of five present-day humans from different parts of the world, first analysis showed that among all the positions where present-day humans were fixed for a derived state and where Neandertals carry the ancestral alleles, 78 nucleotide substitutions seemed to have change the amino-acid group. Additionally, 5 genes had more than one amino-acid changing substitution, derived in present day humans and ancestral in Neandertals, and interestingly 3/5 of these genes were related to skin morphology and physiology. Also initial analysis to detect genomic regions that may have been affected by positive selection in ancestral modern humans, suggested that genes involved in metabolism and in cognitive and skeletal development, could have been involved in this process (Green et al. 2010).

Furthermore, in an attempt to obtain better quality data to infer amino-acid changes occurring specifically within the modern human lineage, ~14,000 protein-coding positions assumed to have changed on the human lineage since the last common ancestor shared with

chimpanzees were captured and sequenced in one Neandertal and 50 present-day humans samples (Burbano et al. 2010). Information was successfully retrieved from around 10,000 positions from which all present-day humans had a change different from the chimpanzee sequence, whereas for 9525 (87%) the Neandertal sample had the same nucleotide as humans and thus represented changes common to both hominin lineages. Notwithstanding, there were 88 positions were Neandertal had the ancestral chimpanzee like nucleotide, and all the present-day humans had the same mutation but different from that of the ancestral consensus. This 88 positions occurring within 83 genes, represented a list of recently fixed and unique substitutions to our evolutionary lineage.

Although this represented a first catalogue of those genetic changes underlying at least some of the traits possible responsible for "what makes us humans", such a list of changes was difficult to interpret since the genes where they occurred did not had any associated function or pathway in common. Additionally, the high incidence of *post mortem* modifications, the sparse regions of the genome with overlapping reads (i.e. less than 50% of the genome has at least 1 read covering that region) and its depreciable coverage (1.3X), prevented further analysis to be undertaken as well as fine tuning of previous estimations.

2.6 High-coverage genome

A major technical breakthrough involved a novel library preparation method that exploited single-stranded DNA and greatly increased the yield of sequencing from ancient samples. Briefly, instead of building the libraries exclusively from double-stranded DNA - where only sequences without "nicks" or single strand breaks can be incorporated into NGS sequencing libraries - the new method first denaturates DNA fragments and incorporates the single strands of DNA into NGS libraries, allowing for the recovery of significantly more DNA molecules than hitherto possible. By applying this new method a 30X

coverage genome from the same Denisovan sample (Meyer et al. 2012) and a 54X coverage genome from a female Neandertal toe bone (Prüfer et al. 2014) also from Denisova Cave - known as the Altai Neandertal - were generated.

Having high-quality genome data does not only offer refined insights into Neandertal relatedness to modern humans, but also allows us to start addressing questions concerning their diversity and demographic history, something that could not be done with low coverage data. For instance, under a no gene flow scenario, the date of the split of the archaic and modern human populations, which by necessity is more recent that sequence divergence, can be estimated. Recently, mutation rates have been a subject of great debate (Scally and Durbin 2012). Based on a mutation rate of 1.03×10^{-9} derived from the fossil record, the population split between Denisovans, Neandertals and modern humans probably occurred between 383,000–257,000 years ago, while the populations that evolved into Neandertals and Denisovans separated roughly ~236,000–190,000 years ago (Prüfer et al. 2014).
In addition, by using high-quality archaic genome data, a more precise idea of how and when the admixture with archaic humans occurred is beginning to emerge. By coupling high coverage archaic and modern human genomes, the amount of DNA introgressed from Neandertals into non-Sub Saharan Africans has been refined to a range of 1.5–2.1% of Neandertal ancestry in present-day populations (Sankararaman et al. 2014; Prüfer et al. 2014). It has also been observed that Neandertal-derived DNA in all non-Africans is more closely related to a low coverage genome from the Mezmaiskaya skeleton in the Caucasus than to the Altai or to the Vindija genome (Prüfer et al. 2014). The linkage disequilibrium (LD) pattern of haplotypes of suspected Neandertal origin suggests a date of admixture date between 37,000 and 82,000 years ago (Sankararaman et al. 2012).
Altogether these observations seemed to indicate that a currently unsampled Middle Palaeolithic Neandertal population living in the

Levant and/or Western Asia probably encountered modern humans as they were migrating out-of-Africa, and subsequently spreading the signature of introgression as they populated the rest of the world (**Figure 18**).

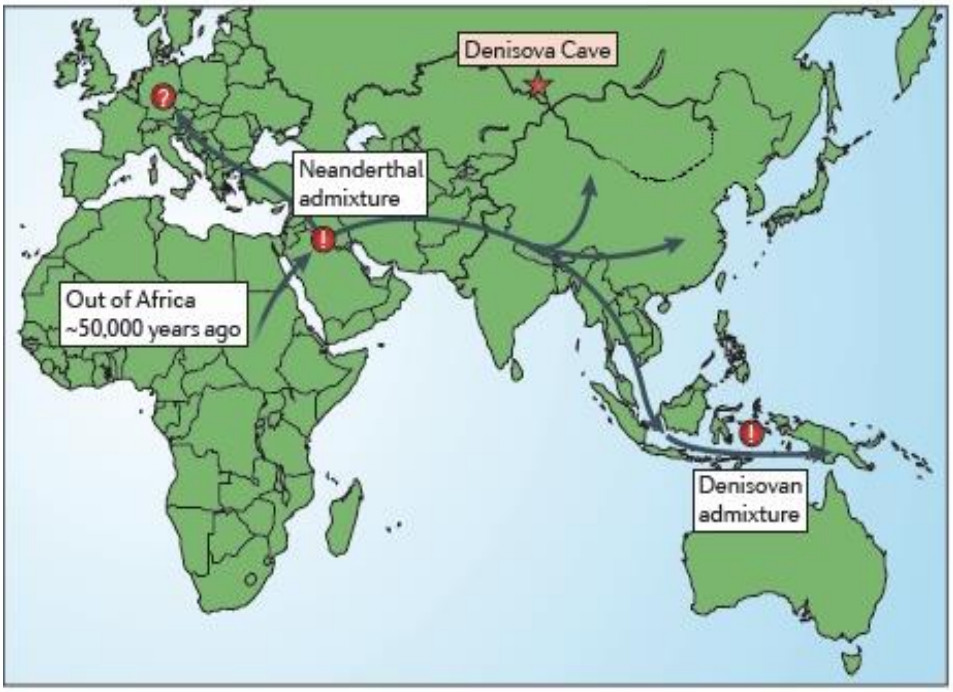


Figure 18: Dispersal of modern humans from Africa [Adapted from (Green et al. 2010)].

Furthermore, it has recently been shown that East Asians and Native Americans may have 20% more Neandertal admixture than other non-African populations, suggesting that a second introgression event could have taken place after European and Asian modern human populations diverged (Meyer et al. 2012; Wall et al. 2013). This latter finding is unexpected given the archaeological evidence of a long-term occupation of Neandertals in Europe, and a possible late overlap with early human migrations to the same continent as well. Moreover, late Palaeolithic and Mesolithic modern human genomes have so far failed to demonstrate a closer relatedness to Neandertals (Raghavan et al.

2014; Olalde, Allentoft, et al. 2014). High coverage genomes of late Pleistocene Europeans - and from other populations too - will be needed in order to estimate accurately, if other admixture events could have occurred with Neandertals or Denisovans. Interestingly, some lines of evidence suggest that interbreeding may have been limited by genetic incompatibilities (see below) and thus an increase in Neandertal admixture would only be temporally observed close to the interbreeding events(s) (Sankararaman et al. 2014).

A further interesting application of the high coverage genomes is to investigate in detail the archaic introgressed regions, and see whether they harbor genetic variants that could be biologically important to modern humans. Several recent publications suggest that some archaic variants could have been advantageous or at least functionally relevant after been introgressed into modern humans (Abi-Rached et al. 2011; Mendez, Watkins, and Hammer 2012; Mendez, Watkins, and Hammer 2013). For instance, Neandertal haplotypes in European and East Asians are enriched for genes harboring keratin filaments - a protein expressed in skin, hair and nails - suggesting that skin adaptation to colder, non-African environments could have been enhanced after the admixture event. Inversely, there seem to be large "deserts" of Neandertal ancestry, implying that purifying selection may have acted to remove genetic material derived from Neandertals which could have been deleterious on a modern human background (Sankararaman et al. 2014; Vernot and Akey 2014). Furthermore, genes that are more highly expressed in testes than in any other tissue are especially reduced in Neandertal ancestry, and that there is an approximately five-fold reduction of Neandertal ancestry on along the X chromosome (Sankararaman et al. 2014); facts which can be interpreted as selection eliminating genes that have may have reduced male fertility. Furthermore, the known differences in effective population size between East Asians and Europeans, could have resulted in less efficient selection to remove Neandertal-derived deleterious alleles and thus be the cause for the excess of Neandertal signal observed in East

Asians populations (Sankararaman et al. 2014); however (Vernot and Akey 2014) state it was more likely to have occurred from further interbreeding in the East.

Finally, the high-coverage Neandertal and Denisova genomes provided a sound basis to identify genomic changes specific to modern humans and with that a more accurate list of substitutions accountable for "what makes us modern humans" is beginning to emerge.
By comparing data from 1,094 present-day humans (1000 genomes data), the high-coverage of both Neandertal and Denisovan genomes as well as the great apes genomes, a genome-wide high-quality catalogue of the sites unique to modern human lineage has been finally created. While this list was previously consider to be 3-4 times bigger, only 31,389 single nucleotide of such substitutions and 4,113 short insertions and deletions (indels) are shared by all present-day humans (the numbers increases if changes shared also by archaic humans and apes and segregating at low frequency in humans are taken into account); however it has been surprising that this list of genetic changes, that distinguish us from our nearest extinct relatives is so small. For instance, if one focuses on functional relevant changes, barely 96 fixed amino acid substitutions (occurring within 87 protein sequences) are common to all present day humans, while only in the order of three thousand fixed changes could potentially influence gene expression. Therefore, in order to understand how those genetic changes that are unique to our evolutionary lineage translate into the morphological and biological differences that clearly distinguish us as "modern humans", one has to comprehend the functional importance and impact of those genetic substitutions.
A way to prioritize changes in the catalogue for functional studies is to identify those that show signs of having arisen to high frequency rapidly, as they may have been affected by positive selection. (Prüfer et al. 2014) scanned the genome for regions where the Neandertal and Denisovan genomes fall outside of the variation of present-day humans,

in order to look those such regions. They found that 63 regions contain 2,123 substitutions and 61 indels that are fixed or of high-frequency (.90%) in modern humans, and were both ancient humans are homozygote for the ancestral allele, are likely candidates to have been affected by positive selection. Interestingly, one gene (RB1CC1) is essential for the maintenance of neuronal stem cells in the adult sub-ventricular zone like VCAM1. Functional investigations will be necessary to clarify to which extent these and other changes affect any phenotypes in present-day humans.

In addition, further insights from the Denisovan and Neandertal high-coverage genomes, have allowed us to take an initial glance into the demography of archaic humans and compare it to our own.

The demographic history of the population can be reconstructed from the distribution of the times to the most recent common ancestor of the two the copies of the genome that a single person carries. Therefore by using a pairwise sequentially Markovian coalescent (PSMC) model (H. Li and Durbin 2011), (Prüfer et al. 2014) inferred that the ancestral populations of Neandertal and Denisovans decreased in size for hundreds of thousands of years sometime after around half a million years ago, while the population ancestral to present-day humans increased in size (**Figure 19**).

Moreover and quite interestingly, the Altai Neandertal individual presents long stretches of homozygosity which translates to an inbreeding coefficient of 1/8, suggesting that probably that the parents this individual were as closely related as half-siblings (Prüfer et al. 2014). Furthermore, the Neandertal high-coverage autosomal genome carries 1.7–1.8 heterozygous sites per 10,000 bp. This is 84% of the number of heterozygous sites in the Denisovan genome, 22–30% of that in present-day non-African genomes, and 16–18% of that in present- day African genomes. is observed in the Denisovan high coverage genome (Prüfer et al. 2014).

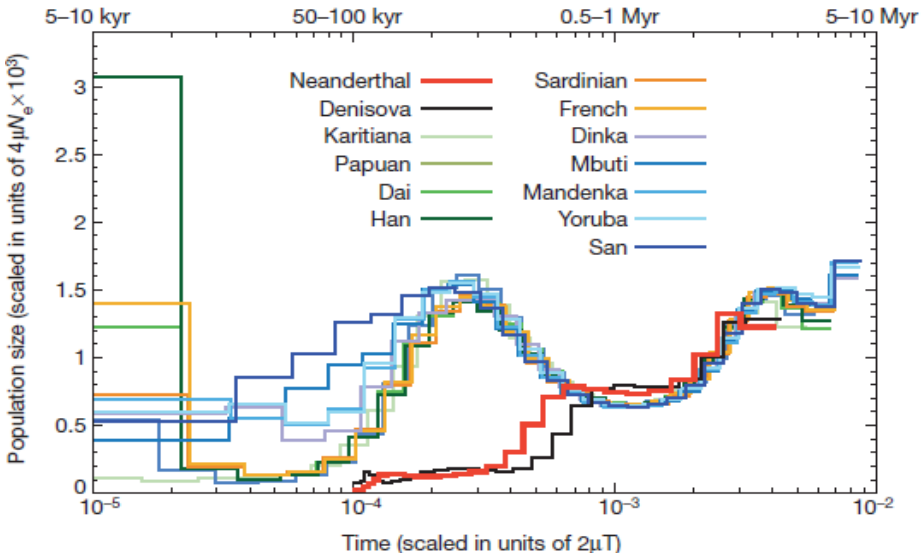


Figure 19: Inference of population size change over time.
The y axis specifies a number proportional to the population size Ne. The x axis specifies time in units of divergence per base pair (along the top in years for mutation rates of 0.5 x 10^{-9} to 1.0x 10^{-9} per site per year). [Adapted from (Prüfer et al. 2014)]

2.7 Super archaic DNA

A further interesting aspect of the evolutionary history has been revealed by the sequence analysis of the mitochondrial genome of ca. 400,000-year-old hominin from the Sima de los Huesos in Atapuerca (Spain). As mention before, the skeletal remains of this hominin, previously been classified as *H. heidelbergensis* and initially thought to represent the ancestor of Neandertals, have now been shown to present an almost full set of Neandertal derived traits in their face and teeth, whereas the braincase still retained "primitive" features. The fact that derived braincases shapes as well as other Neandertal morphological attributes are not found in Europe before ~200,000 years ago, supports an Accretion origin of Neandertal, which suggests that Neandertal features did not evolve as a block but rather their derived traits were gradually fixated (Arsuaga et al. 2014). Intuitively one will expect that

the sequence of hominin from Sima de los Huesos would be phylogenetically close to Neandertals, if a Neandertals lineage where already present at that time or clade basal to all know Neandertal mtDNA genetic diversity if they represent an ancestral to Neandertals.
Interestingly and enigmatically, the mtDNA sequence from one of these Sima de los Huesos individuals seems to be phylogenetically most similar to that of Denisovans (Meyer et al. 2014) found thousands of miles away in Siberia, than to Neandertals most commonly found in Europe (**Figure 20**). Whereas this surprisingly finding demand the nuclear genome sequence of these specimens in order to ascertain their precise relationship to archaic and modern humans, it's fascinating to realized that the relatedness of archaic and modern humans is far more complicated than previously thought.
Additionally and also remarkably, this study (Meyer et al. 2012) provided evidence that aDNA techniques have become sensitive enough to recover and analyze DNA from Middle Pleistocene hominin remains, even from non-permafrost environments.

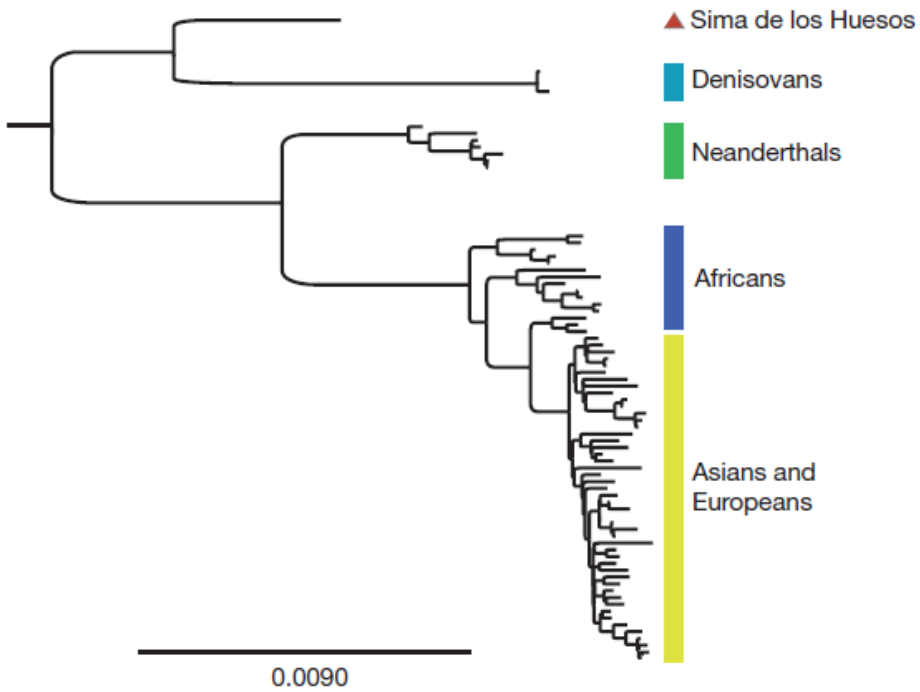


Figure 20: Inferred phylogenetic tree of hominin mitochondrial sequences including the Sima de los Huesos mtDNA genome [Adapted from Meyer et al 2014].

Almost 20 years of Neandertal paleogenetics and paleogenomics have significantly increased our knowledge both about their evolutionary history, as well as about our own. Neandertal genetic data have confirm that not only Neandertals are our closest know hominin relative, but that their relatedness to modern humans is not as simple as previously conceived. Neandertals seemed to have interbred with the ancestors of non-sub Saharan present day humans to an extent that although it didn't left any traces in the mitochondrial lineages from contemporary humans, it is still detected as a Neandertal ancestry of 1.5–2.1% in the nuclear genomes of those humans. Moreover, Neandertal´s origin is still not fully understood. While morphological findings (Rosas et al. 2006; J. Hublin 2009; Arsuaga et al. 2014) point towards gradual appearance of its traits starting around half a millions ago, genetic data of a their suspected "ancestor" neglects a recent common origin with any known Neandertal sequence. Additionally, comparison of the high coverage genomes of Neandertal and Denisova to databases of present-day human genome-wide variation, have permitted to accurately infer changes specific to our evolutionary lineage, and thus has open the possibility to start addressing the functional implications of those genetic variants that might "make us modern humans".

Furthermore, by addressing the mitochondrial genetic diversity from a geographically broad sampling of Neandertal specimens, in addition to the genetic variability and inbreeding insights from the Altai genome, a preliminary conception about their genetic diversity and population size is beginning to emerge.
However, there are still aspects about Neandertals evolutionary history that remain elusive. For instance, it is unfortunate that when the admixture analysis with modern humans was undertaken, only three non-African populations where selected, disregarding other human populations with more complex demographic histories.

Moreover, a subject that has always fascinated scientist and public alike are the known morphological differences between modern humans and Neandertals. While the availability of the high coverage Neandertal and Denisova genomes have allowed us to define a list of derive changes specific to our lineage, such a list is still unavailable for archaic humans. Obtaining a list of reliable changes specific to the Neandertal or the archaic lineage from several individuals could shred light for identifying the genetic basis underlying some of their particular phenotypic traits. Such a list of genetic changes could be of great interest, not only to understand why a certain trait did eventually become fixed but also to better comprehend the genetic changes responsible for the morphology of such features in present-day humans.

Furthermore, even though their mitochondrial genetic diversity as well as the genetic variability from the Altai genome suggests their effective population size was low and probably decreasing in time, the lack of a broader geographical and temporal sampling, prevents us to address if such observations reflect the demographic history of a population or rather they are the consequences of local events.

In addition, it will also be of interest to address to which extent the putative decreasing effective population size could be reflected in the patters of genetic variation of Neandertal samples from later time points, as well as its implication on their fitness and to which extent they contribute to their eventual extinction.

Finally, the genetic composition of Neandertals social groups has not been characterized before and thus the extent of which the patterns of genetic diversity from its members potentially reflect the dynamics of the population remains largely unaddressed.
The aim of the present thesis is to contribute to clarify some of the latter mentioned issues by better understanding their demographic and evolutionary history at different resolution levels.

III METHODS

3.1 How to obtain DNA from ancient remains?

Recovering, extracting and analyzing DNA from ancient remains is not an easy and straightforward endeavor. If reliable results want to be produced, sufficient effort and special considerations should be devoted in order not only to avoid as much environmental or laboratory contamination as possible, but also to ensure the quality of the data been generated. This methodological chapter aims to describe both the experimental proceedings to neatly extract and sequence DNA from ancient specimens, as well as to process and filter NGS data in order to produce good quality sequence data. Throughout the following sections, emphasis will be strengthened on how to elude contamination, while producing trustworthy results by undertaking different approximations.

The experimental section starts reviewing the general practices to avoid contamination during the excavation, extraction and isolation of DNA from fossils (**section 3.2**). Later on through sections **3.3 and 3.5,** a brief review on the nature, advantages and disadvantages of the main technological approximations employed to recover DNA from ancient remains (namely from PCR to capture-enrichment techniques) is discussed.

In addition, sections **3.4 and 3.6-3.12** will focus on major strategies to recognize, process, map, and analyze aDNA sequences produced by NGS technologies. Important technical processing and filtering procedures to ensure the quality of the sequence data being produced (**section 3.6-3.11**), as well as, major considerations to guarantee the authenticity of the DNA sequences retrieved and estimate human contamination (**section 3.12**), will also be reviewed.

Finally, it is worth to briefly discuss the general outline that a NGS aDNA project usually has. Historically, aDNA projects starts by receiving the samples of interest to a laboratory with appropriate

facilities in order to cautiously recover its DNA. However, a recent study has suggested recovering fossils directly from archeological sites in the most pristine possible conditions (see **section 3.2**). Next steps within an aDNA project involve extracting and isolating the DNA, and unless there is *a priori* knowledge of the amount of endogenous DNA in the sample, the following step should be to carry out a NGS screening shotgun run. Information from that first sequencing run will allow the researcher to estimate how much endogenous DNA, and laboratory (usually human contamination), as well as, environmental contamination (microorganism contamination) is present in the sample.

Once the nature of the DNA is understood, a decision should be taken towards choosing the most efficient and feasible approximation in order to recover as much relevant data as possible, depending on the proportion of endogenous DNA estimated and the scientific question pursued for the project. For instance, if a sample with a high fraction of endogenous DNA is recovered and the project aims to recover a whole genome, direct sequencing is recommended. On the other hand, if a sample has very little DNA from the target organism or the scientific scope involves the recovering of only a certain genomic region, target enrichment methods are preferred.

Finally, after sequencing has been performed, the major concerns are to accurately map, process and call genetic variants, in order to obtain reliable sequence data that can be unbiasedly interpreted by the researchers.

3.2 Sample retrieval and DNA extraction: general considerations to avoid contamination

3.2.1 Recovering aDNA and major anti-contamination precautions

An ideal approach to ensure that a sample is free from contamination, should be to access an archeological site and retrieve the fossil remains

in the most pristine and sterile possible conditions, in order to avoid direct contact from excavation personal and to isolate them from the enviroment, as much as possible. Although it hasn't been widely applied within the aDNA community, a recent published anti-contamination excavation protocol by Fortea et al (2008) (Fortea et al. 2008), suggests wearing full ancient DNA lab gear, clean excavation tools with sodium hypochlorite, avoid washing remains and immediately after their discovery introduce the samples into sterile plastic bags and instantly freezing them at -20°C until its arrival to the aDNA laboratory. By carrying out this approximation, (Fortea et al. 2008) showed that samples retrieved with their protocol had significantly lower contamination estimates, than those which that had been excavated before its implementation (**Figure 21**).

Furthermore, once samples are retrieved they should be transported as soon as possible to an aDNA laboratory. An ancient DNA laboratory is an extremely sterile environment, exclusively dedicated to aDNA extraction and where no previous DNA amplification has been undertaken. Appropriate clothing to be used exclusively in this area, should be worn at all times, including full-body sterile footwear and a plastic mask protection (Alan Cooper and Poinar 2000).
An aDNA laboratory should be physically separated from any molecular biology laboratory (ideally different buildings or floors), have a positive air pressure system, overnight UV light irradiation and be continually cleaned with bleach on the bench surfaces (Alan Cooper and Poinar 2000). Furthermore, all sample and reagent handling has to be performed in a laminar flow cabinet routinely irradiated with UV light and use sterile tips at all times, in order to help avoid intra-laboratory derived contamination.

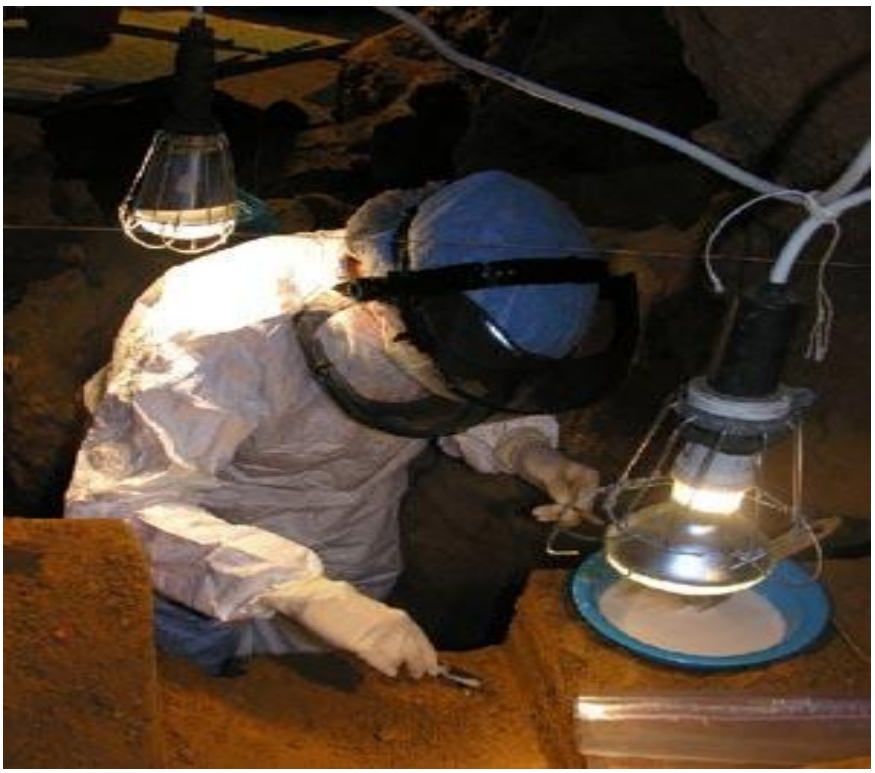


Figure 21: Archaeologist excavating in El Sidrón; clean-room gear was provided in order to reduce exogenous DNA contamination as for of the anti-contamination protocol from Fortreat el at (2008). [Adapted from Fortea et al 2008]

The concern for contamination is such that some aDNA laboratories have implemented routine checks of their reagents, as well as, other working material, in order to detect any possible source of contaminating DNA.

Table 2 shows four general concerns for aDNA studies, as well as, the main procedures by which one can address them, in order to enhance the probabilities to recover authentic aDNA sequences.

PRODUCING AUTHENTIC ANCIENT DNA SEQUENCES	
Aim	Procedure
Prevent modern DNA contamination	excavate and curate ancient specimens under 'clean' conditions
	work in a properly equipped and physically isolated clean laboratory facility for DNA extraction and NGS library preparation, in which no PCR work is ever done
	decontaminate surfaces, reagents, and tools frequently, using UV light and/or bleach
	decontaminate ancient specimens before processing
Detect and quantify modern DNA contamination	use multiple *blank* controls in extractions and PCRs
	clone PCR products and sequence multiple clones
Show that results are consistent with ancient DNA	quantify template DNA
	show that amplifiable PCR products are short (<500 bp)
	demonstrate inverse relationship between amplification length and efficiency, reflecting degradation and damage in the ancient DNA template
	ideally, observe age-dependent patterns in both sequence diversity and DNA damage
Show reproducibility	repeat findings within laboratory
	have results reproduced independently by another laboratory

Table 2: Recommended considerations in order to authenticate data from aDNA studies. [Adapted from Chapter 4 of (Jobling et al. 2014)].

3.2.2 DNA extraction

Techniques employed to extract DNA from ancient specimens have varied accordingly to bone type being sampled and new technological improvements being implemented. Notwithstanding, most methods rely on an initial overnight exposure to EDTA to remove mineral salts, followed by a protein digestion of the sample to release DNA (with proteinase K) and a subsequent DNA purification step using either organic solvents (mainly phenol-chloroform protocol) or the DNA binding properties of silica (Krings et al. 1997; Rohland and Hofreiter 2007). Moreover, it is usual that before starting the DNA extraction procedure (especially when assessing ancient human samples), the specimen is 'decontaminated' using various techniques, such bleach

surface cleaning or UV light surface irradiation, aimed at removing any surface contaminants. Following decontamination, bones, teeth, and other ancient samples, are powdered using a drilling machine. Grinding has the effect of increasing surface area, and thus maximizing the contact surface from which DNA can be retrieved (Alan Cooper and Poinar 2000; Ramírez et al. 2009).

As mentioned before, data from massive parallel sequencing of DNA from ancient remains have allowed to portrait the nature of the whole spectrum of genetic sequences present in a sample. Key observations from some analysis are that aDNA is present very small fragments (>40-50 nt) (Poinar et al. 2006) -previously inaccessible by PCR approximations- (**Figure 29**), and that there is an inverse relationship between the number of surviving DNA fragments and the length of the sequences (Sawyer et al. 2012). Furthermore, improvements into library preparation methods now permit the recovery of even shorter DNA sequences, by focusing on the sequencing of single stranded molecules, instead of only relying on double stranded fragments (see **sections 2.6 and 3.4**).

A breakthrough regarding the extraction of aDNA was recently attained by (Dabney et al. 2013), when by modifying the chemical conditions of existing silica-based extraction protocols, significantly increased the total yield of aDNA molecules recovered from a sample. By employing this approach, the authors confirm not only that majority of aDNA information seems to be stored in small DNA fragments (less than 40-50 nt long), but that they were also able to sequence the first Middle Pleistocene mtDNA genome, thus further incrementing the temporal range from which aDNA from non-permafrost enviroment can be recovered.

3.3 Breakthrough technologies in aDNA studies

3.3.1 PCR times

For almost two decades the golden approach to recover DNA from ancient remains, was the target-single-locus recovery of genetic sequences by means of PCR amplification followed by conventional Sanger sequencing (Krings et al. 1997).
Briefly, in the PCR, the genomic template DNA is denatured by heating, and then the reaction is cooled to a specific temperature to allow the primers to anneal to their specific target sequence. The temperature is then raised in the extension phase, and a thermostable DNA polymerase, isolated from a thermophilic ("heat-loving") bacterium such as *Thermus aquaticus* (*Taq*), carries out DNA synthesis from the primers, using the genomic DNA as template. The same process is repeated several times, and as it progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified (Mullis Kb 1987).
Although not review here, PCR-based protocols employed in aDNA studies, usually carry out PCR product purification and molecular cloning before sequencing, in order to efficiently recover the amplified DNA products, as well as, to unravel the heterogeneity of the amplicons generated.

Furthermore, until the 21 century all DNA sequencing was carried out by Sanger sequencing, named after its inventor Frederick Sanger, and often also referred as chain-termination, dideoxy, or capillary sequencing (Sanger and Nicklen 1977). Shortly, the method relies upon the incorporation of labeled dideoxynucleotides into a growing daughter DNA strand, which allows the detection of DNA fragments of different size by terminating their replication at one of the four modified nucleotide bases. Initially, detection was assessed by

autoradiography of radiolabeled fragments separated by gel electrophoresis (Sanger and Nicklen 1977). However, the development of fluorescent labels that could be detected by laser excitation, and the replacement of gels by capillary electrophoresis of samples in 96- or 384-well microtiter plates, permitted the necessary automation and increased throughput for the sequencing of larger genomic regions.

Notwithstanding, a combined approximation of PCR-based protocols and Sanger sequencing employed for aDNA studies, had several inconvenient. First, although by amplifying specific regions harboring either functional variants or small contiguous sequences, specific questions about the biology of extinct specimens or their relatedness to their extant counterparts could be investigate (Hoss et al. 1996; A Cooper et al. 1992; Hanni, Laudet, and Stehelin 1994; Lalueza-Fox et al. 2007), such an approach seemed extremely unviable to recover whole DNA molecules, necessary to infer more accurate evolutionary conclusions. Another problematic was that for each independent PCR performed, additional sample material was needed; expending the already scarce DNA present in the sample. A couple of publications (Krause et al. 2006) tried to circumvent both of these shortcomings, by designing multiplex PCR to amplify several mitochondrial regions at the same time. However, given the fragmentary nature of aDNA, that approach implied to carry out several tens of independent multiplex PCR rounds in order to recover i.e. a mtDNA genome, a time-consuming strategy that again will expend lots of precious sample material (Briggs et al. 2009).
Moreover, while applying this technology to ancient non-human samples was only hindered by the inability to efficiently produce enough sequencing data, amplifying DNA from Neandertals and ancient modern human samples was also affected by contamination. For instance, given that a set of primers should be of at least 20 bp of length in order to be highly specific and that aDNA is usually presented in very short fragments, only DNA molecules of at least of 50-70 bp

could be in principle recovered; however preferentially amplifying the longer DNA fragments (**Figure 22**). Moreover, given that recent human contaminating DNA is not degraded and found in higher proportions than endogenous DNA molecules, unless there was a phylogenetic signal within each amplified region (or to the complementary sequence of each binding-primer), there was no way of distinguishing endogenous vs exogenous molecules. Finally, although an ultimate approach would have been to sequence directly all DNA available in a sample, such an approximation was at that time extremely unrealistic, due the high sequencing costs and the limited sequencing throughput that aDNA genomic would have implied.

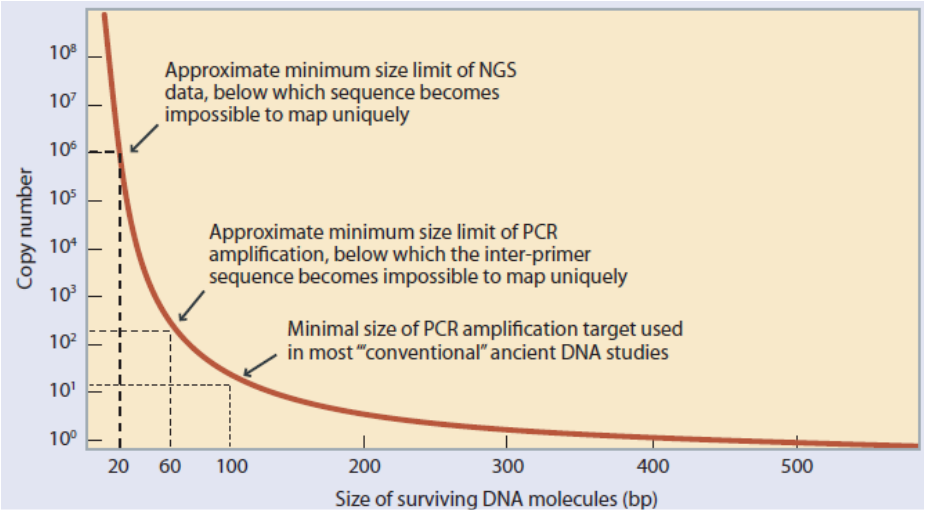


Figure 22: Relationship between size of surviving DNA fragments and copy number in ancient samples [Adapted from Chapter 4 of (Jobling et al. 2014)

3.3.3 Next Generation Sequencing

With Sanger sequencing being regarded as a first-generation technology, these methods became known as NGS methods (sometimes even called second-generation or SGS), and examples of such first NGS machines included the 454 pyrosequecer (renamed to FLX after the acquisition by Roche), the Solexa (acquired by Illumina) and the

SOLiD (life technologies) (Metzker 2010). Moreover, although each of these SGS technology bears unique chemistry and sample preparation procedures, they share common features, like the construction of sequencing libraries, the massive amplification of such libraries and the sequencing-by-synthesis technology. This sequencing technology is based on the fact that tens-to-hundreds of millions of clusters of small ssDNA templates are 'read' simultaneously by sequentially incorporating modified nucleotides to its complementary bases, and emitting a fluorescent signal each time one is added (Metzker 2010). The synthesis process is captured in a series of images that are then analyzed by base-calling algorithms that infer the actual nucleotide composition based on the fluorescence-intensity data for each cluster of DNA templates. Finally, a measure of uncertainty (or quality score) to each base call is assigned and short-reads are constructed from each DNA fragment sequenced, that are then assembled or map into to a reference genome.

Furthermore, following SGS´s arrival, a `third generation` of sequencing (TGS) technologies emerged; platforms in this category included the HeliScope (Helics BioSciences Corporation), the GridIon (Oxford´s Nanopore technologies) and PacBio (Pacific Biosciences), though many more are in active development (Glenn 2011; Schadt, Turner, and Kasarskis 2010). In contrast to SGS, these technologies directly sequence individual DNA molecules without the need to amplification; avoiding in this way, many biases that can potentially emerge in this step (see **section 3.9.1**). Although it´s implantation to sequence aDNA molecules sounds appealing given its likely benefits, it yet remains to be observed how single molecule sequencing technologies can further efficiently and without-error transform the field of paleogenomics (Orlando et al. 2013).

For the purpose of this thesis I will center exclusively in the Illumina sequencing, currently the most widely used platform in aDNA and the NGS technology employed for one of the research chapters I am presenting

3.3.3 Illumina sequencing

The SOLEXA technology, initially introduced to the market in 2006, has evolved into several commercially available high-throughput platforms (i.e. Genome Analyzer IIx, HiSeq series and MySeq) since its initial acquisition by Illumina in 2007. All these sequencing machines rely on the same principles: in situ amplification of adapter-ligated DNA molecules (Fedurco et al. 2006) and the use of fluorescent reversible deoxyribonucleotides terminators (Guo et al. 2008; Bentley et al. 2008). Moreover, after adapters are ligated on both ends of DNA fragments, the strands are attached and immobilized in a flow cell. Bridge amplification is then carried out, generating clusters of few thousands of identical sequences randomly scattered across the flow cell. With the generation of clusters, a fluorescent signal is emitted every time a base is incorporated during DNA synthesis, but in an amplified fashion and more easily detected by its optical device (Bentley et al. 2008).

Specific to Illumina sequencing, nucleotides are incorporated one at a time through repeated cycles of polymerase-directed single base extension. Each nucleotide is labeled with a different removable fluorophore, and after each cycle of incorporation the base is determined by lase-induce excitation of fluorophore and imaging.
Following fluorophore excitation, both the florescent dye and the terminating group are removed and a 3´ hydroxyl group is regenerated allowing the addition of a new nucleotide in the next cycle, until all selected sequencing cycles have been performed (Guo et al. 2008) (see **Figure 23**).
Finally, after the end of all sequencing process, Illumina *de facto* base-calling algorithms transform the fluorescence-intensity data for each cluster of DNA templates at each cycle into a sequence of nucleotides (Metzker 2010).

For the Illumina platform, indel –insertion or deletion mutations- errors are rare, but the overall miscall error rate can be as high as 1%. The main disadvantage of this technology is that towards the final cycles of a run, different copies of DNA templates within the same cluster can become desynchronized and thus the base calling becomes less accurate in later cycles as the extent of asynchrony is exacerbated (Nielsen et al. 2011).

Moreover, in order to accurately identify real nucleotides and avoid sequencing errors, base-calling algorithms produce per-base quality scores by using noise estimates from image analysis. Some sequencing platforms adopt quality values that are defined specifically for the platforms, but those quality values can be easily transformed into the standard Phred[16] quality score, given by $Q\text{Phred} = -10 \log10\ P(\text{error})$ (Metzker 2010). Thus, given this logarithmic scale, a Phred score of 20 corresponds to a 1% error rate in base calling, while for instances a Phred score of 30 will correspond to 1 out of 1000 bases been incorrectly called. Reducing the error rate of base calls and improving the accuracy of the per-base quality score have important implications for assembly, polymorphism detection and downstream population-genomic analyses. As such, several base-calling algorithms have been developed to optimize data acquisition for most of the NGS platforms. For instance, while Ibis (Kircher, Stenzel, and Kelso 2009) or BayesCall (Kao and Song 2011; Kao, Stevens, and Song 2009) have been developed for Illumina, Pyrobayes (Quinlan, Stewart, and Marth 2008) and Rsolid (Wu, Irizarry, and Bravo 2010) , have been developed for 454 and SOLiD respectively (Nielsen et al. 2011).

The length of Illumina reads ultimately depend on the number of cycles of the sequencing run, with one sequence read per cluster, and all reads having the exact same length. In addition, each molecule can be read from both sides by a process called paired-end (PE) sequencing. After a first round of single-end (SE) sequencing, the recently synthesized strand is removed and clusters are regenerated so the complementary

strand can then be read on a second sequencing run. Although PE sequencing is commonly used to map structural variation or assemble genomes, this particular feature of Illumina sequencing is of special interest for the field of aDNA, by only keeping DNA molecules sequenced from both sides and that have a minimum number of nucleotides overlapping sequencing from each read ends, ensuring the quality of the nucleotide selected for downstream analysis.

The throughput of Illumina machines depends on the particular platform of choice, the number of lanes each one has and the sequencing scheme (PE vs Se sequencing). For instance, it ranges from the ~5Gb of a MiSeq PE run ,to the ~600Gb per run of a HiSeq2000 PE one; making it the most cost effective choice when genomic scale projects are considered, and one of the reasons of why it was been so widely apply for aDNA genome-wide studies.

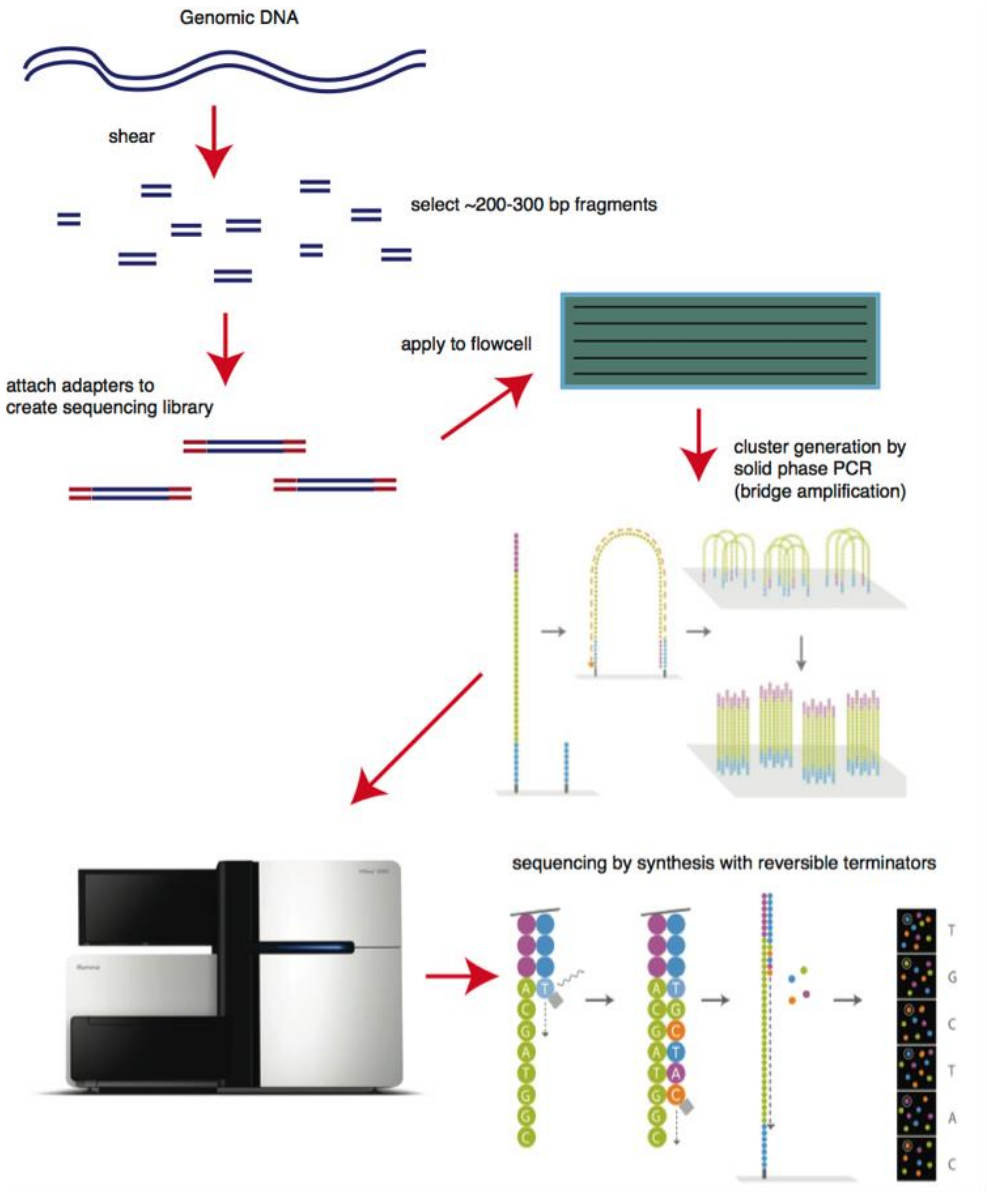


Figure 23: Illumina sequencing flow
[Adapted from http://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/]

3.4 Preparing a NGS DNA Library and its optimizations from the ancient projects

When employing parallel sequencing to recover DNA from ancient remains, a key step to ensure the legitimacy of the aDNA sequences is to make sure that only DNA extracted at the aDNA laboratory is going to be sequenced; without a careful control, sequencing runs could easily be overwhelm with exogenous sequences that would diluted the presence of the already scarce ancient DNA molecules.

Sequencing library preparation protocols vary according to the type of sample assessed, the sequencing platform of election and the objective of the experiment. Nonetheless, the ultimate goal of typical modern DNA library preparation protocols is to ligate adaptors sequences to the ends of each double stranded DNA fragment of a given extract or sample. These adaptors employed are a pair of double stranded oligonucleotides (Adaptors "A" and "B") that not only supply priming sequences to later amplification and nucleotide sequencing, but also provide a unique 4-6 base non palindromic sequencing key (known as index or barcode) used to calibrate the system's software, for base calling and to precisely identify legitimate library reads. By employing such an experimental procedure, an immortalize group of sequences representing a certain condition or a sample can be generated, and normally referred, as "*library constructs*". Moreover, within the field of aDNA and until recently, such technological approach have been the *defacto* approach to prepare sequencing libraries (see below); usually referred as *double stranded* library preparation protocol (dsDNA-prot).

Notwithstanding, as mentioned before, every aDNA sequencing project start with exploratory sequencing runs to screen several library constructs, in order to assess the preservation state of the endogenous DNA in a sample. In addition, given that probably other NGS projects will be carry out at the sequencing facilities, it becomes necessary to pool multiple samples in a single lane, in order to efficiently exploit all

the machine´s sequencing throughput. In case of paleogenomic projects, different indexing and barcoding protocols specific for aDNA sequencing libraries have been developed (Knapp, Stiller, and Meyer 2012). Another important technological innovation of aDNA sequencing library preparation protocols have been the possibility to experimentally remove *post mortem* damage in order to avoid biasing downstream analysis (describe in detail in **sections 3.10-3.11**) (Briggs et al. 2010). What is more, improvements to library preparation methods now permit the recovery of even shorter genetic sequences and far more endogenous genetic material, by building sequencing libraries based on single stranded molecules; this library preparation procedure is technically referred within the aDNA field as single-stranded library preparation protocol (ssDNA-prot) (Meyer et al. 2012; Gansauge and Meyer 2013) (**section 2.6**).

3.5 Capture-enriching methods

Despite the significant improvement NGS represented to the field of ancient genomics, the restrictions posed by the low proportion of endogenous DNA found in most non-permafrost samples still challenge the accessibility to DNA from most ancient samples (Stoneking and Krause 2011).

Although initially conceived for sequencing functional segments of the genome of modern samples (i.e. exomes), target enrichment methods (capture-enrichment methods), circumvent such limitations by utilizing user-defined targets to "capture" by hybridization and enriched specific regions, prior to sequencing. Thus, by employing this technological innovation, not only the amount of environmental contaminating DNA can be limited, but also a high throughput sequencing of ancient samples becomes economically feasible (Burbano et al. 2010; Castellano et al. 2014; Schuenemann et al. 2011).

Capture probes representing the regions of interest can be either immobilized on a solid surface (on-array capture) (Hodges et al. 2007) or biotagged and selectively retrieved using streptavidin-coated magnetic beads (in-solution capture) (Gnirke et al. 2009). Furthermore, two main types of capture-enrichment methods are gaining popularity within aDNA investigation: either large-scale capture using commercial methods (in-solution or on-array), or non-commercial methods represented by smaller scale in-solution approximations, such as primer extension capture (PEC) (Briggs et al. 2009) and its modified version Capture On Beads (COBs) (Maricic, Whitten, and Pääbo 2010) (**see Figure 24**).

Some examples of higher-scale target enrichment aDNA projects -which have relied mainly on technology manufactured by Agilent-, have been directed to either to target ~14,000 protein-coding positions assumed to have changed on the human lineage since the last common ancestor shared with chimpanzees (Burbano et al. 2010), to recover the chromosome 21 from an early modern human from the China (Fu et al. 2013) or to capture and sequence the exome from two template poorly preserved Neandertals samples (Castellano et al. 2014) (see **Chapter 2**).

Furthermore, both the PEC and the COB method rely on either synthetizing an oligonucleotide primer with a biotinyl group attached to its 5´end (which is very expensive) or link the biotinyl group *a posteriori* to a long range fragmented PCR amplicons, as baits, instead of manufacturing the probes. Additionally, even though both of these approaches are an efficient alternative to commercial target enrichment aDNA methods, either because of their prohibited cost or its underlying technological basis, both PEC and COB are limited to mitochondria or small DNA molecules (**see Figure 24**).

In recent times, a whole genome capture method (WISC) that uses home-made biotynilated RNA probes as bait (which significantly reduces the cost of probe design) has been developed (Carpenter et al. 2013). While this approach sounds attractive, currently it seems to

introduce a bias against shorter DNA molecules, as well as an enrichment of probes mapping within repetitive regions; both of these issues will have to be addressed before it can be fruitfully applied to samples with highly degraded DNA (Avila Arcos *et al* personal communication).

Finally, while capture-enrichment technologies (unless an extremely well-preserved sample is found) will undoubtedly become the gold-standard approach to recover aDNA from more template environments, technological improvements to economically retrieved as much endogenous DNA as possible, while avoiding capturing repetitive sequences or longer contaminating DNA fragments, need to develop (M. Avila-Arcos et al. 2014).

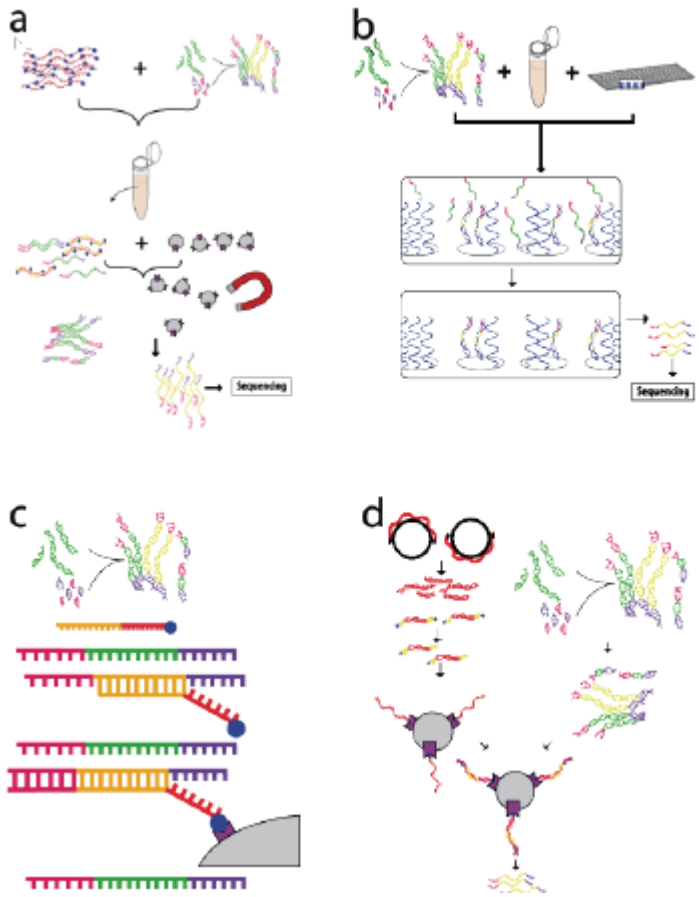


Figure 24: Target enrichment methods.
All methods use a sequencing library as starting material and predefine baits or capture probes targeting the regions of interest (ROI). Targeted regions are represented by yellow insert and background molecules by green insets a) **In-solution targeted enrichment.** Sequencing library is incubated with biotynilated RNA baits (in red) targeting the ROI. Hybridized DNA is retrieved by streptavidin coated magnetic beads. RNA is degraded and the targeted enriched library is ready for library amplification and sequencing b**) On-array target enrichment.** Library molecules are hybridized with probes (in blue) attached to a solid surface targeting the ROI. Unbound DNA is washed off and library molecules are eluted with heat c) **Primer-extension capture (PEC).** Biotynilated DNA primers hybridize with the ROI in the library molecules. Primer bound molecules are immobilized in streptavidin-coated magnetic beads and an extension reaction is carried out. Unattached molecules are washed off and targeted molecules are eluted d) COBs approach relies on construction of `baits` as part of the protocol. Amplified DNA from the ROI (mitochondrial genome in this case) by long range PCRs is converted into baits (in red) by fragmentation and ligation to biotynilated adapters. Baits are attached to streptavidin-coated beads. Library molecules are incubated with bead-bound baits, background are washed off and hybridized molecules are recovered and eluted. [Adapted from (Ávila-Arcos 2012) reproduced with permission]

3.6 NGS data mapping, processing and filtering

Along with the development of NGS technologies, the need for computational and algorithmic resources increased drastically in order to extract reliable and meaningful information from sequencing data. As a result of these demands, several tools have been developed to handle the data at every stage: from its generation and manipulation, to its analysis, interpretation and visualization. **Figure 25** represents the steps needed for the analysis of NGS data from single genome, which are linked together in a bioinformatic pipeline; the final steps involve a functional validation of the predicted genetic variants. Given the particularities of aDNA and depending on the quality of the data, some parts of this pipeline have been modified or skipped and even new software specific for aDNA characteristics has been developed (Briggs et al. 2009; P. Skoglund et al. 2014).

The focus of the second part of **section 3** is on the manipulation of the data, once it has been generated by basecalling algorithms, roughly from removing adaptors and uniquely map reads (**sections 3.7-3.8**), to variant calling and obtaining contamination estimates (**section 3.9-3.12**).

Once sequence reads are generated, downstream analysis will depend on the platform used, the amount of data generated and if the mapping will be *de novo* (no reference sequence to align against exists) or is it a "resequencig" experiment (there is a sequence to which reads can be aligned).

All the aDNA projects carried out during this thesis have relied on mapping reads to a known reference; hence *de novo* mapping strategies are not discussed.

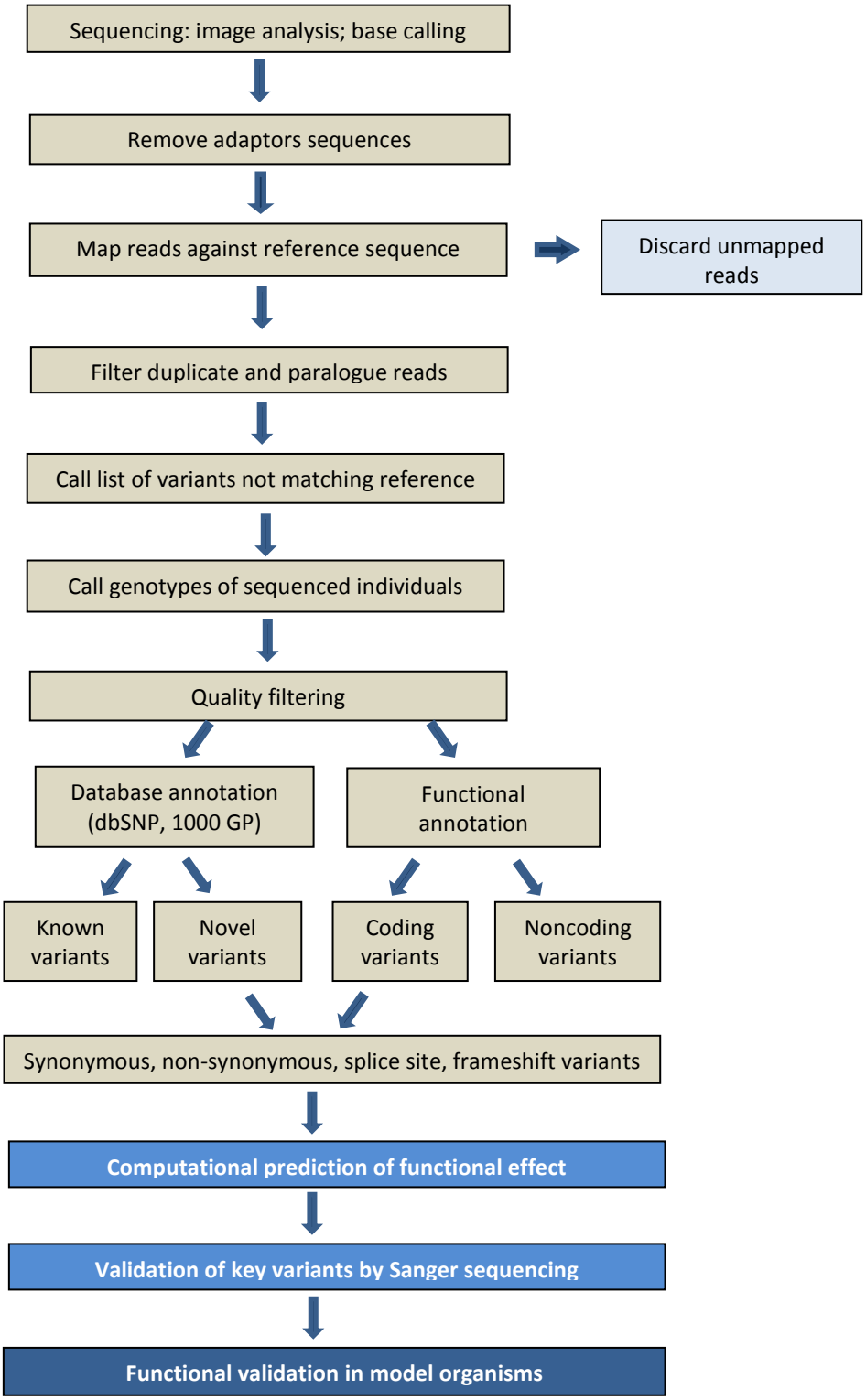


Figure 25: Steps in the generation an analysis of NGS data from base calling to a functional validation of predicted variants. [Adapted from (Nielsen et al. 2011)]

3.7 Adapter removal

Regardless of the platform, once sequence reads are generated, there are several steps in a standard analysis pipeline that need to be carried out before comparing them to a reference. The first of such steps, include the sorting of reads by the index or barcode previously described and the removal of adapter sequences. The former can be incorporated into standard Illumina´s basecalling protocols (Real Time Analysis `RTA` + CASAVA), and the latter can by either directly implemented by user or carried out recently publish software, *AdapterRemoval* (Lindgreen 2012).

Removing adapter sequences is a crucial step when dealing with aDNA data. Given its fragmented nature, the number of cycles of the sequencing run usually exceeds the length of aDNA insert molecules (short molecules), resulting in reads carrying adapter sequence at its 3´end. For instance while the average read length from a typical Neandertal samples is 40-50 base pairs (bp) and the average size of sequences from permafrost samples is ~60-70 bp, standard SE HiSeq2000 runs currently involve 100 cycles, thus 100 nt.

As consequence, aDNA reads need to be "cleaned" from adapters before downstream mapping and aligning. Not carrying out this step, will cause that many reads most likely will not match a reference sequence due to misalignment of the adapter region. In addition, adapter-dimers molecules should be eliminated from downstream analysis. Several programs are currently available for adapter trimming and adapter dimer removal (Falgueras et al. 2010; Schmieder and Edwards 2011). Among these, *AdapterRemoval* recently developed by (Lindgreen 2012), is a software that apart from trimming adapter sequences, it also removes adapter-dimers as well as low quality reads and trims low quality stretches and *Ns* at the ends of the reads. Other adjustable parameters, include the threshold for minimum length after trimming and number of allowed mismatches, just to mention a few (Lindgreen 2012).

3.8 Read mapping

As soon as adapter sequences have been trimmed, sequencing data is ready to be analyzed either by mapping reads to a reference, *de novo* assembly into contigs, or a hybrid strategy. The method of choice will depend on the existence or not of a reference sequence that belongs to the same species from which DNA was extracted, or to a species that has a close evolutionary relationship. For instance in the absence of a Neandertal reference genome, reads from this hominin have been mapped to the human or chimpanzee genomes (Green et al. 2010).

The accuracy of the alignment has a crucial role in genetic variant detection. Incorrectly aligned reads may lead to false positive variant detection (variant calling error), so it is important for alignment algorithms to be able to cope with sequencing errors, potentially real genetic site variation, and in the case of aDNA samples, with deamination modifications.

Several strategies for mapping NGS reads to references have been developed in recent years. Most alignment algorithms for NGS data are based on either 'hashing' or an effective data compression algorithm called the 'Burrows–Wheeler transform' (BWT) (Burrows and Wheeler 1994). BWT-based aligners (for example, Bowtie (Langmead et al. 2009), SOAP2 (R. Li, Li, et al. 2009) and BWA (H. Li and Durbin 2009) are fast, memory-efficient and particularly useful for aligning to a certain extent repetitive reads; however, they tend to be less sensitive than the state-of-the-art hash-based algorithms (for example, MAQ (H. Li, Ruan, and Durbin 2008), Novoalign and Stampy (Lunter and Goodson 2011)). The Novoalign and Stampy aligners currently produce the most accurate overall results, while also being practical in terms of running time (please review (Fonseca et al. 2012; Trapnell and Salzberg 2009; H. Li and Homer 2010) for a detailed comparison of the performance of various aligners).

Currently, only one software has been develop (Briggs et al. 2009) which takes into account aDNA *post mortem* modifications in order to

map sequencing reads; however, its use has been limited to only a few studies (Briggs et al. 2009; Green et al. 2010). As an alternative, most paleogenomic projects have opted for strategies that optimize the use of short reads mappers, in order to take into account aDNA peculiarities (Schubert et al. 2012). Moreover, presently BWA is the mapper of choice not only for aDNA sequencing, but also for many modern NGS projects as well.

Furthermore, depending on the assumed evolutionary distance between the sample and the reference being employed (but also the level of genetic diversity for the organism under study) several mismatches and gap openings can be allowed, which can help optimizing the number of hits for aDNA data. For instance, given the short length of typical aDNA reads, a typical fine-tuning trick when mapping with BWA is to deactivate the minimum seed length requisite (Schubert et al. 2012); by employing this approach, the whole read is aligned as a whole to the reference genome, instead of using sub-oligomers of the read which could be map anywhere. As the aligning process finishes (in the case of BWA), an output file in the SAM (Sequence Alignment/Map) format (or its binary version BAM) is generated (H. Li et al. 2009). SAM is the standard format for reporting short read alignments to references genomes; as a consequence is the *de facto* input format for many downstream analysis tools.

A key step of aligning NGS data is to produce well-calibrated alignments (or mapping quality values), as variant calls as well as other downstream analysis depend on those scores. *Mapping quality values* are a measurement of the confidence that a read actually comes from the position where it was originally aligned by the mapping software, as well as other parameters such as base quality and the number of mismatches (H. Li et al. 2009; Nielsen et al. 2011). Consequently, as any confidence estimate, it is associated to a probability that the read was incorrectly mapped. Therefore, mapping quality values are normally reported in a logarithmic scale by a number associated to the

probability a given read was incorrectly aligned (usually refer as MAPQ or MAPping Quality Phred-scaled scores) (Nielsen et al. 2011; H. Li et al. 2009). Thus, as mentioned before for the Phred base qualities, the higher the Phred score is the lower the chances that a read is incorrectly mapped. Moreover, filtering to retain only the most likely correctly aligned reads can be attained by a combinatory strategy of *in-house* scripting on SAM files and/or using tools design for such purposes such as Samtools (H. Li et al. 2009), GATK or PICARD tools. As a rule of thumb, a minimum MAPQ score of 25 is usually recommended as trade-off between the reducing the bias of erroneously align reads for downstream analysis and recovering as much data as possible.

Finally, as mentioned before one of the crucial steps throughout the undertaking of a paleogenomic projects, is to estimate the DNA preservation in sample in order to select the best experimental methodology to recover as much genetic material as possible. One of the most common approximations to calculate DNA preservation is to estimate the **sample sequencing efficiency** (Poinar et al. 2006; Green et al. 2010; Olalde, Allentoft, et al. 2014). Sequencing efficiency is usually determined by the mere comparison of the total number of reads mapped to the reference out of the total reads in the library. However, although is widely used as a proxy to determinate how much endogenous material stemming from the organism of interest is present, this estimation still includes PCR read duplicates as well as unspecific map reads which could potentially bias such preservation estimates. The next section describes in detail the relevance of removing PCR clones and paralogue sequences from the analysis.

3.9 Removing duplicates and paralogue: going to the unique.

3.9.1 Duplicate removal

NGS libraries usually require an amplification step with few cycles (~8 to 12) in order to reach the amounts of DNA specified by manufacturers. For aDNA, however, such numbers of cycles often need to be increased due to the little endogenous DNA content. Increased number of cycles may lead to an excess of clonal molecules in the sequencing reads, biasing depth coverage values and affecting polymorphism identification, damage estimation, as well as the enrichment quantifications in capture experiments. Therefore, it is critical to remove such duplicates at the library level before carrying on downstream analysis. The most straightforward approach for eliminating such clonal artifacts is by "collapsing" all reads on the same strand with identical external coordinates into a single sequence. Such *clonal collapse* can be performed with standard analysis tools such as *rmpdup* tool of the Samtools or the Mark Duplicates routine in Picard package; both tools identify PCR clones and retain the read with the highest mapping quality. Levels of clonality vary vastly across experiments and depend mostly on the amount of endogenous DNA and the number of cycles performed. Capture experiments are particularly susceptible to high clonal duplication levels (due to the nature of their protocol); therefore caution is advice when capture enrichment levels are reported without performing the clonal removal steps (M. C. Avila-Arcos et al. 2011).

3.9.2 Paralogue removal

A common characteristic of large genomes is the high content of repetitive elements (e.g. maize: ~80%, human: ~50% (Lander et al. 2001), making them difficult to be assemble with short fragments such

as those generated by Illumina platforms. Even though PE reads with different insert sizes can alleviate to some extent this limitation, it becomes more difficult for ancient samples, where DNA molecules are highly fragmented.

In consequence reads from aDNA experiments within repetitive regions are difficult to assign unambiguously to unique genomic coordinates. Therefore it is recommended to exclude such ambiguous hits from downstream analyses since it is impossible to confidently assess their exact genomic location. Reads within repetitive regions can be removed from BAM/SAM files by setting a filter for mapping quality.

In addition, reads with multiple hits can also be further removed by controlling for particular tags in the BAM files. Tags are optional pieces of information of the SAM format that depend on the mapper used to align the reads. In particular, BWA includes thee tags that are informative of the "repetitiveness" of the read: XT, XA and X0 (H. Li et al. 2009). When a read aligns equally well to more than one position in the genome, BWA randomly picks one to report in the coordinate field, however, it keeps track of such finding by reporting it in the XA,XT and X0 tags (H. Li and Durbin 2009).

Therefore at this point, one can unbiasedly estimate the amount of unique DNA molecules preserved in a sample, by calculating the proportion of DNA endogenous sequences in a *library construct*. Although there is no consensus on how to report DNA conservancy, estimates based on the percentage of endogenous unique DNA sequences are always preferred.

Finally, it is worth mentioning that processing and analyzing aDNA NGS data is become more user friendly, especially for those researchers without a bioinformatic background. Recently (Schubert et al. 2014) have publish a pipeline –called PALEOMIX- design to provide a full bioinformatic service from removing adapters and mapping, to calling variants and genotypes or estimating DNA preservation and contamination, while taking into account aDNA

special characteristics; especially recommended for those without a bioinformatics background.

3.10 Postmortem fragmentation patterns from NGS data

Probably the most valuable insight obtained from analyzing NGS aDNA data (Briggs et al. 2007; Poinar et al. 2006), has been an almost full understanding of the misincorporation patterns occurring in aDNA. As mentioned before, one of the hallmarks of aDNA noticed already from the very early studies, was that some of it nucleotides presented chemical modifications, particularly cytosine deaminations (Hofreiter et al. 2001). Results from the first NGS aDNA projects allowed a further understanding of the nature of deamination events; such *post mortem* damages were observed more frequently towards the last 10 nucleotides at the edge of the sequencing reads in the form of C to T and G to A substitutions at the 5' and 3' ends, respectively (Brotherton et al. 2007; Briggs et al. 2007). Later on, it became evident that the observed increase of G → A changes at the 3´ end, didn't represented a real genetic change of G into A, but rather a technical artifact generated when filling-in 5' the overhanging ends on the complementary strand, during the double stranded library protocol preparation. Therefore, such changes represent real C to T substitutions on the complementary 5-ends of the original template molecule (Briggs et al. 2007) (see **Figure 26)**.

Furthermore, a recent sequencing library technological breakthrough (Meyer et al. 2012) has permitted to characterize the genetic composition of single stranded aDNA fragments, and with that confirm that cytosine deaminations indeed occur more frequently with last 10 nt at each edge of the sequencing reads in ancient remains (Briggs et al. 2007) (see **Figure 28**) (see **section 3.4**).

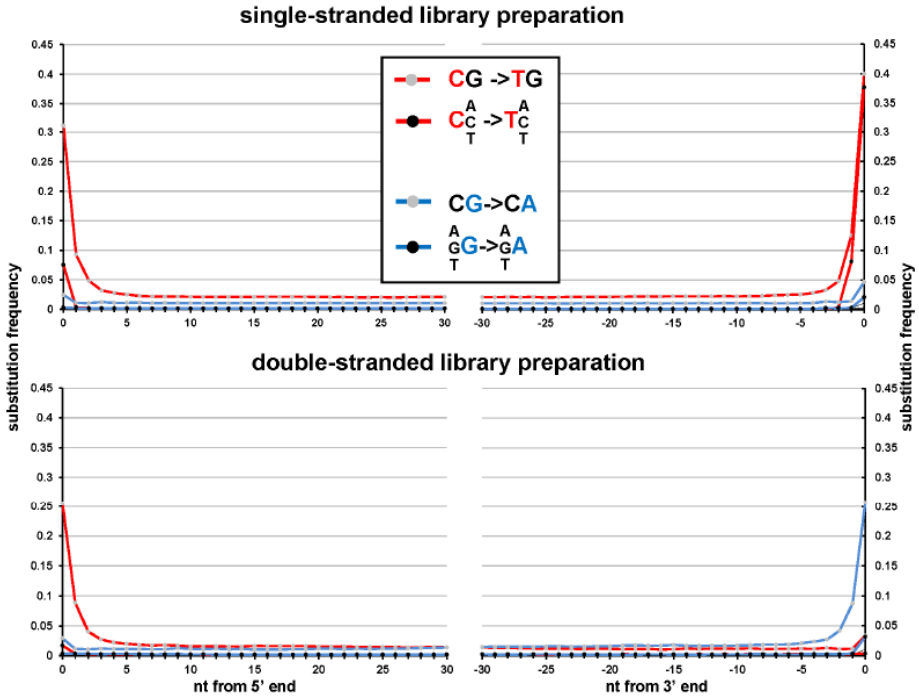


Figure 26: Misincorporation patterns observed from aDNA data.

Position dependence of C->T and G->A substitution frequencies at 5'- and 3'-ends of Denisovan

Additionally, as mentioned before, the amount of *post mortem* damage and its occurrence as a function of distance from the edges of the sequencing reads strongly depend on the preservation conditions where the sample was recovered (Sawyer et al. 2012). In addition, also with the advent of NGS an increased knowledge of the nature of the immediate nucleotides from the reference genome up and downstream of the coordinates where reads map has been attained. Such analyses have led to comprehend that the rate of purines is high in the bases immediate after the read edges into the reference; an observation that agrees with previous findings that depurination is probably related to fragmentation and deamination chemical modifications (Briggs et al. 2007).

3.11 Variant calling, accounting for DNA damage and data authentication

In-resequencig efforts, the most informative analysis arise from detecting single nucleotide variants (SNVs) either for population history or medical and phenotypic inferences. For instance, depending on the nature of the genomic project, some NGS strategies couple the accurate detection variants and genotypes with elucidating their putative functional relevance or its possible phenotypic consequences. Either way, accurately detecting genetic changes and filtering less likely variants are key steps within NGS projects to ensure the veracity of the biological implications of the data (Nielsen et al. 2011). **Figure 26** illustrates key steps from registering base calling of NGS to assessing the putative function of genetic variants.

Once a reliable SAM/BAM file is produced and only uniquely mapped reads have been retained, genetic variants and genotypes can be called. Moreover, obtaining well-calibrated base quality scores is important, as variant and genotype calling at a specific position in the genome depend on both the base calls and the per-base quality scores of the reads overlapping the position. Therefore, base quality recalibration and local realignment is of great importance to warrant the quality of the data produced (McKenna et al. 2010; Nielsen et al. 2011).

For instances, while in SOAPsnp (R. Li, Yu, et al. 2009; R. Li, Li, et al. 2009), per-base quality scores are recalibrated by comparing a sequenced genome to the reference genome at sites with no known SNPs, a related alignment-based recalibration algorithm has been implemented in the GATK software (McKenna et al. 2010) which also takes into account several covariates such as machine cycle and dinucleotide context. The rationale behind this approximation is to reassess the quality of call bases, given their raw quality score, the position within the read, the genomic context and the sequence batch

with which it was sequence; any base call difference to the previous nucleotide prediction is pondered by these factors.

Furthermore, indels (especially with low or moderate coverage) might cause some reads not to be accurately map to the reference genome given local structural variation; this scenario is usually surpassed by realigning locally all reads that map close to known polymorphic indels, minimizing mismatches to the reference genome and thus erroneous calls. The process of converting base calls and quality scores into a set of genotypes for each individual in a sample is often divided into two steps: variant calling (also known as SNP calling) and genotype calling; here the terminology '*calling*' is used to make reference to the estimation of one specific genotype or genetic variant. While variant calling aims to determine in which positions there is a true variant, a polymorphism or at which positions at least one of the bases differs from a reference sequence, genotype calling is the process of determining the genotype for each individual (Nielsen et al. 2011).

Moreover, genotype and variant calling assessments can be carried out as easily as by counting alleles at each site and using simple cutoff rules for when to call a variant or genotype/employing algorithmic methods which use a probabilistic framework. In this later approximation, so-called 'genotype likelihoods' — which incorporate errors that may have been introduced in base calling, alignment and assembly — are coupled with prior information, such as allele frequencies and patterns of LD.. The result is a variant and genotype call and an associated measure of uncertainty (which is often described by a 'quality score'), both of which have a concrete statistical interpretation (Nielsen et al. 2011). Nonetheless, as with all assessments related to NGS data, the higher the sequence depth the higher the chances that the variant and genotype predictions will be accurate. Examples of software commonly used to call variant and genotypes from NGS data,

are Samtools or GATK, many among others. For an extended review on SNP and genotype calling please referrer to (Nielsen et al. 2011).

Even though a better understanding of the nature of aDNA molecules and its importance when aligning them to a reference genome it is now attained, it is unfortunate that no a software has been developed to take into account *post mortem* damage as a prior in order to accurately make variant or genotype calls. Therefore, since the availability of NGS data, the aDNA community has implemented different alternative procedures in order to warrant that *post mortem* damage will not bias calling variants.

A commonly wide-used strategy - based on the known features of aDNA- identifies all putative damaged bases within their expected range of occurrence (i.e. a C -> T change at the 5´ or a G -> A change at the 3´, when using the lib-dsDNA-prot i.e.), and then lower the quality of such variants from their corresponding positions in a SAM files (Sánchez-Quinto et al. 2012; Pontus Skoglund et al. 2012); by having such low base quality *prior*, those positions will be filtered from downstream analysis. Nonetheless, although this approximation sounds effective, a problem arises when removing all possible deamination events, since some of these changes might actually represent real polymorphism, and thus their removal might cause some underestimates of genetic diversity or divergence. Alternatively some research groups have also opted to reduce the quality of all thymine or all thymines and guanines within 5-10 nt from the edge of the read ends (Castellano et al. 2014). However, this action has the opposite effect, because while is efficient to avoid biasing downstream analysis, with this approximation several true genetic variants can be lost and unfortunately a fair amount of the data is wasted.

A couple of programs have been generated, which largely model the probability that a certain genetic variant occurred due to DNA damage or not (Jónsson et al. 2013; P. Skoglund et al. 2014). Thus, given the location of a genetic substitution within the edge of a read and the

known patterns of genetic diversity for that specie, these programs build a statistical model taking those parameters (among others) into account to estimate the likelihood that a genetic changes at a given position along a read is due a deamination modification.

Yet another strategy to remove misincorporation events, relies on experimentally removing all uracils found within aDNA sequences (Gansauge and Meyer 2013; Meyer et al. 2012). Given that in fact, cytosines are deaminated into uracils (although read as thymines by sequencing machines), by removing all uracils and incorporating the correct nucleotide i.e. cytosine, the future influence of *post mortem* modifications is prevented. Unfortunately, even though removing uracils also sounds as an attractive measurement to avoid calling false positive variants, it is only useful when sufficient coverage is obtained and human contamination estimates are negligible for aDNA projects. For instance, if contamination is high or coverage is too low, and then *post mortem* patters are removed, there will be no way of distinguishing true endogenous sequences from contaminant sequences (P. Skoglund et al. 2014)(see below).

Once putative deamination modifications have been taken into account and the possibility that they will influence downstream analysis is minimized, then the next steps will exclusively depend on the depth and coverage of the data (in addition to its quality) in order to accurately call variants/genotypes; a process for which any of the previous mentioned software/strategies for analyzing modern DNA can be employed.

On the other hand, given the reduced percentages of endogenous DNA from most ancient samples, most paleogenomic projects are usually characterized by low average coverage and reduced amounts of data; genotype will not be able to be called for most of the data and most positions only one read will cover each position. Therefore for such aDNA genomic projects the aDNA community has developed specific approaches to accurately identify genetic variation. In these circumstances, a general approximation is to couple the action of

lowering the base qualities of changes suspected of being due to deaminations and randomly select bases that pass stringent base quality filers for each covered position when possible (Sánchez-Quinto et al. 2012; Pontus Skoglund et al. 2012). Thus a "consensus sequence" will be created wherever data is available.

Lastly, although there is still room for improvement in order to maximize the recovery of aDNA data and its quality by appropriately modeling *post mortem* damage when calling variants, currently, the limiting step to aDNA paleogenomic projects will continue to be to obtain as much depth of coverage as possible from DNA extracted from ancient remains.

3.12 Data authenticity

As it has been reviewed in **sections 2.1, 3.2 and 3.10**, ancient DNA is present in very small amounts, its seriously damaged, and is difficult to recover; in contrast, modern DNA is plentiful, ubiquitous, in relatively good condition, and therefore very easy to amplify. Therefore, contamination can occur either while handling a specimen after its excavation, at the ancient aDNA laboratory while performing experiments to extract DNA, and even it can be introduced while sequencing at the sequence facility.

One precaution to avoid major sources of contamination is to process only reads which have the short project-specific bar codes (index) corresponding to the project being undertaken (Knapp, Stiller, and Meyer 2012) (see **section 3.4**); thus, restricting the analyses to sequences which seem to have originated exclusively at the clean room and limiting the possible sources of contamination.

As mentioned before, contaminating DNA can be either environmental (microbial or fungi DNA) or human contaminating DNA; this later types is also known as laboratory contaminating DNA, since it is introduce at aDNA laboratories. While the amount of environmental DNA can be control i.e. by employing capture-hybridizing techniques

(among other approaches (Green et al. 2010)), modern human DNA contamination is almost inevitable, and thus methods to elucidate the authenticity of an ancient sample are of paramount importance (Jónsson et al. 2013; P. Skoglund et al. 2014).

There are three main procedures that can be employed to bioinformaticaly evaluate the fidelity of sequence data produced, and which can be divided in: direct and indirect measurements of estimating contamination. While direct approaches are based on inferring contamination directly from *diagnostic positions* within the sequence data, indirect estimates rely on analyzing if the sequence data present patterns typical of aDNA samples. The following sections briefly detail how to estimate the amount of human contaminating DNA in aDNA samples from each of the different approaches.

3.12.1 Direct estimates

3.12.1.1 Sequence contamination estimates

Assuming that the major source of laboratory contamination arises from modern human DNA, estimating contamination will depend invariably on the type of sample being analyzed. If samples are known *a priori* to be divergent to modern humans (i.e. there is a *phylogenetic signal*), more straightforward approaches to elucidate the authenticity of ancient remains (Shapiro and Hofreiter 2014) can be undertaken. On the other hand, if samples genetic variation modern humans are been surveyed, other approaches can be undertaken.

On the one hand, regarding samples know to fall outside the genetic variation of modern humans, i.e. Neandertals (see **section 2.2**), given that there are base substitutions observed in most or all Neandertal mtDNAs that are not seen in present-day human, and vice versa (see **Figure 27**) (Krings et al. 1997; Green et al. 2008), by having sufficient independent reads at such *diagnostic positions*, an average ratio of endogenous vs contaminant sequences along the mtDNA could be

elucidated and an approximate estimator of the percent of human contaminating DNA be estimated (see **Figure 27**) (Green et al. 2008). However, although mtDNA sequences can yield reliable estimates of contamination, the level of nuclear DNA contamination can be under- or over-estimated if the contaminating DNA source contains less or more mtDNA, respectively, than the endogenous DNA (Green et al. 2009). Therefore, if working with nuclear data, estimates of nuclear contamination as also recommended. Fortunately, current large genomic projects such as 1000 genomes have largely characterized the genetic variation in modern humans, and have provided information about sites with fixed or nearly fixed mutations that can also be surveyed. When working with Neandertals, *diagnostic positions* with autosomic and sexual chromosomes are worth being addressed, in order to obtained unbiased contamination estimates.

On the other hand, other approaches requiring no *a priori* information on diagnostic positions have also been developed; some of them rely for instance on the concordance of the morphological determined gender with genetic data of a given ancient sample. For example when morphological evidence suggests that any ancient remains belong to a female, sequences mapping to the Y chromosome can be surveyed. Thus, by focusing in the genome-wide unique regions of the Y chromosome, and comparing the number of such sequences to the total number of sequences that you would expect to map to the Y chromosome if the individual was a male, it is possible to obtain estimate male contamination from female bones (Green et al. 2009). The opposite is also possible when the remains are suspected to belong to a male individual. In the case of male, as males are haploid for the X chromosome, heterozygosity should not be observed in overlapping reads mapped to X chromosome (Green et al. 2009). Therefore, by restricting to unique regions from the X-chromosome and with a minimum coverage threshold (to avoid bias by sequencing errors), estimates of contamination can be calculated by dividing the number of heterozygous positions with the total number of such positions (Olalde,

Allentoft, et al. 2014). Notwithstanding, for any contamination estimates based on sexual chromosomes estimates, special consideration should be placed towards avoiding regions of high similarity between X and Y chromosomes, as well as carefully filtering any misaligned sequence or repetitive regions.

Finally, although previously unconceivable, *phylogenetic signal*-like analyses can also be applied for ancient modern human samples analyses. Briefly, by having enough depth-coverage at positions defining main mitochondrial or Y modern human lineages, such "*diagnostic site"*-like positions can be examine for reads not matching the inferred haplotype for the sample, and thus be a putative source of contamination (Olalde, Allentoft, et al. 2014).

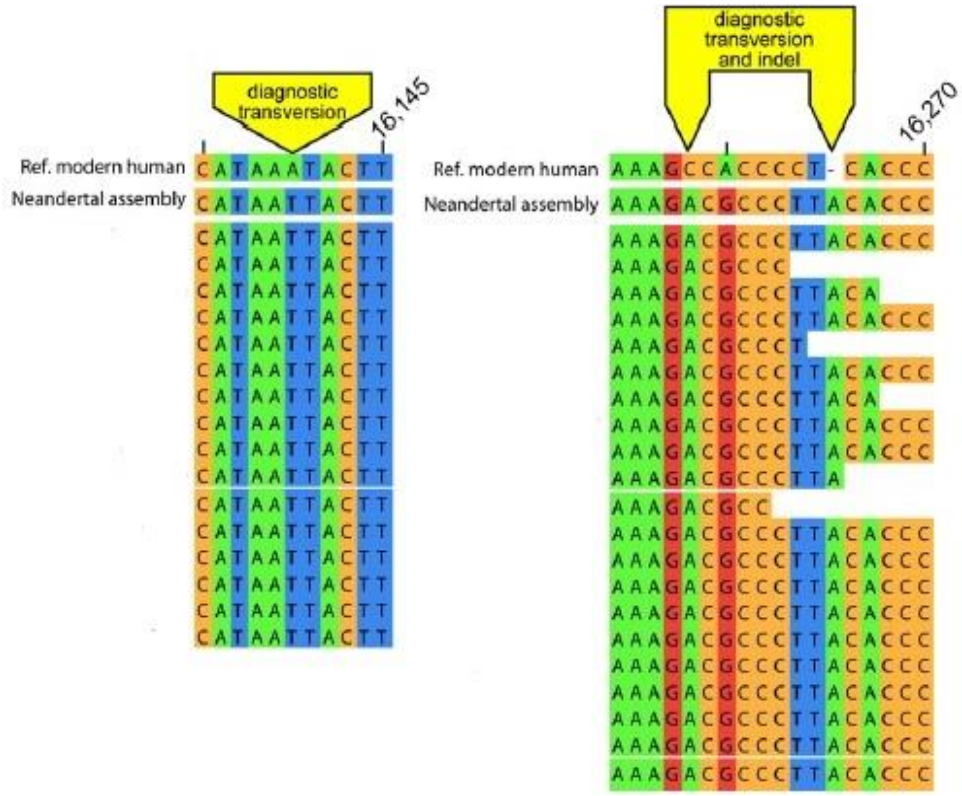


Figure 27: Estimating contamination from modern human-Neandertal diagnostic positions.

On the image above each line represents an independent read align either to the modern human or the Neandertal assembles at positions from the mtDNA where all or nearly all (>99%) modern humans different from Neandertals. The authenticity of a sample can be assessed by calculating the average ratio of reads equal to the Neandertal or the modern human mtDNA genome. [Adapted from (Green et al. 2008)]

3.12.2 Indirect estimates

3.12.1 Misincorporation pattern plots

The introduction of massive parallel sequencing has allowed the characterization of the whole landscape of DNA molecules present in ancient remains. By carefully identifying endogenous DNA fragments and thoroughly analyzing the genetic substitutions most frequent to each position, a more complete understanding of the nature of aDNA molecules and *post mortem* damage is now attained (Briggs et al. 2007). For instance, it is now understood that misincorporation pattern seem to increase with time (Sawyer et al. 2012), and thus a common practice to authenticate aDNA NGS data, has been to correlate the frequency of C to T deaminations at the edge of the sequencing read ends (or of both C to T and G to A substitutions depending on the experimental procedure) with the putative archeological age of the sample. The empirical rule of thumb is that the older the sample, the higher the frequency of the deamination events should be at the edge of the reads end (Sawyer et al. 2012) (see **Figure 28**).

However, *post mortem* DNA modifications depend on site specific environmental conditions, and thus DNA extracted from samples coming from specific latitudinal locations or climatic conditions might bear misincorporation patterns that are similar to those from samples recovered from sites with better "preserving environmental conditions". For instance, while a sample ~500 years old from the Caribbean might have recorded the same amount of cytosine deaminations than a sample 8,000 years old from a more template location within Europe, a sample from the permafrost in Greenland from ~5,000 years ago, might have as much damage as a hair sample from ~200 years ago kept in a Museum. Thus, although useful, this approximation to authenticate the origin of DNA recovered from an ancient sample should be put into context and be used more as additional evidence, than a formal way to estimate contamination. Furthermore, mapDamage 2.0 pipeline (Jónsson et al.

2013) has a special feature with which one can easily obtain the misincorporation patterns plots from aDNA NGS data. MapDamage 2.0 is highly recommended given that it generates statistical models that take into account several parameters in order to calculate the likelihood that an observed deamination is actually true.

Finally, recently an *in-silico* approach has been developed in order to separate endogenous DNA sequences from putative contamination in aDNA studies (P. Skoglund et al. 2014). This strategy, has allowed the identification of authentic Neandertal sequences from historic samples, and it opens the possibility of analyzing Neandertal remains that were previously discarded for genetic studies due to their high level of human contamination.

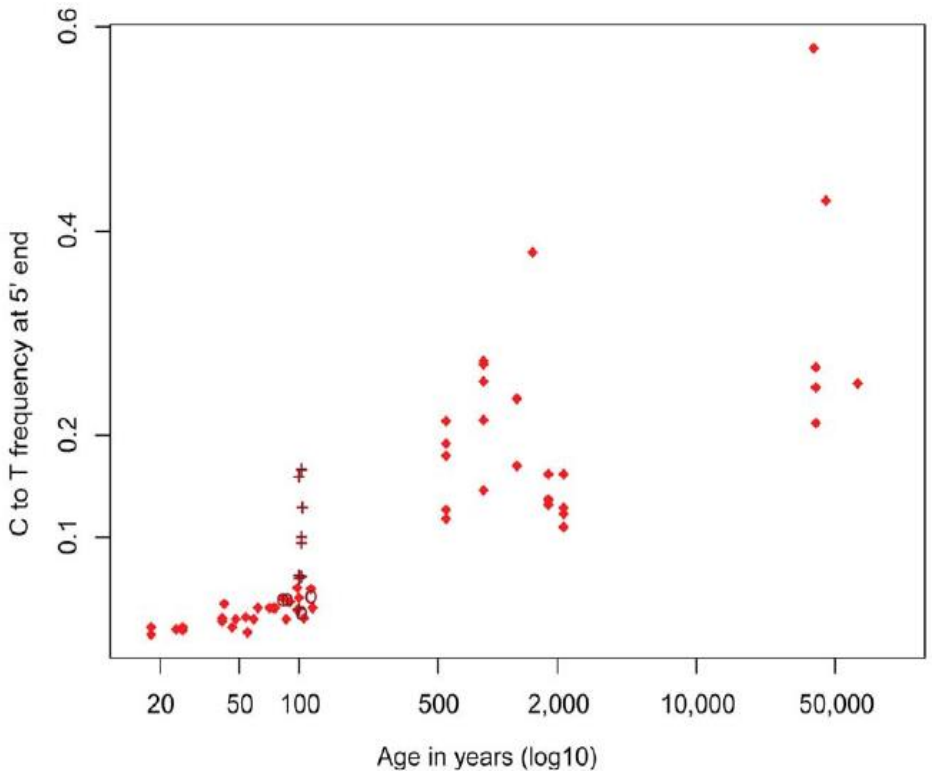


Figure 28: Occurrence of misincorporations damages in time.
X axis reflects the temporality of samples, while the Y axis indicates the extent of post mortem modifications for that specific sample. [Adapted from (Sawyer et al. 2012)]

3.12.2.2 Read length distribution

Given fragmentary nature of DNA, it will be expected that most of the DNA molecules will be found in small pieces, while very few sequences with have a medium-large length (100-200 nt) (Sawyer et al. 2012). Notwithstanding, as mentioned before, different preservation conditions could cause that DNA would not appear as degraded as it should, or on the other hand, appear more degraded than expected, in relation to the putative age of the sample. Therefore, by assessing if the read length histogram of the sequencing data from aDNA sample has a modal distribution (i.e. it has only one mean) one can hypothesize that most of the sequencing data corresponds to DNA from our sample of interest. Moreover, reads longer than expected (given several parameters to estimate DNA decay and degradation) and distribution outliers are usually regarded as putative contaminants (Green et al. 2009). **Figure 29** portraits sequence data from a sample that is heavily contaminated and thus a bimodal distribution of the sequence read length is obtained; the first peak of the histogram represents most of the endogenous DNA fragments from the sample, while the second stems from modern human contaminating DNA.

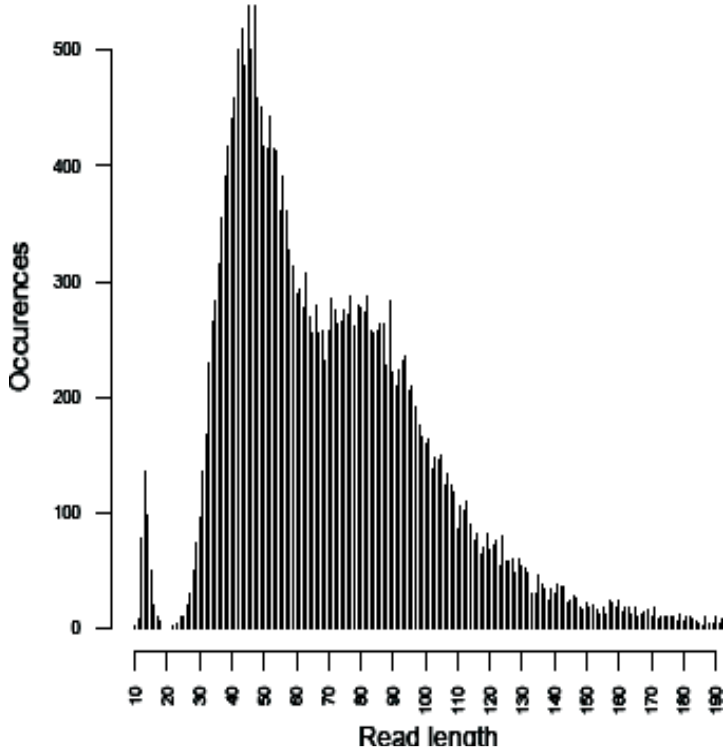


Figure 29. Histogram build from a contaminated aDNA sample of Louis XVI.
[Data use to generate plot was obtained from sequencing data from (Olalde, Sánchez-Quinto, et al. 2014)]

IV OBJECTIVES

Objectives

1) Analyse genetic data from populations not previously addressed by Neandertal introgression studies and determine to which extent these present-day human populations carry the fingerprint of admixture with Neandertals. This will help us to better understand the hybridization event with modern humans.

2) Obtain a general idea of the coding diversity patterns of Neandertals to identify genetic variants specific to their lineage that could be related to their known morphological appearance and to elucidate to which extent their patterns of genetic diversity reflect their evolutionary history.

3) Understand the genetic diversity and possible kinship relationships within a putative contemporaneous group of Neandertals in order to obtain information on the reproductive strategy and demographic dynamics of Neandertal groups

V Results

Chapter 1

North African Populations Carry the Signature of Admixture with Neandertals

Federico Sánchez-Quinto*, Laura R. Botigué*, Sergi Civit, Conxita Arenas, María C. Ávila-Arcos, Carlos D. Bustamante, David Comas, Carles Lalueza-Fox

Published

PLoS One. 2012;7(10):e47765.
doi:10.1371/journal.pone.0047765. Epub 2012 Oct 17.

Sánchez-Quinto F, Botigué LR, Civit S, Arenas C, Avila-Arcos MC, Bustamante CD, Comas D, Lalueza-Fox C. North African populations carry the signature of admixture with Neandertals. PLoS One. 2012;7 (10):e47765. doi: 10.1371/journal.pone.0047765.

Patterns of coding variation in the complete exomes of three Neandertals

Sergi Castellano, Genís Parraa, **Federico Sánchez-Quinto**, Fernando Racimo, Martin Kuhlwilm, Martin Kirchera, Susanna Sawyer, Qiaomei Fua, Anja Heinze, Birgit Nickel, Jesse Dabney, Michael Siebauer, Louise White, Hernán A. Burbano, Gabriel Renaud, Udo Stenzel, Carles Lalueza-Fox, Marco de la Rasilla, Antonio Rosas, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Matthias Meyer, Janet Kelso, Aida M. Andrés, and Svante Pääbo

Published

Proc Natl Acad Sci U S A. 2014 May 6;111(18):6666-71. doi:10.1073/pnas.1405138111. Epub 2014 Apr 21.

Castellano S, Parra G, Sánchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, Sawyer S, Fu Q, Heinze A, Nickel B, Dabney J, Siebauer M, White L, Burbano HA, Renaud G, Stenzel U, Lalueza-Fox C, de la Rasilla M, Rosas A, Rudan P, Brajković D, Kucan Ž, Gušic I, Shunkov MV, Derevianko AP, Viola B, Meyer M, Kelso J, Andrés AM, Pääbo S. Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci U S A. 2014 May 6;111(18):6666-71.
doi:10.1073/pnas.1405138111.

Chapter 3

Genetic evidence for patrilocal mating behavior among Neandertal groups

Carles Lalueza-Fox, Antonio Rosas, Almudena Estalrrich, Elena Gigli, Paula F. Campos, Antonio García-Tabernero, Samuel García-Vargas, **Federico Sánchez-Quinto**, Oscar Ramírez, Sergi Civit, Markus Bastir, Rosa Huguet, David Santamaría, M. Thomas P. Gilbert, Eske Willerslev, and Marco de la Rasilla.

Published

Proc Natl Acad Sci U S A. 2011 Jan 4;108(1):250-3. doi:10.1073/pnas.1011553108. Epub 2010 Dec 20.

Lalueza-Fox C, Rosas A, Estalrrich A, Gigli E, Campos PF, García-Tabernero A, García-Vargas S, Sánchez-Quinto F, Ramírez O, Civit S, Bastir M, Huguet R, Santamaría D, Gilbert MT, Willerslev E, de la Rasilla M. Genetic evidence for patrilocal mating behavior among Neandertal groups. Proc Natl Acad Sci U S A. 2011 Jan 4;108(1):250-3. doi: 10.1073/pnas.1011553108

VI Discussion

Discussion

Almost twenty years have passed since the first Neandertal sequences were recovered, and albeit the study of their evolutionary history was initially motivated to use their genetic information as a benchmark to identify those unique changes to our evolutionary lineage, a deeper understanding about their own evolutionary history is beginning to emerge.

The analyses of DNA recovered from Neandertal remains, to date, suggest that although they were a distinct hominin population to modern humans (Krings et al., 1997), a certain degree of genetic admixture occurred between both lineages (Green et al., 2010). Furthermore, recent evidence suggests that archaic introgressed material could have been important for modern humans to adapt to new environments (Williams et al., 2014). Additionally, their patterns of genetic variation from mitochondrial genetic sequences and a high-coverage genome suggest that Neandertals probably had a low effective population size throughout their evolutionary history, inferred from their low levels of genetic variation and non-expected accumulation of non-synonymous changes in their mitochondrial lineage (Briggs et al., 2009; Green et al., 2008).

It is clear that different evolutionary processes from population growth, bottlenecks and contractions, to adaptive events and admixture affected the dynamics of their evolutionary lineage, and maybe even contributed to their eventual demise. However, without a full characterization of their genetic uniqueness as a population, and an understanding of their demography and social structure, the extent to which these processes influenced their evolution will remain largely unknown.

In the current thesis we have investigated the Neandertal evolutionary history at three different resolution levels, from: clarifying their relatedness to previously unscreened modern human populations, to characterizing their demographic history from their patterns of coding variation, to taking a first glance at their population dynamics inferred from the genetic data recovered from a contemporaneous and putative social Neandertal group.
Over the following sections, a thorough analysis and discussion of each subject is provided in order to assess the impact and general considerations of the research chapters here presented, to better understand the Neandertal evolutionary history and start to address with genetic data the reasons behind their eventual demise.

Relatedness to modern humans and the implications of archaic genetic introgression

The first steps in order to study the evolutionary history of an organism are to comprehend which might be its closest phylogenetic relatives and to set a date to its most recent common ancestor.
Another matter of importance is to determine whether its evolutionary lineage remained in complete genetic isolation or not after the split from its most proximate evolutionary relatives. This notion is of importance not only to define the genetic uniqueness of its evolutionary lineage, but to understand the functional and biological implications of the hybridization event(s), as well as to address the origin and evolution of particular morphological and biological features on its lineage.

After nearly two decades of Neandertal paleogenetics studies, their phylogenetic relatedness to modern humans, as well as to other archaic humans, begins to be well understood.

Initial analysis of mitochondrial Neandertal sequences showed no evidence of interbreeding between them and modern humans, at least to

a level sufficient to be detected within the modern human mtDNA gene pool (Krings et al., 1997). More than 15 years after that initial assessment, paleogenomic studies have allowed not only to describe a previously unknown archaic hominin population –termed Denisovans, named after Denisovan Cave where the first fossil remains were found-, but also to discover that while Denisovans, Neandertals and modern humans shared a common ancestor around 800,000 years ago, the Denisovan and Neandertal genomes were more closely related to each other as sister species, their sequences started to diverged somewhere around 600,000 years ago (Reich et al., 2010).

Furthermore, insights from the high coverage genomes of a Siberian Neandertal and a female infant Denisovan individuals show that the population split between Denisovans, Neandertals and modern humans probably occurred between 383,000–257,000 years ago, while the populations leading to the Neandertal and Denisovan lineages separated roughly 236,000–190,000 years ago (Prüfer et al., 2014).
These high-quality archaic genomes in addition to the high-coverage genomes of modern human individuals from representative populations worldwide, have been useful to redefine previous assessments of Neandertal and Denisova admixture with modern humans (Green et al., 2010; Reich et al., 2010), to more accurate estimations of 1.5-2.1% of Neandertal genetic introgression into the ancestors of current non-Sub Saharan modern humans and ~ 3-6% of genetic admixture of Denisovans (or a hominin carrying Denisova-like DNA) with ancestors of present-day Melanesians.

In the case of the hybridization with Neandertals, several lines of evidence (Green et al., 2010; Prüfer et al., 2014; Sankararaman, Patterson, Li, Pääbo, & Reich, 2012) suggest that the admixture events could have taken place in the Middle East and/or Western Asia, when Neandertals encountered modern humans as they migrated out of Africa. However, even though a worldwide sampling of present-day

human populations was carried out in order to fully characterize the nature of hybridization events with archaic humans, it is unfortunate that modern human populations with more complex demographic histories and potentially different estimates of i.e. Neandertal genetic introgression were not considered for the analysis.

One of such modern human groups, are some North African populations. Henn et al. (2012) have recently shown that North African populations have an native genetic component not derived from late European or Middle Eastern migrations but rather stemming from a migration that probably took place somewhere between 12,000–40,000 years ago (Henn et al., 2012). Therefore, even though the genetic makeup of cosmopolitan populations like Egyptians, Algerians or Moroccans clearly reflect historical migrations from Middle East and Europe, other populations such as the Tunisian Berbers and the Saharawi (West Sahara) have high estimates of this autochthonous North African component.

In light of this evidence, it is tempting to hypothesize that if indeed the indigenous component observed in some North African populations originates from a Late Paleolithic/Mesolithic back-to-Africa migration, then the ancestors of these individuals could have been descendants from the populations that first interbreed with Neandertals about 37,000–86,000 years ago somewhere in the Middle East (Sankararaman et al., 2012), and thus carry the signature of admixture with Neandertals (see **Figure 30**).

Further characterizing the hybridization events with Neandertals is of importance, not only to understand to which extent other unaccounted admixture events might have occurred, but because the functional relevance that some of the introgressed regions seem to have had in some modern humans (Mendez, Watkins, & Hammer, 2012; Sankararaman et al., 2014; Vernot & Akey, 2014; Williams et al.,

2014) implies that other undocumented introgression events could have introduced the same or different genetic sequences of biological relevance for modern humans.

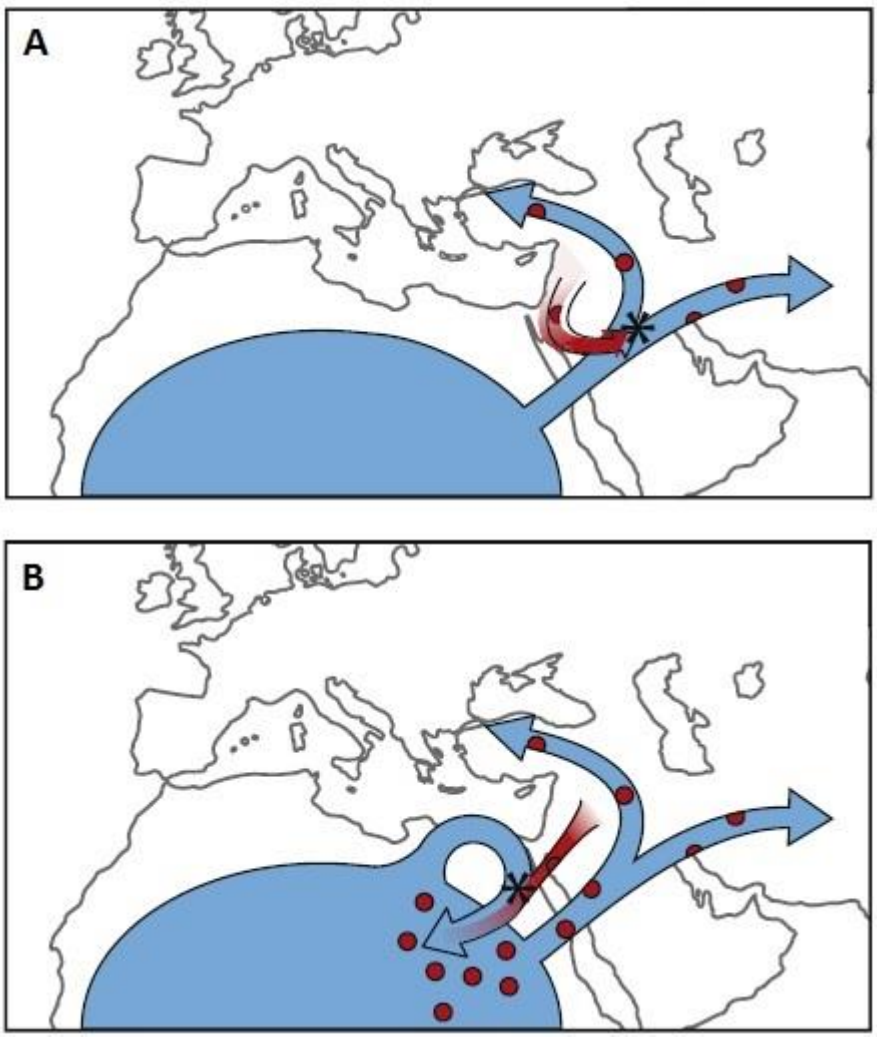


Figure 30: A scheme representing two complementary scenarios in which the ancestors of modern human population acquired genomic regions of Neandertal origin.

(A) Portrays the most known hypothesis were Neandertals and modern humans exchanged genes probably somewhere around the Middle East and/or Western Asia shortly after modern humans left Africa. **(B)** Shows a scenario at which North African populations could have traces of Neandertal DNA due to a Late Paleolithic/Mesolithic back-to-Africa migration of individuals bearing the signature of admixture with Neandertals [Adapted

Chapter 1 describes a study that aims at estimating the extent to which individuals representing different North African locations carry the signature of admixture with Neandertals, compared to other non-African present-day human populations.

Our results showed that North African populations vary in the percentage of Neandertal inferred admixture, primarily depending on the amount of European or Near Eastern ancestry they present. As expected, populations like North Moroccans and Egyptians which have a higher European and Near Eastern component, have also a higher amount of Neandertal ancestry. On the contrary, South Moroccans, that exhibit the highest Sub-Saharan component, showed the lowest Neandertal signal. Intriguingly, the analysis of the Tunisian Berbers, specially ten of them which have almost all of their genome assigned to the North African autochthonous ancestry, showed a higher Neandertal ancestry component than any other North African population and comparable to that of other Eurasian populations.

A further correlation, between the amount of Neandertal ancestry estimated in each North African population and the ancestry proportions inferred per population (i.e. the faction of European, Sub-Saharan, Middle Eastern and native North African heritage), revealed that a high European or Middle Easter ancestry is not sufficient to account for the predicted Neandertal ancestry in the North African populations and thus the North African native component contributes to the signal of gene flow from Neandertals. Assumed the aforementioned origin of the North African autochthonous component, a likely scenario is that individuals bearing this genetic ancestry also carry the signature of admixture with Neandertals; stemming from the same or other hybridization events with Neandertals throughout the Middle East and West Eurasia.
Moreover, although this is an insightful attempt to characterize Neandertal admixture events in other previously unaddressed modern

human populations, the origin of the putative admixture signal could have alternative explanations. Skoglund et al. (2011) state that studies using SNP array data to investigate the amount of genetic introgression from Neandertals, tend to report a higher similarity to Neandertals increasing with distance from Africa; an observation that they argument could be explained by SNP ascertainment bias plus a strong genetic drift in East Asian populations. Notwithstanding, the fact that other authors validated those initial results employing unbiased sequencing data, suggests that at least some of the putative admixture signal from Skoglund et al. (2011) should be correct (Meyer et al., 2012; Prüfer et al., 2014; Wall et al., 2013). However, more complex, population-biased, ascertainment schemes might have additional and foreseen effects (i.e. bottlenecks or isolation and drift), and thus the fact that the Tunisian population has been reported to be a genetic isolate (Henn et al., 2012), could imply that part of the archaic introgressed signal detected could actually be due to genetic drift.

Nevertheless, even though genetic drift could inflate the Neandertal ancestry predictions, it would not generate any false positives in admixture estimations, since to our best knowledge North African individuals have not been part of any array design panel and thus spurious high derived allele frequencies will be not expected. On the other hand, the fact that the signal is observed in populations with a high North African native component and which do not have a known demographic history of isolation (i.e. North Moroccans and Saharawis), advocates in favor of a previously undescribed Neandertal ancestry in those populations. What is more, given that SNP arrays are based on common alleles and probably the relevant admixture information is encoded within less frequent variants, the potential bias, if anything, will reduce the power to detect ancient hominin admixture, as shown in previous studies (Reich et al., 2010, 2011).

Therefore, while the finding of putative Neandertal admixture traces in the North Africa indigenous component demonstrate the necessity of a more complete sampling of modern human populations when assessing admixture events with archaic humans, the nature of the data employed for this analysis prohibit us to confirm our results. A complete clarification of this situation can only be achieved by employing sequencing data (preferably high quality data) of North African populations in order to authenticate the presence of Neandertals alleles in the individuals bearing the aforementioned native genetic component.

Clearly elucidating the interbreeding process of Neandertals (but also of other archaic hominins) with modern humans has two main applications in order to better understand archaic and modern human evolutionary histories. First, investigating to which extent other independent and undescribed admixture events took place, as well as determining the degree of shared and private Neandertal ancestries within all non-Sub Saharan populations (ideally including DNA data from Paleolithic modern humans) (Vernot & Akey, 2014; Wall et al., 2013) can shed light on the nature of hybridization event(s) with Neandertals; and on the degree of the genetic isolation between the two hominin groups (Sankararaman et al., 2014; Vernot & Akey, 2014). This latter notion is of great importance, not only to achieve a broader comprehension of the essence of genetic incompatibilities and their molecular mechanisms of action, but also contributed to our understanding of the biological definition of species. Do the skeletal remains found in Middle East with part Neandertal-like and AMH morphological features represent real archaic-modern human hybrids?

Furthermore, the biological relevance of introgressed sequences is also of great interest. Huerta-Sánchez et al. (2014), Sankararaman et al. (2014), Vernot et al. (2014) and Williams et al. (2014) have shown evidence that some modern human genomic regions of functional

relevance related to either the formation of keratin filaments, the susceptibility to certain genetic diseases or the capacity to transport oxygen at high altitudes, could be of Neandertal and Denisova origin. Therefore, a further understanding of the evolutionary origin of biological relevant or disease causing variants, as well as the functional meaning of archaic introgressed sequences could help us comprehend not only if these regions could have been advantageous after having been introgressed into modern humans, but also to characterize the etiology of present-day human´s diseases and detrimental traits.

Finally, as aforementioned before, the importance of the research carried out in **Chapter 1** resides at the interest to continue to detail the close relationship of Neandertals (as well as of other archaic hominins) with modern humans and its biological and evolutionary implications.

Inferences of Neandertal lineage specific changes and demographic history from the patterns of coding variation of three Neandertal exomes

Once Neandertal relatedness to modern humans has been defined by means of their divergence and extent of genetic admixture, the next level of resolution to study Neandertal evolution, is to elucidate their patterns of genetic diversity and population dynamics in order to investigate to which extent their demographic history could have affected their evolutionary history.

Section 2 provides a thorough review on what has been learned about Neandertal history throughout the past seventeen years of Neandertal aDNA studies. Undoubtedly, while the genetic data retrieved from Neandertal remains have set a milestone to our understanding of their evolution much remains to be understood on the impact of their population history on the dynamics of their evolutionary lineage. For

instance, while inferences from mitochondrial genetic sequences suggest that Neandertals had a low effective population size inferred from the reduced levels of genetic diversity and unexpected accumulation of non-synonymous changes, the fact that this data only reflects the evolutionary history from one genetic locus prevents us to make accurate inferences about their demographic history. Moreover, it is interesting that demographic inferences from the Altai Neandertal genome further support a similar scenario of a reduction of the effective population size in time. However, it is not clear to which extent this inference reflects the complete demographic history of Neandertals as a population or rather the outcome of isolated and inbred lineage (Prüfer et al., 2014).

In order to tackle this puzzling scenario and at the same time deal with the low levels of endogenous DNA recovered from typical non-permafrost samples, **Chapter 2** analyses the patterns of coding variation from the hybridized-capture exome regions of two Neandertal samples (El Sidrón and Vindija), together with the exomized regions from the high coverage genomes of the Altai Neandertal and the Denisovan individual, and compares them to those of three modern human individuals from Africa, Europe and Asia/Pacific.
It could seem implausible to investigate the Neandertal evolutionary history from the pattern of genetic variation of only three representative individuals, since large amounts of data are needed to reconstruct recent events from the evolutionary history of a population. However, information from few individuals is sufficient to reconstruct much older events. Therefore, by employing this approach, not only genetic changes specific to the Neandertal evolutionary lineage and possibly underlying some of their particular morphological features have been assessed, but also key insights into the demographic history of Neandertals have been investigated.

Neandertal specific lineage changes and the genetic basis underlying Neandertal morphological traits

While the high-coverage Neandertal and Denisova genomes provided a sound basis to identify genomic changes specific to modern humans, the exomes of the three Neandertals, the Denisovan individual and data from world-wide modern humans (1000 genomes) allow us for the first time to identify derived amino acid changes shared by three Neandertals as well as the Denisovan individual that are not seen, or only occur at a very low frequency in present-day humans, and thus are Neandertal specific or archaic specific changes.
By calculating the fraction of all amino acid changes specific to either the archaic or modern human lineages for each different gene-based phenotypic category in the *Human Phenotype Ontology database*, an estimation of the enrichment of amino acid changes for each phenotype of each archaic lineage has been be calculated (Castellano et al., 2014).

Results of this analysis suggested that while genes involved in skeletal morphology may have changed more on the Neandertal and Denisova lineages than on the preceding lineage from the common ancestor shared with chimpanzees, on the modern human lineage there is an over representation of genetic changes within genes related to traits such as "hyperactivity" or "aggressive behavior". It is interesting to note that the only category, for which there seems to be an enrichment of non-synonymous changes in the lineage leading to Neandertal, is *hyperlordosis*. This is a fascinating result, since a morphological feature common to most of all complete skeletal Neandertal remains (see **section 1.2.1**) is that they had a reduced lordotic curvature - the curvature of the lumbar and cervical spine-. Although it is not clear if those genetic variants reduce or enhance the expression of this phenotypical feature, a relationship between the genes harboring the genetic changes enriched in the Neandertal lineage and the morphological trait seems likely.

Unfortunately there is not enough morphological evidence from Denisovans to assess if a higher accumulation of non-synonymous changes specific to the lineage leading to both archaic humans and found within genes related to skeletal and morphological features could have been related to their phenotypical appearance.

Notwithstanding, it is interesting that while several phenotypic categories are enriched with non-synonymous changes in the lineage that leads to Neandertals and Denisovans, only hyperlordosis seems to have accumulated significantly more non-synonymous changes on the lineage leading to Neandertals than on the preceding lineage from the common ancestor shared with chimpanzees. This is noteworthy since some of the categories from the former list include morphological traits that can easily be related to a Neandertal phenotypical appearance like "*thorax anomalies*" or "*short distal limbs*".

This observation can be explained by mainly two different but complementary scenarios, related to the fact that this test is specific to derived changes on the shared evolutionary history of the Neandertal linages, and thus they must be present at high frequency or within all the Neandertal exomes analyzed. First, it could be possible that since the three Neandertal samples stem from different times and encompassing their complete geographical range, their most common recent ancestor could predate the appearance of some of these morphological characteristics. Second, even if contemporaneous samples would have been analyzed, the fact that the typical Neandertal features appear gradually in time, suggest that depending on the age of the samples analyzed, the underlying basis of certain derived morphological features might not be able to be addressed.

A subject of interest is to investigate whether adaptive events or genetic drift caused the unexpected enrichment of non-synonymous changes

associated to a given trait within each evolutionary lineage. However, the aforementioned analysis of the rate of accumulation of non-synonymous changes on each lineage is not an evolutionary test, and thus the responsible evolutionary force cannot be inferred.
A straightforward approximation to address this scenario will be to analyze if the genomic regions surrounding the non-synonymous changes at high frequency/fixated and related to a given phenotypic category have reduced patterns of genetic diversity, when compared to other non-synonymous changes that are also at high frequency or fixed derived, but which are not specifically related to any particular phenotype category. The rationale behind this approach is that if any of the aforementioned variants were important for survival, then positive selection would have caused an increase in frequency of the advantageous variant as well as their associated haplotypes. Therefore, a reduction in diversity surrounding the alleles and genes contributing to the phenotypic trait will be expected since linked variation will be inherited as a block.

Although carrying out this approach for traits enriched with genetic substitutions on the modern human lineage seems plausible (due to an increasing amount of worldwide sequencing data available), such a strategy directed towards answering the same question but for changes in the lineage leading to Neandertal or to both archaic humans is far more difficult to address. For instance, probably there will be no power to determine whether positive selection or genetic drift caused that morphological traits on the archaic lineage accumulated more mutations than expected (given that the fingerprints of an adaptive event are believed to last at most ~200,000 years), the same approximation to identify the evolutionary force behind the fixation of genetic variants related to hyperlordosis as well as other phenotypical features, will demand nuclear genetic data from several contemporaneous samples. However, as mentioned before, even though nuclear genetic data from contemporary samples eventually become

available, the fact that most Neandertal features evolve progressively in time, constrains that this approximation could only be undertaken with late Neandertal samples showing a full display their classical traits.
Another limitation will be the type of data needed for such analysis. While large contiguous regions of sequence data from high coverage genomes are usually preferred to detect drops of genetic diversity and to deduce if certain genomic regions might have experienced a higher selective constraint, currently such genetic data has only been recovered from a very high latitude Southern Siberian Neandertal sample. What is more, even though the limitation of having a reduced amount of endogenous DNA obtained from typical non-permafrost Neandertal samples, has been recently overcome by the ability to retain and sequence Neandertal DNA employing capture-hybridization methods, methodological and economical restrictions of such approximations hinders the possibility of scaling up the specific recovery of whole ancient genomes with current technology.

Moreover, the fact that current genetic and morphological Neandertal data suggest that genetic drift had a strong influence shaping the dynamics of the Neandertal evolutionary lineage, means that the accumulation of genetic changes related to the angle of their lordotic curvature might have also occurred as a consequence of demographic process rather than adaptive events.

It is important to keep in mind that the relevance of identifying the genetic basis of Neandertal specific phenotypical characteristics, or that of other archaic humans, is not only to better understand the evolution of their biological and morphological features, but also to pinpoint the underlying variants responsible for such traits in modern humans. Moreover, in order to accurately reconstruct their phenotypical appearance from genetic data, other types of genetic variants i.e. regulatory changes (Fraser, 2013), as well as other sources of

biologically relevant data such as ancient epigenetic modifications (Gokhman et al., 2014; Pedersen et al., 2014) need to be investigated.

At the end, our understanding of the biology of ancient humans, in this case Neandertals, will be not limited by the inaccessibility of data but by our interpretation of modern human genomes from a functional perspective (Hawks, 2013).

Demographic inferences from the patterns of coding variation from three Neandertal exomes

A comparison of the patterns of coding variation from Neandertal exomes with respect to that of modern humans has revealed interesting insights about their population history.
On the one hand, an analysis of the genetic variation, the extent of their runs of homozygosity (ROH) and the genetic differentiation among individuals measured by means of Fst, suggests that Neandertals effective population size was one-third of that of modern humans, and that they probably lived in small and relatively isolated population (Castellano et al., 2014).
On the other hand, an assessment of the evolution of their coding variants to investigate the extent to which their population history was reflected in their genetic makeup, reports an accumulation of deleterious/tolerable variants at SNPs with different frequencies for the derived allele and thus confirms previous predictions (Green et al., 2008) that Neandertals were evolving under a lower selective-constraint. The low levels of Neandertal nuclear genetic diversity, in addition to a predicted declining demographic trajectory (Prüfer et al., 2014) could imply that Neandertals had been evolving under long-term low effective population sizes. A reduced population size over a long time would decrease the efficacy of purifying selection and contribute to a larger fraction of slightly deleterious alleles, particularly at low frequency. It is interesting that such expectation is concordant with the

observed ratios of damaging/benign variants at SNPs believed to have been at low frequency among Neandertals, because they are seen once among the six chromosomes.

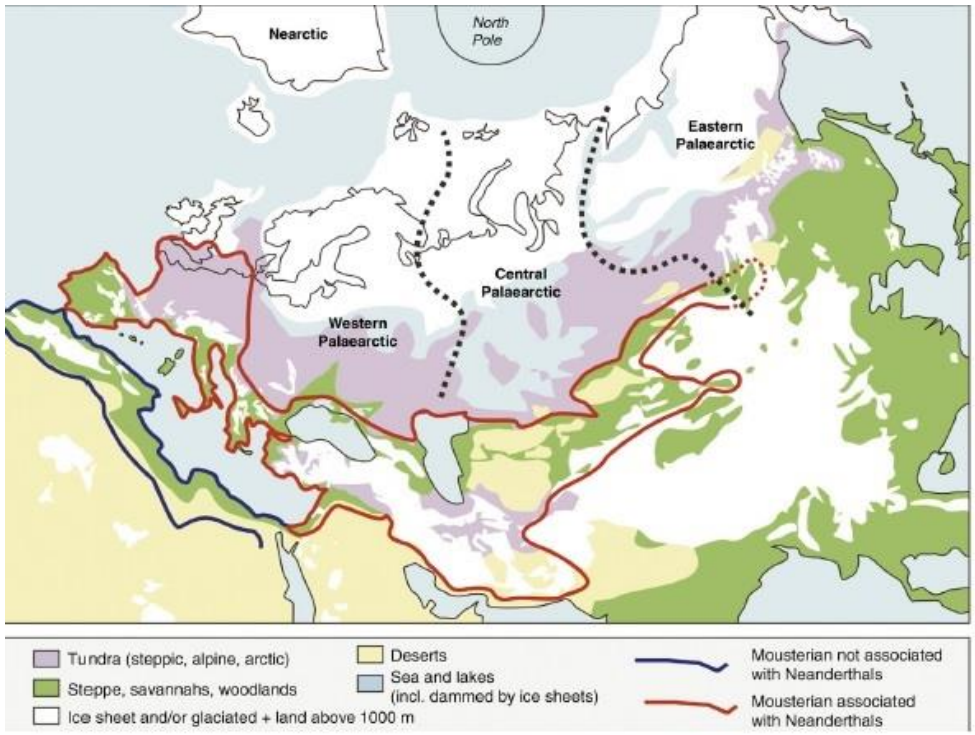
However, the fact that Neandertal individuals analyzed are not contemporaneous (and probably differing in time by as much as 20,000 years), could account for some of the observations.
For instance, a lack of contemporaneity of the samples could cause an overestimation of the genetic differences between the Neandertal individuals and thus, erroneous inferences about the relatedness of the Neandertal samples separated in time (Pontus Skoglund, Sjödin, Skoglund, Lascoux, & Jakobsson, 2014).
In addition, under such a scenario, a high ratio of deleterious/benign alleles at SNPs thought to be at low frequency in Neandertals or the large fraction of damaging variants per individual observed in Neandertals (Castellano et al., 2014), could be actually the reflection of demographic processes affecting each Neandertal lineage independently, rather than the fingerprint of a population evolving under a long-term low effective population size scenario.
On the other hand, a lack of coetaneous samples would not underestimate the assessments of genetic diversity. Therefore, genetic variation estimates are most likely a conservative prediction of the actual value; which seems to be a third of what is seen in present-day humans and in accordance with previous estimates from the mitochondrial lineage (Green et al., 2008).

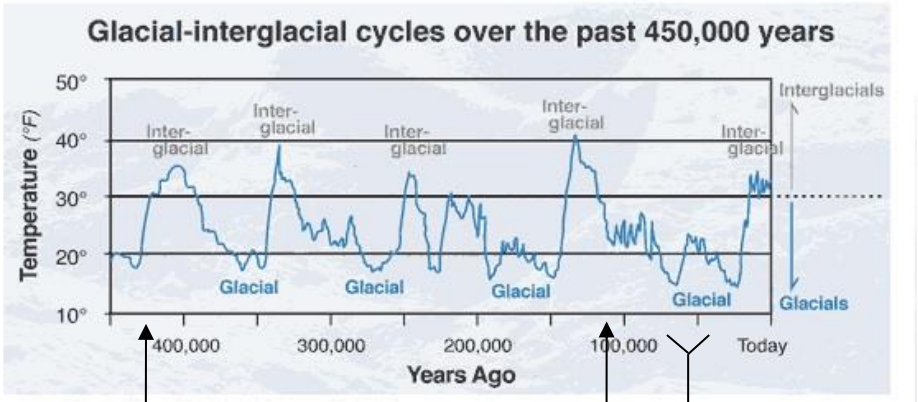
Moreover, samples from different time points will not bias the observation of a higher proportion of derived non-synonymous changes occurring at conserved positions which are homozygous in the all individuals within the "population" in Neandertals than in modern humans. Therefore this finding implies that at least one major demographic event caused a reduction of the efficiency of purifying

selection leading deleterious alleles to become fixed in the ancestral line to the three Neandertal populations.
Finally, current observations hinder us from inferring that Neandertals lived in isolated populations given their estimates of genetic differentiation. However, the fact that they present longer runs of homozygosity than modern humans (and even presenting significant levels of inbreeding), suggest that their interaction with other Neandertal groups might have been reduced.

Another subject of much interest is how inferences from genetic data reconcile well with archaeological and anthropological findings about the general aspects of their evolutionary history.
For instance, insights from both mitochondrial and nuclear data from Late Pleistocene Neandertal samples suggest that their population faced at least one serious bottleneck (Dalén et al., 2012; Prüfer et al., 2014). This observation coupled to the aforementioned progressive evolution of their morphological features, hints at that their evolutionary history was influenced by major environmental climate changes. Most likely manifested by population expansions and reductions related to the spread and contraction of ice sheets and their respective permafrost areas, during glacial maxima and inter glacial periods (see **Figure 31**).

It is surprising that while the geographical extent and duration of major ice ages occurring through the Middle and Late Pleistocene unquestionably shaped the dynamics of their evolutionary lineage, such harsh conditions in isolation were not reflected into any further speciation events with this hominin lineage. This observation is contrary to what has been reported from other species such as the brown bear or the European hedgehog, which while inhabiting a similar geographical region throughout roughly the same period of time, diverged into different subspecies (Liu et al., 2014; Santucci, Emerson, & Hewitt, 1998; Talbot & Shields, 1996).





Estimated age of Sima de los Huesos (Spain) hominins

TMRCA of all mtDNA Neandertal genomes recovered to date

Temporal range from which samples bearing DNA have been assessed

Figure 31: A cartoon depicting how Europe might have looked during an ice age and the temporal span of major glacial-interglacial cycles during the last half million years.

Upper image displays how the Northern Hemisphere might have looked during the Last Glacial Maximum, and thus it portrays the severity of the environmental conditions were also experience during past ice ages. Bottom image also displays the date of some important Neandertal samples from which DNA has been extracted or the time to the most common recent ancestor. [Upper image is adapted from (C. Finlayson & Carrión, 2007)and bottom from http://geology.utah.gov/surveynotes/gladasked/gladice_ages.htm]

Such scenario suggests that either Neandertal population dynamics still continued during periods of severe environmental conditions (avoiding a significant genetic isolation between possible structure populations), or expansions and retractions involved major population replacements which absorbed the surviving individuals from northern glacial refugia and imposed the evolutionary lineage of the migrating populations. While both scenarios could have occurred, the gradual emergence of their typical morphological features together with evidence of important population bottlenecks invites to think that expansion and colonization from southern ice age refugia was a key aspect that shaped the development of their evolutionary history (see **Figure 31**).

STRUCTURE-like analyses of genetic data from several Neandertal individual samples from different time points and geographical locations, correlated with a complete record of climatic changes, will be paramount to elucidate the extent to which a series of population reductions and expansions affected the dynamics of their evolutionary history. For now, an analysis of patterns of coding variation from three Neandertals throughout their geographical range confirms that they had a different demographic history to that of modern humans.

Finally, a subject of much interest is to discuss to which extent the particularities of the Neandertal demographic history could have influenced their extinction process.

An example of the genetic burden of a population from events of their demographic history, is the initial settlement of the current French-Canadian population in the East coast of Canada (Casals et al., 2013). Most of present-day *Québécois* descend from a French immigrant founder group of very few individuals and less than 20 generations ago. By analyzing the whole-exomes of both French-Canadian and French individuals, Casals et al. (2013) showed that in less than 20 generations of genetic isolation from the French population, the genetic pool of

French-Canadians showed reduced levels of diversity, higher homozygosity, and an excess of rare variants with low variant sharing with Europeans. Furthermore, the French-Canadian population contains a larger proportion of putatively damaging functional variants, which could partially explain the increased incidence of genetic disease in the province.

In light of this evidence, it is straightforward to think that since Neandertals were evolving under a low selective constraint, in addition to their predicted population isolation and inbreeding practices, an increase in their overall deleterious load within the population towards the end of their evolutionary time could have contributed to their eventual demise.

The deleterious load of an individual can be defined by a reduction of the individual fitness which is directly related to the number of deleterious alleles carried by such individual. A first approximation to assess the deleterious load per individual from the exomic data of the three Neandertals, revealed that the total number of genes associated to recessive traits with heterozygous or homozygous derived alleles inferred to be deleterious is not statistically different between Neandertal and present-day individuals, and therefore the susceptibility of Neandertals to any specific genetic disorder cannot be inferred from these data. This observation could have two different explanations. From one side, Simons et al. (2014) state that the individual deleterious load is insensitive to recent population history, given that depending on the degree of dominance and the strength of selection, an accumulation of deleterious variants will only be transitory since demographic processes will eventually weed out the damaging alleles. On the other hand, as with the previous assessments of accumulation of deleterious alleles at lower frequency, it could be that a lack of contemporaneity between the samples biases the estimation of this analysis, unlike the characterization of the French-Canadian population in which multiple coetaneous samples were analyzed.

Until genomic data from several contemporaneous individuals per site encompassing different temporal points becomes available, a final word on the consequences of their particular demography and population dynamics (i.e. a higher susceptibility to recessive disorders) will not be able to be inferred.

Pattern of mitochondrial genetic variation from a social Neandertal group

A ground level to study Neandertal evolution is to understand the genetic variation and kinship relationships from its social groups. As mentioned before, the genetic diversity within a population is determined by the biological capacity of a species to generate genetic variation and the distribution of such diversity, and by the interaction of different processes such as migration, genetic drift, adaptation and admixture.
Additionally both the over action of one or several of these processes, as well as specific environmental, behavioral, hereditary or social factors may create a population structure. Therefore, by studying the genetic makeup, the kinship relationships and the distribution of sexes within a group of contemporaneous individuals, the factors that contributed to determine such levels of genetic diversity and population structure can be elucidated.

Finding several DNA-bearing Neandertal individuals in order to study their local genetic diversity and social structure has been virtually impossible in the past. While the scarcity of Neandertal fossils and the fragmentary nature of the skeletal remains hinders the identification of coetaneous samples, the harsh environmental conditions from which most skeletal remains are usually recovered (even when such samples are available) has prevented the characterization of the genetic relatedness between the individuals. Fortunately, there is a Neandertal site with fossil remains belonging to individuals of different age and sex which could help us to study such scenario, since they may

represent the remains of a social/family group. Discovered and excavated since the year 2000 and dated to around 49,000 YA, the Spanish site of El Sidrón is thought to be a synchronic accumulation of 12 Neandertals (a further possible individual has been recently identified last year) including 6 adults, three adolescents, two juveniles and one infant (see **Figure 33**).
While the age and gender distribution inferred from morphological features of El Sidrón individuals are similar to those found in contemporary hunter-gatherer societies, without genetic data their individual relatedness and a characterization of their demographic social dynamics cannot be attained.

Chapter 3 describes the test for presence or absence of the Y chromosome and the analysis of mtDNA sequences in the El Sidrón individuals, to obtain information about the kinship, gender distribution, and genetic diversity of a Neandertal group.

Results show that the 12 individuals stem from three different maternal lineages (seven bearing one lineage, four another and only one having a different third mitochondrial haplotype). Such observations, in addition to the genetic characterization of their sex, revealed that although the three adult males carried the same mtDNA lineage, each of the three adult females had different mitochondrial lineages. Additionally, pair-wise differences of the mitochondrial sequences of El Sidrón individuals suggest that genetic variation was reduced within a contemporaneous social group of Neandertal individuals. Altogether these findings provide evidence advocating that Neandertal groups not only were small and characterized by low genetic diversity but also were likely to have practiced patrilocal mating behavior.

However, although these observations provide tantalizing clues about the Neandertal group structure and social dynamics, some caveats of the study and aspects of the data should be considered into account in order to validate the previous inferences.

For instance, the method employed to infer the gender of the individuals can generate false-positive results for females, since an absence of amplification of Y-chromosome sequences could be due to insufficient genomic coverage in a particular sample. Such an incorrect sex determination could bias the results and the conclusions from the analyses herein presented. Nonetheless, the fact that the genetic estimations are concordant with the morphological characterization brings further support to the veracity of these assessments. A final and accurate gender characterization can only be attained by mapping high quality nuclear data to either sexual chromosome from each sample.

In addition, another confounding factor of the study is that this approximation only focuses in a few diagnostic positions and thus further unidentified genetic variability could have been largely overlooked and affected both the relatedness and genetic diversity estimations. However, phylogenetically informative positions where inferred using all the mitochondrial genetic diversity data available at the time. Moreover, as data stems from HVR I-II from mtDNA, where the mutation rate is one order of magnitude higher than in the coding region, it would indicate that if variation was not found in the regions where the most phylogenetic informative data was expected, not much data would remain unsampled. Nonetheless, even if additional information is collected, the fact that the resulting haplotypes stem from the aforementioned El Sidrón A, B, and C lineages, should not affect significantly the pattern of diversity observed in this study.

One of the most interesting questions emerging from genetic data retrieved from a synchronic contemporaneous group of Neandertals, is to which extent the individuals investigated were related to each other, and whether or not their genetic affinities represent those of a family group.

While both the morphological and genetic data from the study suggest a social structure configuration where direct consanguinity relationships or at least partially relatedness could be inferred by having two groups of several close relatives in the female line, a close maternal relatedness between the individuals -specially the adult ones- could have other interpretations. Under the aforementioned demographic scenario for Neandertals, it could be possible that two individuals seem more closely related to each other genetically by either a recent and drastic population bottleneck or historically reduced levels of genetic variation, rather than belonging to the same kin structured group. While this alternative explanation intended to be addressed in the analysis by showing that the probability of recovering such a structure configuration from a large sample of HVRs sequences from a big database of unrelated modern humans was low, the fact that modern humans and Neandertals had such different demographic histories could be biasing the results. Therefore, although accurately inferring kinship relations among El Sidrón Neandertal group awaits the recovery of genome-wide nuclear data from the same individuals, the low levels of genetic diversity observed among these individuals does not seem to be experimentally underestimated.

Investigating the structure of a possible social Neandertal group provides not only exciting hints about the demography and behavior of this hominin group, but it also summed up information which may be helpful when choosing more demographic parameters to generate models for Neandertal population dynamics in order to investigate their eventual demise (see **Figure 32**). Therefore, a significant effort should be placed towards recovering genome-wide data from the El Sidrón individuals. The fact that their fossils are dated towards the time of their extinction and that there is minimal probability of finding another Neandertal group with similar characteristics as El Sidrón individuals, highlights the opportunity of retrieving genomic DNA from such specimens. We stand before a unique chance to obtain a detailed

account of the consequences of their demographic history on their final days.

Figure 32: An illustration representing the appearance of the possible social Neandertal group from El Sidrón. [Adapted from the exposition "Los 13 del Sidrón" at http://www.museoarqueologicodeasturias.com].

Concluding remarks

Albeit the initial interest to extract DNA from Neandertals remains was probably motivated to identify the genetic basis of our uniqueness as a species, an accurate reconstruction of their evolutionary history (as well as that of other hominins) will continue to shed light about the singularity of our lineage, by addressing why, after living over such a long evolutionary time, did they eventually got extinct whereas modern humans prevailed.

Without question the next step to complete an accurate and detailed picture from the evolutionary history of Neandertals is to recover and analyze high quality nuclear data from several individuals at different time points. These data will help us address the major pending inquiries about their evolutionary history. By obtaining DNA from samples of different temporal ranges, a clearer understanding about the origin of their lineage can be elucidated by investigating when their MRCA lived. Such a collection of sequencing data will also allow us to obtain a thorough reconstruction of their population.

In addition, high quality genome-wide data from several individuals displaying their classic phenotypic features will be useful to elucidate the genetic basis of their distinctive morphological traits.
Middle Eastern samples with part Neandertal-like and AMH morphological features will be of particular interest to address to which extent they represent real archaic-modern human hybrids and to further characterized the degree of the genetic isolation between hominin groups. However, as mentioned before, to obtain a full clarification on this issue, analysis of a wide collection of modern human samples will be required.
To address the previously mentioned questions an extensive sampling of new fossils will be needed, and even though ongoing archaeological excavations will hopefully continue to produce material for aDNA

studies, it is clear that a number of Neanderthal samples of interest may be stored within museums under less than ideal conditions, or may not have been excavated and handled with enough care to prevent contamination. Two main caveats arise from this: first many specimens will probably have low endogenous DNA contents, and second, they might have been contaminated significantly with modern human DNA.

As samples from older periods are screened in search for genetic material, even sequencing a mitochondrial genome may require significant amounts of bone tissue (Meyer et al., 2014), which may enter in conflict with conservation purposes. Target capture techniques have proven to be the most efficient in accessing samples with low endogenous DNA. However, only certain genomic regions (e.g., mtDNA or exomes) have been retrieved with high-coverage using this approach. While WISC has been recently employed to recover genome-wide data from ancient samples, its bias against shorter DNA molecules (like the ones found in most Neandertal fossils) is an issue that needs to be addressed before it can be fruitfully applied to samples with highly degraded DNA.

Therefore the challenge for the upcoming years, not only for Neandertal paleogenomics but for any genome-wide aDNA approximation that is dealing with extremely degraded and possibly highly contaminated samples, is to develop new experimental and *in silico* approaches to cope with these issues.

For instance, a recently published statistical and computational approach aimed to separate putative endogenous sequences from contaminant material (P. Skoglund et al., 2014). However, this approximation does not preclude the sequencing of contaminant material, which might not be suitable if a high number of poorly handled and preserved samples have to be screened.

An experimental solution could be on the same line as the a recently presented methodological design by Gansauge and Meyer (2014), which intended to enrich the amount of endogenous material, by

experimentally retaining only the informative damaged molecules(Gansauge & Meyer, 2014). Although this method will greatly improve the suitability of the samples for deeper shotgun sequencing, the exact proportion of uracil-containing molecules varies among fossils and cannot be accurately predicted from the age of a sample alone. More approaches with similar scopes will be very welcome by the aDNA community.

At the end, all aDNA studies will still be limited by the amount of endogenous DNA present in the sample. Until new methodological approaches are available, target capture and shotgun sequencing will no doubt continue to be used, depending on the preservation condition of the samples and the scientific questions being addressed. Moreover, it remains to be seen to which extent TGS technologies can transform the field of aDNA and hominin palaeogenomics.

Finally, from all the aforementioned subjects regarding the evolutionary history of Neandertals, understanding the reasons behind their population collapse is probably the most appealing and controversial subject.
While some lines of evidence suggest that climate changes might have caused their extinction, other hypotheses propose that Neandertals were probably overrun by modern humans (G. Finlayson et al., 2006; Mellars & French, 2011). However, an immediate third possibility is that a combination of both factors contributed to certain extent. It could be that the consequences of their particular population history shaped by harsh environmental conditions left Neandertals in such a demographic predicament that after the arrival of modern humans they were easily displaced, absorbed or succumbed to infectious diseases to which they have never been previously exposed.
Genetic data will be paramount to assess such hypothesis. In particular to understand to what extent Neandertals were affected by their small population size, relative isolation and inbreeding practices. For

example, the new data might allow us to observe if they displayed a significant accumulation of variants associated with recessive disorders in comparison with modern humans. While this is just a hypothesis, it could be that an accumulation of genetic deleterious variants associated with decreased effective population size, exacerbated by inbreeding practices in the last Neandertals, may have contributed to their final demise.

Bibliography

Abi-Rached, Laurent, Matthew J Jobin, Subhash Kulkarni, Alasdair McWhinnie, Klara Dalva, Loren Gragert, Farbod Babrzadeh, et al. 2011. "The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans." *Science (New York, N.Y.)* 334 (6052) (October 7): 89–94.

Andrés Moreno-Estrada. 2009. "Analysis of Human Genetic Variation in Candidate Genes under Positive Selection on the Human Lineage". Universitat Pompeu Fabra.

Arsuaga, JL, I Martínez, L J Arnold, A Aranburu, A Gracia-Téllez, W D Sharp, R M Quam, et al. 2014. "Neandertal Roots: Cranial and Chronological Evidence from Sima de Los Huesos." *Science (New York, N.Y.)* 344 (6190) (June 20): 1358–63.

Ávila-Arcos, Maria C. 2012. "Data Analysis Methods for Paleogenomics". University of Copenhagen. National History Museum of Denmark.

Avila-Arcos, María C, Enrico Cappellini, J Alberto Romero-Navarro, Nathan Wales, J Víctor Moreno-Mayar, Morten Rasmussen, Sarah L Fordyce, et al. 2011. "Application and Comparison of Large-Scale Solution-Based DNA Capture-Enrichment Methods on Ancient DNA." *Scientific Reports* 1 (January): 74.

Avila-Arcos, Maria, Hannes Sandoval-Velasco, Marcela Schroeder, Anna-Sapfo Carpenter, Meredith L Malaspinas, Nathan Wales, Carlos D Peñaloza, Fernando Bustamante, and M. Thomas P Gilbert. 2014. "Comparative Performance of Two Whole Genome Capture Methodologies on Ancient DNA Illumina Libraries." *bioRxiv* (July 24).

Barbujani, G, A Magagni, E Minch, and L L Cavalli-Sforza. 1997. “An Apportionment of Human HLA Diversity.” *Proc. Natl. Acad. Sci.* 94 (April) (April): 4516–4519.

Benazzi, Stefano, Katerina Douka, Cinzia Fornai, Catherine C Bauer, Ottmar Kullmer, Jiří Svoboda, Ildikó Pap, et al. 2011. “Early Dispersal of Modern Humans in Europe and Implications for Neanderthal Behaviour.” *Nature* 479 (7374) (November 24): 525–8.

Bentley, DR, S Balasubramanian, HP Swerdlow, HP Smith, J Milton, CG Brown, KP Hall, et al. 2008. “Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry.” *Nature* 456 (November).

Bermudez de Castro, J. M. 1997. “A Hominid from the Lower Pleistocene of Atapuerca, Spain: Possible Ancestor to Neandertals and Modern Humans.” *Science* 276 (5317) (May 30): 1392–1395.

Bordes, Jean-guillaume, David Cochard, Isabelle Couchoud, David Dubrasquet, Erik Trinkaus, and Les Rochers-de-villeneuve. 2005. “A Late Neandertal Femur from Les.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (20): 7085–7090.

Briggs, AW, J M Good, RE Green, J Krause, T Maricic, U Stenzel, C Lalueza-Fox, et al. 2009. “Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes.” *Science (New York, N.Y.)* 325 (5938) (July): 318–21.

Briggs, AW, Udo Stenzel, Philip L F Johnson, Richard E Green, Janet Kelso, Kay Prüfer, Matthias Meyer, et al. 2007. “Patterns of Damage in Genomic DNA Sequences from a Neandertal.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (37) (September 11): 14616–21.

Briggs, AW, Udo Stenzel, Matthias Meyer, Johannes Krause, Martin Kircher, and Svante Pääbo. 2010. “Removal of Deaminated

Cytosines and Detection of in Vivo Methylation in Ancient DNA." *Nucleic Acids Research* 38 (6) (April): e87.

Brotherton, P., P. Endicott, J. J. Sanchez, M. Beaumont, R. Barnett, J. Austin, and a. Cooper. 2007. "Novel High-Resolution Characterization of Ancient DNA Reveals C > U-Type Base Modification Events as the Sole Cause of Post Mortem Miscoding Lesions." *Nucleic Acids Research* 35 (17) (August 23): 5717–5728.

Brown, P, T Sutikna, M J Morwood, R P Soejono, E W Saptomo, and Rokus Awe Due. 2004. "A New Small-Bodied Hominin from the Late Pleistocene of Flores , Indonesia." *Nature* 431: 1055–1061.

Brunet, Michel, Franck Guy, David Pilbeam, Hassane Taisso Mackaye, Andossa Likius, Djimdoumalbaye Ahounta, Alain Beauvilain, et al. 2002. "A New Hominid from the Upper Miocene of Chad , Central Africa": 145–152.

Burbano, H. a., E. Hodges, R. E. Green, a. W. Briggs, J. Krause, M. Meyer, J. M. Good, et al. 2010. "Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture." *Science* 328 (5979) (May 6): 723–725.

Burrows, M, and DJ Wheeler. 1994. "A Block-Sorting Lossless Data Compression Algorithm." *Digital*.

Caramelli, David, Carles Lalueza-Fox, Silvana Condemi, Laura Longo, Lucio Milani, Alessandro Manfredini, Michelle de Saint Pierre, et al. 2006. "A Highly Divergent mtDNA Sequence in a Neandertal Individual from Italy." *Current Biology : CB* 16 (16) (August 22): R630–2.

Caramelli, David, Carles Lalueza-fox, Cristiano Vernesi, Martina Lari, Antonella Casoli, Francesco Mallegni, Brunetto Chiarelli, et al. 2003. "Evidence for a Genetic Discontinuity between Neandertals and 24 , 000-Year-Old Anatomically Modern Europeans."

Proceedings of the National Academy of Sciences 100 (11): 6593–6597.

Carpenter, Meredith L., Jason D. Buenrostro, Cristina Valdiosera, Hannes Schroeder, Morten E. Allentoft, Martin Sikora, Morten Rasmussen, et al. 2013. "Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries." *The American Journal of Human Genetics* 93 (5) (November): 852–864.

Casals, Ferran, Alan Hodgkinson, Julie Hussin, Youssef Idaghdour, Vanessa Bruat, Thibault de Maillard, Jean-Christophe Grenier, et al. 2013. "Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans." *PLoS Genetics* 9 (9) (January): e1003815.

Castellano, Sergi, Genís Parra, Federico a Sánchez-Quinto, Fernando Racimo, Martin Kuhlwilm, Martin Kircher, Susanna Sawyer, et al. 2014. "Patterns of Coding Variation in the Complete Exomes of Three Neandertals." *Proceedings of the National Academy of Sciences of the United States of America* 111 (18) (May 6): 6666–71.

Cooper, A, C Mourer-Chauviré, G K Chambers, A von Haeseler, A C Wilson, and S Pääbo. 1992. "Independent Origins of New Zealand Moas and Kiwis." *Proceedings of the National Academy of Sciences of the United States of America* 89 (18) (September): 8741–4.

Cooper, Alan, and Hendrik N Poinar. 2000. "Ancient DNA: Do It Right or Not at All." *Science* 289 (5482) (August 18): 1139.

Dabney, Jesse, Michael Knapp, Isabelle Glocke, Marie-Theres Gansauge, Antje Weihmann, Birgit Nickel, Cristina Valdiosera, et al. 2013. "Complete Mitochondrial Genome Sequence of a Middle Pleistocene Cave Bear Reconstructed from Ultrashort DNA Fragments." *Proceedings of the National Academy of Sciences of the United States of America* 110 (39) (September 24): 15758–63.

Dalén, Love, Ludovic Orlando, Beth Shapiro, Mikael Brandström-Durling, Rolf Quam, M Thomas P Gilbert, J Carlos Díez Fernández-Lomana, Eske Willerslev, Juan Luis Arsuaga, and Anders Götherström. 2012. "Partial Genetic Turnover in Neandertals: Continuity in the East and Population Replacement in the West." *Molecular Biology and Evolution* 29 (8) (August): 1893–7.

Davis, JB. 1867. *Thesaurus Craniorum: Catalogue of the Skulls of the Various Races of Man, in the Col- Lection of Joseph Barnard Davis*. London.

Durand, Eric Y, Nick Patterson, David Reich, and Montgomery Slatkin. 2011. "Testing for Ancient Admixture between Closely Related Species." *Molecular Biology*: 1–53.

Eriksson, Anders, and Andrea Manica. 2012. "Effect of Ancient Population Structure on the Degree of Polymorphism Shared between Modern Human Populations and Ancient Hominins." *Proceedings of the National Academy of Sciences of the United States of America* 109 (35) (August 28): 13956–60.

Eriksson, Anders, and Andrea Manica. 2014. "The Doubly-Conditioned Frequency Spectrum Does Not Distinguish between Ancient Population Structure and Hybridisation." *Molecular Biology and Evolution*: 1–13.

Excoffier, L, P E Smouse, and J M Quattro. 1992. "Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data." *Genetics* 131: 479–91.

Falgueras, Juan, Antonio J Lara, Noé Fernández-Pozo, Francisco R Cantón, Guillermo Pérez-Trabado, and M Gonzalo Claros. 2010. "SeqTrim: A High-Throughput Pipeline for Pre-Processing Any Type of Sequence Read." *BMC Bioinformatics* 11 (January): 38.

Fedurco, M, A Romieu, S Williams, I Lawrence, and G Turcatti. 2006. "BTA, a Novel Reagent for DNA Attachment on Glass and Efficient Generation of Solid-Phase Amplified DNA Colonies." *Nucleic Acids Research* 34 (3) (January): e22.

Ferring, Reid, Oriol Oms, Jordi Agustí, Francesco Berna, Medea Nioradze, Teona Shelia, Martha Tappen, Abesalom Vekua, David Zhvania, and David Lordkipanidze. 2011. "Earliest Human Occupations at Dmanisi (Georgian Caucasus) Dated to 1.85-1.78 Ma." *Proceedings of the National Academy of Sciences of the United States of America* 108 (26) (June 28): 10432–6.

Finlayson, Clive, and José S Carrión. 2007. "Rapid Ecological Turnover and Its Impact on Neanderthal and Other Human Populations." *Trends in Ecology & Evolution* 22 (4) (April): 213–22.

Finlayson, Geraldine, Ethel Allue, Javier Baena Preysler, Isabel Ca, Christopher P Gleed-owen, Francisco J Jimenez Espejo, Pilar Lo, et al. 2006. "Late Survival of Neanderthals at the Southernmost Extreme of Europe." *Nature* 443 (October): 850–853.

Fonseca, NA, J Rung, A Brazma, and JC Marioni. 2012. "Tools for Mapping High-Throughput Sequencing Data." *Bioinformatics (Oxford, England)* 28 (24) (December 15): 3169–77.

Fortea, Javier, Marco de la Rasilla, Antonio García-Tabernero, Elena Gigli, Antonio Rosas, and Carles Lalueza-Fox. 2008. "Excavation Protocol of Bone Remains for Neandertal DNA Analysis in El Sidrón Cave (Asturias, Spain)." *Journal of Human Evolution* 55 (2) (August): 353–7.

Fraser, Hunter B. 2013. "Gene Expression Drives Local Adaptation in Humans." *Genome Research* 23 (7) (July): 1089–96.

Fu, Qiaomei, Matthias Meyer, Xing Gao, Udo Stenzel, Hernán a Burbano, Janet Kelso, and Svante Pääbo. 2013. "DNA Analysis of an Early Modern Human from Tianyuan Cave, China."

Proceedings of the National Academy of Sciences of the United States of America 110 (6) (February 5): 2223–7.

Gansauge, Marie-Theres, and Matthias Meyer. 2013. "Single-Stranded DNA Library Preparation for the Sequencing of Ancient or Damaged DNA." *Nature Protocols* 8 (4) (April): 737–48.

Gansauge, Marie-Theres, and Matthias Meyer. 2014. "Selective Enrichment of Damaged DNA Molecules for Ancient Genome Sequencing." *Genome Research* (July 31).

Gibbons, A. 1994. "Possible Dino DNA Find Is Greeted With Skepticism." *Science* 266 (November): 1994.

Gilbert, M Thomas P, Lynn P Tomsho, Snjezana Rendulic, Michael Packard, Daniela I Drautz, Andrei Sher, Alexei Tikhonov, et al. 2007. "Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts M." *Science* 317.

Glenn, Travis C. 2011. "Field Guide to next-Generation DNA Sequencers." *Molecular Ecology Resources* 11 (5) (September): 759–69.

Gnirke, Andreas, Alexandre Melnikov, Jared Maguire, Peter Rogov, Emily M LeProust, William Brockman, Timothy Fennell, et al. 2009. "Solution Hybrid Selection with Ultra-Long Oligonucleotides for Massively Parallel Targeted Sequencing." *Nature Biotechnology* 27 (2) (February): 182–9.

Gokhman, David, Eitan Lavi, Kay Prüfer, Mario F Fraga, José a Riancho, Janet Kelso, Svante Pääbo, Eran Meshorer, and Liran Carmel. 2014. "Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan." *Science (New York, N.Y.)* 344 (6183) (May 2): 523–7.

Green, RE, Adrian W Briggs, Johannes Krause, Kay Prüfer, Hernán a Burbano, Michael Siebauer, Michael Lachmann, and Svante Pääbo. 2009. "The Neandertal Genome and Ancient DNA

Authenticity." *The EMBO Journal* 28 (17) (September 2): 2494–502.

Green, RE, Johannes Krause, Susan E Ptak, Adrian W Briggs, Michael T Ronan, Jan F Simons, Lei Du, et al. 2006. "Analysis of One Million Base Pairs of Neanderthal DNA." *Nature* 444 (7117) (November): 330–6.

Green, RE, A S Malaspinas, J Krause, A Briggs, P L F Johnson, C Uhler, M Meyer, et al. 2008. "A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing." *Cell* 134: 416–426.

Green, RE, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Jeffrey D Jensen, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22.

Gunz, P, S Neubauer, S Maureille, and J.J Hublin. 2010. "Brain Development after Birth Differs between Neanderthals and Modern Humans." *Current Biology* (20): 921–922.

Guo, Jia, Ning Xu, Zengmin Li, Shenglong Zhang, Jian Wu, Dae Hyun Kim, Mong Sano Marma, et al. 2008. "Four-Color DNA Sequencing with 3'-O-Modified Nucleotide Reversible Terminators and Chemically Cleavable Fluorescent Dideoxynucleotides." *Proceedings of the National Academy of Sciences of the United States of America* 105 (27) (July 8): 9145–50.

Hagelberg, E, MG Thomas, and EC Charles. 1994. "DNA from Ancient Mammoth Bones." *Nature* 370: 333–34.

Handt, O, M Krings, R. H. Ward, and S Pääbo. 1996. "The Retrieval of Ancient Human DNA Sequences." *Am. J. Hum. Genet.* 59: 368–376.

Hanni, Catherine, Vincent Laudet, and Dominique Stehelin. 1994. "Tracking the Origins of the Cave Bear (Ursus Spelaeus) by

Mitochondrial DNA Sequencing." *Proc. Natl. Acad. Sci.* 91 (December): 12336–12340.

Hawks, John. 2012. "Dynamics of Genetic and Morphological Variability within Neandertals." *Journal of Anthropological Sciences* 90 (January): 81–97.

Hawks, John. 2013. "Significance of Neandertal and Denisovan Genomes in Human Evolution*." *Annual Review of Anthropology* 42 (1) (October 21): 433–449.

Henn, Brenna M, Laura R Botigué, Simon Gravel, Wei Wang, Abra Brisbin, Jake K Byrnes, Karima Fadhlaoui-Zid, et al. 2012. "Genomic Ancestry of North Africans Supports Back-to-Africa Migrations." *PLoS Genetics* 8 (1) (January): e1002397.

Henn, Brenna M, L L Cavalli-Sforza, and Marcus W Feldman. 2012. "The Great Human Expansion." *Proceedings of the National Academy of Sciences of the United States of America* 109 (44) (October 30): 17758–64.

Herries, Andy IR. 2011. "A Chronological Perspective on the Acheulian and Its Transition to the Middle Stone Age in Southern Africa : The Question of the Fauresmith" 2011.

Higham, Thomas. 2011. "European Middle and Upper Palaeolithic Radiocarbon Dates Are Often Older than They Look: Problems with Previous Dates and Some Remedies." *Antiquity* 85: 235–249.

Higuchi, R, B Bowman, M Freiberger, OA Ryder, and AC Wilson. 1984. "DNA Sequence from the Quagga, an Extinct Member of the Horse Family." *Nature* 312 (5991): 282–4.

Hodges, Emily, Zhenyu Xuan, Vivekanand Balija, Melissa Kramer, Michael N Molla, Steven W Smith, Christina M Middle, et al. 2007. "Genome-Wide in Situ Exon Capture for Selective Resequencing." *Nature Genetics* 39 (12) (December): 1522–7.

Hodgson, Jason A, Christina M Bergey, and Todd R Disotell. 2007. "Neandertal Genome : The Ins and Outs of African Genetic Diversity." *Current Biology* 20 (12): R517–R519.

Hoffecker, John F. 2009. "The Spread of Modern Humans in Europe" 2009.

Hofreiter, M, V Jaenicke, D Serre, A Haeseler, and S Pääbo. 2001. "DNA Sequences from Multiple Amplifications Reveal Artifacts Induced by Cytosine Deamination in Ancient." *Nucleic Acids Research* 29 (23): 4793–4799.

Holliday, TW. 1997. "Postcranial Evidence of Cold Adaptation in European Neandertals." *American Journal of Physical Anthropology* 104 (2): 245–58.

Hoss, Matrhias, Amrei Dilling, Andrew Currantt, and Svante Paabo. 1996. "Molecular Phylogeny of the Extinct Ground Sloth Mylodon Darwinii." *Proceedings of the National Academy of Sciences* 93 (January): 181–185.

Howell FC. 1957. "The Evolutionary Significance of Variation and Varieties of 'Neanderthal' Man." *The Quarterly Review of Biology* 32 (4): 330–347.

Hublin, Jean-Jacques. 2014. "How to Build a Neandertal?" *Science* 344 (6190) (June 20): 1338–9.

Hublin, JJ. 1998. *Neandertals and Modern Humans in Western Asia*. Edited by Bar-Yosef O Akazawa T, Aoki K. Plenum, New York.

Hublin, JJ. 2009. "The Origin of Neandertals." *Proceedings of the National Academy of Sciences* 106 (38): 16022–16027.

Jobling, E Hollox, M Hurles, T Kivisild, and C Tyler-Smith. 2014. *Human Evolutionary Genetics*. Second. Garland Science.

Jobling, MA, ME Hurles, and C Tyler-Smith. 2004. *Human Evolutionary Genetics: Origins, Peoples and Disease*. London/New York.

Jónsson, Hákon, Aurélien Ginolhac, Mikkel Schubert, Philip L F Johnson, and Ludovic Orlando. 2013. "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters." *Bioinformatics (Oxford, England)* 29 (13) (July 1): 1682–4.

Jorde, L B, W S Watkins, M J Bamshad, M E Dixon, C E Ricker, M T Seielstad, and M a Batzer. 2000. "The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data." *American Journal of Human Genetics* 66 (3) (March): 979–88.

Kao, Wei-Chun, and Yun S Song. 2011. "naiveBayesCall: An Efficient Model-Based Base-Calling Algorithm for High-Throughput Sequencing." *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 18 (3) (March): 365–77.

Kao, Wei-Chun, Kristian Stevens, and Yun S Song. 2009. "BayesCall: A Model-Based Base-Calling Algorithm for High-Throughput Short-Read Sequencing." *Genome Research* 19 (10) (October): 1884–95.

Kircher, Martin, Udo Stenzel, and Janet Kelso. 2009. "Improved Base Calling for the Illumina Genome Analyzer Using Machine Learning Strategies." *Genome Biology* 10 (8) (January): R83.

Knapp, Michael, Mathias Stiller, and Matthias Meyer. 2012. "Generating Barcoded Libraries for Multiplex High-Throughput Sequencing." In , edited by Beth Shapiro and Michael Hofreiter, 840:155–170. Methods in Molecular Biology. Totowa, NJ: Humana Press.

Krause, Johannes, Paul H Dear, Joshua L Pollack, Montgomery Slatkin, Helen Spriggs, Ian Barnes, Adrian M Lister, Ingo

Ebersberger, Svante Pääbo, and Michael Hofreiter. 2006. "Multiplex Amplification of the Mammoth Mitochondrial Genome and the Evolution of Elephantidae." *Nature* 439 (7077) (February): 724–7.

Krause, Johannes, Carles Lalueza-Fox, Ludovic Orlando, Wolfgang Enard, Richard E Green, Hernán a Burbano, Jean-Jacques Hublin, et al. 2007. "The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals." *Current Biology : CB* 17 (21) (November 6): 1908–12.

Krause, Johannes, Ludovic Orlando, David Serre, Bence Viola, Kay Prüfer, Michael P Richards, Jean-Jacques Hublin, Catherine Hänni, Anatoly P Derevianko, and Svante Pääbo. 2007. "Neanderthals in Central Asia and Siberia." *Nature* 449 (7164) (October 18): 902–4.

Krings, M, C Capelli, F Tschentscher, H Geisert, S Meyer, A Von Haeseler, K Grossschmidt, G Possnert, M Paunovic, and S Pääbo. 2000. "A View of Neandertal Genetic Diversity." *Nature Genetics* 26 (2): 144–146.

Krings, M, A Stone, RW Schmitz, H Krainitzki, M Stoneking, and S Pääbo. 1997. "Neandertal DNA Sequences and the Origin of Modern Humans" 90: 19–30.

Lalueza-Fox, Carles, Elena Gigli, Marco de la Rasilla, Javier Fortea, and Antonio Rosas. 2009. "Bitter Taste Perception in Neanderthals through the Analysis of the TAS2R38 Gene." *Biology Letters* 5 (6) (December): 809–11.

Lalueza-Fox, Carles, Elena Gigli, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Jaume Bertranpetit, and Johannes Krause. 2008. "Genetic Characterization of the ABO Blood Group in Neandertals." *BMC Evolutionary Biology* 8 (January): 342.

Lalueza-Fox, Carles, and M Thomas P Gilbert. 2011. “Paleogenomics of Archaic Hominins.” *Current Biology : CB* 21 (24) (December 20): R1002–9.

Lalueza-Fox, Carles, Johannes Krause, David Caramelli, Giulio Catalano, Lucio Milani, María Lourdes Sampietro, Francesc Calafell, et al. 2006. “Mitochondrial DNA of an Iberian Neandertal Suggests a Population Affinity with Other European Neandertals.” *Current Biology : CB* 16 (16) (August): R629–30.

Lalueza-Fox, Carles, Holger Römpler, David Caramelli, Claudia Stäubert, Giulio Catalano, David Hughes, Nadin Rohland, et al. 2007. “A Melanocortin 1 Receptor Allele Suggests Varying Pigmentation among Neanderthals.” *Science (New York, N.Y.)* 318 (5855) (November): 1453–5.

Lalueza-Fox, Carles, Antonio Rosas, Almudena Estalrrich, Elena Gigli, Paula F Campos, Antonio García-Tabernero, Samuel García-Vargas, et al. 2011. “Genetic Evidence for Patrilocal Mating Behavior among Neandertal Groups.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (1) (January 4): 250–3.

Lalueza-Fox, Carles, María Lourdes Sampietro, David Caramelli, Yvonne Puder, Martina Lari, Francesc Calafell, Cayetana Martínez-Maza, et al. 2005. “Neandertal Evolutionary Genetics: Mitochondrial DNA Data from the Iberian Peninsula.” *Molecular Biology and Evolution* 22 (4) (April): 1077–81.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822) (February 15): 860–921.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3) (January): R25.

Leakey, LSB, PV Tobias, and JR Napier. 1964. “A New Species of Genus Homo from Olduvai Gorge.” *Nature* 202: 7–9.

Leakey, Walker A &. 1993. *The Nariokotome Homo Erectus Skeleton*. Harvard University Press.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14) (July 15): 1754–60.

Li, Heng, and Richard Durbin. 2011. “Inference of Human Population History from Individual Whole-Genome Sequences.” *Nature* 475 (7357) (July 28): 493–6.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16) (August 15): 2078–9.

Li, Heng, and Nils Homer. 2010. “A Survey of Sequence Alignment Algorithms for next-Generation Sequencing.” *Briefings in Bioinformatics* 11 (5) (September): 473–83.

Li, Heng, Jue Ruan, and Richard Durbin. 2008. “Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores.” *Genome Research* 18 (11) (November): 1851–8.

Li, Ruiqiang, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang. 2009. “SNP Detection for Massively Parallel Whole-Genome Resequencing.” *Genome Research* 19 (6) (June): 1124–32.

Li, Ruiqiang, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. 2009. “SOAP2: An Improved Ultrafast Tool for Short Read Alignment.” *Bioinformatics (Oxford, England)* 25 (15) (August 1): 1966–7.

Lindahl, T. 1993a. “Recovery of Antedilluvian DNA.” *Nature* 365 (365): 700.

Lindahl, T. 1993b. “Instability and Decay of the Primary Structure of DNA.” *Nature* 362: 709–14.

Lindgreen, Stinus. 2012. “AdapterRemoval: Easy Cleaning of next-Generation Sequencing Reads.” *BMC Research Notes* 5 (January): 337.

Liu, Shiping, Eline D Lorenzen, Matteo Fumagalli, Bo Li, Kelley Harris, Zijun Xiong, Long Zhou, et al. 2014. “Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears.” *Cell* 157 (4) (May 8): 785–94.

Lohmueller, Kirk E, Amit R Indap, Steffen Schmidt, Adam R Boyko, Ryan D Hernandez, Melissa J Hubisz, John J Sninsky, et al. 2008. “Proportionally More Deleterious Genetic Variation in European than in African Populations.” *Nature* 451 (7181) (February 21): 994–7.

Lordkipanidze, David, Tea Jashashvili, Abesalom Vekua, Marcia S Ponce De Leo, G Philip Rightmire, Herman Pontzer, Reid Ferring, et al. 2007. “Postcranial Evidence from Early Homo from” 449 (September): 305–310.

Lowery, R. K., G. Uribe, E. B. Jimenez, M. A. Weiss, K. J. Herrera, M. Regueiro, and R. J. Herrera. 2013. “Neanderthal and Denisova Genetic Affinities with Contemporary Humans: Introgression versus Common Ancestral Polymorphisms.” *Gene* 530: 83–94.

Lunter, Gerton, and Martin Goodson. 2011. “Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads.” *Genome Research* 21 (6) (June): 936–9.

Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa a Bemben, Jan Berka, et al. 2005. “Genome

Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057) (September 15): 376–80.

Maricic, Tomislav, Viola Günther, Oleg Georgiev, Sabine Gehre, Marija Curlin, Christiane Schreiweis, Ronald Naumann, et al. 2013. "A Recent Evolutionary Change Affects a Regulatory Element in the Human FOXP2 Gene." *Molecular Biology and Evolution* 30 (4) (April): 844–52.

Maricic, Tomislav, Mark Whitten, and Svante Pääbo. 2010. "Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products." *PloS One* 5 (11) (January): e14004.

McCown, T D, and A Keith. 1939. *The Stone Age Man of Mount Carmel: The Fossil Human Remains from the Levalloiso-Mousterian*. Clarendon Press, Oxford.

McDougall, I, HB Francis, and JG Fleagle. 2005. "Stratigraphic Placement and Age of Modern Humans from Kibish, Ethiopia." *Letters to Nature* 433 (February): 733–736.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9) (September): 1297–303.

Mellars, P. 1996. *The Neanderthal Legacy*. Princeton University Press, Princeton.

Mellars, P, and JC French. 2011. "Tenfold Population Increase in Western Europe at the Neandertal-to-Modern Human Transition." *Science (New York, N.Y.)* 333 (6042) (July 29): 623–7.

Mellars, Paul. 2010. "Neanderthal Symbolism and Ornament Manufacture: The Bursting of a Bubble?" 107 (47): 20147–20148.

Mendez, Fernando L, Joseph C Watkins, and Michael F Hammer. 2012. “A Haplotype at STAT2 Introgressed from Neanderthals and Serves as a Candidate of Positive Selection in Papua New Guinea.” *American Journal of Human Genetics* 91 (2) (August 10): 265–74.

Mendez, Fernando L, Joseph C Watkins, and Michael F Hammer. 2013. “Neandertal Origin of Genetic Variation at the Cluster of OAS Immunity Genes.” *Molecular Biology and Evolution* 30 (4) (April): 798–801.

Metzker, Michael L. 2010. “Sequencing Technologies - the next Generation.” *Nature Reviews. Genetics* 11 (1) (January): 31–46.

Meyer, Matthias, Qiaomei Fu, Ayinuer Aximu-Petri, Isabelle Glocke, Birgit Nickel, Juan-Luis Arsuaga, Ignacio Martínez, et al. 2014. “A Mitochondrial Genome Sequence of a Hominin from Sima de Los Huesos.” *Nature* 505 (7483) (January 16): 403–6.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G Schraiber, et al. 2012. “A High-Coverage Genome Sequence from an Archaic Denisovan Individual.” *Science (New York, N.Y.)* 338 (6104) (October 12): 222–6.

Mitchell, David, Eske Willerslev, and Anders Hansen. 2005. “Damage and Repair of Ancient DNA.” *Mutation Research* 571 (1-2) (April): 265–76.

Mullis KB, Faloona Fa. 1987. “Specific Synthesis of DNA in Vitro via Polymererase-Catalyzed Chain React.” *Methods Enzymol* 155: 335–50.

Nielsen, Rasmus, Joshua S Paul, Anders Albrechtsen, and Yun S Song. 2011. “Genotype and SNP Calling from next-Generation Sequencing Data.” *Nature Reviews. Genetics* 12 (6) (June): 443–51.

Noonan, James P, Graham Coop, Sridhar Kudaravalli, Doug Smith, Johannes Krause, Joe Alessi, Feng Chen, et al. 2006. “Sequencing and Analysis of Neanderthal Genomic DNA.” *Science (New York, N.Y.)* 314 (5802) (November 17): 1113–8.

Noro M, Masuda R, Dubrovo IA, Yoshu. 1998. “Molecular Phylogenetic Ingerence of the Wolly Mammuthus Primigenius, Based on Complete Sequences Mitochondrial Cytochrome B and 12S Ribosomal RNA Genes.” *J Mol E* 46 (3): 314–26.

Olalde, Iñigo, Morten E Allentoft, Federico Sánchez-Quinto, Gabriel Santpere, Charleston W K Chiang, Michael DeGiorgio, Javier Prado-Martinez, et al. 2014. “Derived Immune and Ancestral Pigmentation Alleles in a 7,000-Year-Old Mesolithic European.” *Nature* 507 (7491) (March 13): 225–8.

Olalde, Iñigo, Federico Sánchez-Quinto, Debayan Datta, Urko M Marigorta, Charleston W K Chiang, Juan Antonio Rodríguez, Marcos Fernández-Callejo, et al. 2014. “Genomic Analysis of the Blood Attributed to Louis XVI (1754-1793), King of France.” *Scientific Reports* 4 (January): 4666.

Orlando, Ludovic, Pierre Darlu, Michel Toussaint, Dominique Bonjean, Marcel Otte, and Catherine Hänni. 2005. “Revisiting Neandertal Diversity with a mtDNA Sequence.” *Current Biology* 16 (11): 400–402.

Orlando, Ludovic, Aurélien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, et al. 2013. “Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse.” *Nature* 499 (7456) (July 4): 74–8.

Ovchinniko, IV, A Götherström, GP Romanova, VM Kharitonov, K Lidén, and W Goodwin. 2000. “Molecular Analysis of Neanderthal DNA from the Northern Caucasus” 404 (March): 490–493.

Pääbo, S. 1985. “Molecular Cloning of Ancient Egyptian Mymmy DNA.” *Nature* 314: 644–645.

Pääbo, S. 1989. “Ancient DNA : Extraction , Characterization , Molecular Cloning , and Enzymatic Amplification.” *Proceedings of the National Academy of Sciences* 86 (March): 1939–1943.

Pääbo, S, and AC Wilson. 1991. “Miocene DNA Sequences - a Dream Come True?” *Curr Biol* 1 (1): 45–6.

Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. “Ancient Admixture in Human History.” *Genetics* 192 (3) (November): 1065–93.

Pedersen, Jakob Skou, Eivind Valen, Amhed M Vargas Velazquez, Brian J Parker, Morten Rasmussen, Stinus Lindgreen, Berit Lilje, et al. 2014. “Genome-Wide Nucleosome Map and Cytosine Methylation Levels of an Ancient Human Genome.” *Genome Research* 24 (3) (February 4): 454–466.

Pennisi, Elizabeth. 2010. “More Genomes From Denisova Cave Show Mixing of Early Human Groups.” In , 2010.

Peresani, M, I Fiore, M Gala, M Romandini, and A Tagliacozzo. 2011. “Late Neandertals and the Intentional Removal of Feathers as Evidenced from Bird Bone Taphonomy at Fumane Cave 44 Ky B.P., Italy.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (10): 3888–3893.

Pickering, Robyn, Paul H G M Dirks, Zubair Jinnah, Darryl J de Ruiter, Steven E Churchil, Andy I R Herries, Jon D Woodhead, John C Hellstrom, and Lee R Berger. 2011. “Australopithecus Sediba at 1.977 Ma and Implications for the Origins of the Genus Homo.” *Science (New York, N.Y.)* 333 (6048) (September 9): 1421–3.

Pickrell, Joseph K, Nick Patterson, Po-Ru Loh, Mark Lipson, Bonnie Berger, Mark Stoneking, Brigitte Pakendorf, and David Reich.

2014. “Ancient West Eurasian Ancestry in Southern and Eastern Africa.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (7) (February 18): 2632–7.

Poinar, Hendrik N. 2002. “The Genetic Secrets Some Fossils Hold.” *Acc. Chem. Res* 35 (8): 676–684.

Poinar, Hendrik N, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D E Macphee, Bernard Buigues, Alexei Tikhonov, et al. 2006. “Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA.” *Science (New York, N.Y.)* 311 (5759) (January): 392–4.

Prado-Martinez, Javier, Peter H Sudmant, Jeffrey M Kidd, Heng Li, Joanna L Kelley, Belen Lorente-Galdos, Krishna R Veeramah, et al. 2013. “Great Ape Genetic Diversity and Population History.” *Nature* 499 (7459) (July 25): 471–5.

Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, et al. 2014. “The Complete Genome Sequence of a Neanderthal from the Altai Mountains.” *Nature* 505 (7481) (January 2): 43–9.

Przeworski, M, RR Hudson, and A Di Rienzo. 2000. “Adjusting the Focus on Human Variation.” *Trends Genet* 16: 296-302.

Quinlan, Aaron R, Donald A Stewart, and T Marth. 2008. “Pyrobayes : An Improved Base Caller for SNP Discovery in Pyrosequences” 5 (2): 454–456.

Quintana-Murci, Lluís, and Andrew G Clark. 2013. “Population Genetic Tools for Dissecting Innate Immunity in Humans.” *Nature Reviews. Immunology* 13 (4) (April): 280–93.

Rae, Todd C, Thomas Koppe, and Chris B Stringer. 2011. “The Neanderthal Face Is Not Cold Adapted.” *Journal of Human Evolution* 60 (2): 234–239.

Raghavan, Maanasa, Pontus Skoglund, Kelly E Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, et al. 2014. "Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans." *Nature* 505 (7481) (January 2): 87–91.

Ramírez, Oscar, Elena Gigli, Pere Bover, Josep Antoni Alcover, Jaume Bertranpetit, Jose Castresana, and Carles Lalueza-Fox. 2009. "Paleogenomics in a Temperate Environment: Shotgun Sequencing from an Extinct Mediterranean Caprine." *PloS One* 4 (5) (January): e5670.

Rasmussen, Morten, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu, et al. 2010. "Ancient Human Genome Sequence of an Extinct Palaeo-Eskimo." *Nature* 463 (7282) (February 11): 757–62.

Reich, D, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327) (December 23): 1053–60.

Reich, D, Nick Patterson, Martin Kircher, Frederick Delfin, Madhusudan R Nandineni, Irina Pugach, Albert Min-Shan Ko, et al. 2011. "Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania." *American Journal of Human Genetics* 89 (4) (October 7): 516–28.

Reich, D, SF Schaffner, MJ Daly, G McVean, JC Mullikin, JM Higgins, DJ Richter, ES Lander, and D Altshuler. 2002. "Human Genome Sequence Variation and the G.e Influence of Ene History, Mutation and Recombination." *Nat Gnet* 32: 135–142.

RG, Klein. 1999. *The Human Career: Human Biological and Cultural Origins*. University of Chicago Press.

Rogers, Jeffrey, and Richard a Gibbs. 2014. "Comparative Primate Genomics: Emerging Patterns of Genome Content and Dynamics." *Nature Reviews. Genetics* 15 (5) (May): 347–59.

Rohland, Nadin, and Michael Hofreiter. 2007. “Comparison and Optimization of Ancient DNA Extraction.” *BioTechniques* 42 (3) (March): 343–352.

Rollo, Franco, and Isolina Marota. 1999. “How Microbial Ancient DNA , Found in Association with Human Remains , Can Be Interpreted How Microbial Ancient DNA , Found in Association with Human Remains , Can Be Interpreted.” *Phil. Trans. R. Soc. Lond. B* (January).

Romualdi, Chiara, David Balding, Ivane S Nasidze, Gregory Risch, Myles Robichaux, Stephen T Sherry, Mark Stoneking, Mark A Batzer, and Guido Barbujani. 2002. “Patterns of Human Diversity , within and among Continents , Inferred from Biallelic DNA Polymorphisms”: 602–612.

Rosas, Antonio, Cayetana Martínez-Maza, Markus Bastir, Antonio García-Tabernero, Carles Lalueza-Fox, Rosa Huguet, José Eugenio Ortiz, et al. 2006. “Paleobiology and Comparative Morphology of a Late Neandertal Sample from El Sidron, Asturias, Spain.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (51) (December): 19266–71.

Saiki, R.K, S Scharf, F. Faloona, K.B. Mullis, and G.T. Horn. 1985. “Enzymatic Amplification of Beta-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia.” *Science* 239: 1350–54.

Sánchez-Quinto, Federico, Hannes Schroeder, Oscar Ramirez, María C Avila-Arcos, Marc Pybus, Iñigo Olalde, Amhed M V Velazquez, et al. 2012. “Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers.” *Current Biology : CB* 22 (16) (August 21): 1494–9.

Sanger, F, and S Nicklen. 1977. “DNA Sequencing with Chain-Terminating” 74 (12): 5463–5467.

Sankararaman, Sriram, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. 2014. "The Genomic Landscape of Neanderthal Ancestry in Present-Day Humans." *Nature* 507 (7492) (January 29): 354–7.

Sankararaman, Sriram, Nick Patterson, Heng Li, Svante Pääbo, and David Reich. 2012. "The Date of Interbreeding between Neandertals and Modern Humans." *PLoS Genetics* 8 (10) (January): e1002947.

Santucci, F, BC Emerson, and GM Hewitt. 1998. "Mitochondrial DNA Phylogeography of European Hedgehogs." *Molecular Ecology* 7: 1163–1172.

Sawyer, Susanna, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Pääbo. 2012. "Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA." *PloS One* 7 (3) (January): e34131.

Scally, Aylwyn, and Richard Durbin. 2012. "Revising the Human Mutation Rate: Implications for Understanding Human Evolution." *Nature Reviews. Genetics* 13 (10) (October): 745–53.

Schadt, Eric E, Steve Turner, and Andrew Kasarskis. 2010. "A Window into Third-Generation Sequencing." *Human Molecular Genetics* 19 (R2) (October 15): R227–40.

Schmieder, Robert, and Robert Edwards. 2011. "Quality Control and Preprocessing of Metagenomic Datasets" (2010): 3–5.

Schubert, Mikkel, Luca Ermini, Clio Der Sarkissian, Hákon Jónsson, Aurélien Ginolhac, Robert Schaefer, Michael D Martin, et al. 2014. "Characterization of Ancient and Modern Genomes by SNP Detection and Phylogenomic and Metagenomic Analysis Using PALEOMIX." *Nature Protocols* 9 (5) (May): 1056–82.

Schubert, Mikkel, Aurelien Ginolhac, Stinus Lindgreen, John F Thompson, Khaled a S Al-Rasheid, Eske Willerslev, Anders

Krogh, and Ludovic Orlando. 2012. “Improving Ancient DNA Read Mapping against Modern Reference Genomes.” *BMC Genomics* 13 (January): 178.

Schuenemann, Verena J, Kirsten Bos, Sharon DeWitte, Sarah Schmedes, Joslyn Jamieson, Alissa Mittnik, Stephen Forrest, et al. 2011. “Targeted Enrichment of Ancient Pathogens Yielding the pPCP1 Plasmid of Yersinia Pestis from Victims of the Black Death.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (38) (September 20): E746–52.

Senut, B, M Pickford, D Gommery, P Mein, and K Cheboi. 2001. “First Hominid from the Miocene (Lukeino Formation , Kenya)” 332: 137–144.

Serre, David, André Langaney, Mario Chech, Maria Teschler-Nicola, Maja Paunovic, Philippe Mennecier, Michael Hofreiter, Göran Possnert, and Svante Pääbo. 2004. “No Evidence of Neandertal mtDNA Contribution to Early Modern Humans.” *PLoS Biology* 2 (3) (March): E57.

Shapiro, B, and M Hofreiter. 2014. “REVIEW A Paleogenomic Perspective on Evolution and Gene Function : New Insights from Ancient DNA” 343 (January): 1–7.

Shea, J.S., Brooks, A.S. 2000. *Mousterian*. Edited by Editors Delson E, Tattersall I, Van Couvering JA, Brooks AS. 2nd ed. Ne. Encyclopedia of human evolution and prehistory.

Skoglund, P., B. H. Northoff, M. V. Shunkov, a. P. Derevianko, S. Pääbo, J. Krause, and M. Jakobsson. 2014. “Separating Endogenous Ancient DNA from Modern Day Contamination in a Siberian Neandertal.” *Proceedings of the National Academy of Sciences* 111 (6) (January 27): 2229–34.

Skoglund, Pontus, and Mattias Jakobsson. 2011. “Archaic Human Ancestry in East Asia.” *Proceedings of the National Academy of*

Sciences of the United States of America 108 (45) (November 8): 18301–6.

Skoglund, Pontus, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M Thomas P Gilbert, Anders Götherström, and Mattias Jakobsson. 2012. "Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe." *Science (New York, N.Y.)* 336 (6080) (April 27): 466–9.

Skoglund, Pontus, Per Sjödin, Tobias Skoglund, Martin Lascoux, and Mattias Jakobsson. 2014. "Investigating Population History Using Temporal Genetic Differentiation." *Molecular Biology and Evolution* 31 (9) (June 17): 2516–2527.

Slimak, L, JI Svendsen, J Mangerud, H Plisson, HP Heggen, A Brugère, and PY Pavlov. 2014. "Late Mousterian Persistence near the Arctic Circle." *Science* 841 (2011).

Smith, J M, and J Haigh. 1974. "The Hitch-Hiking Effect of a Ourable Gene." *Genet Es*.

Stoneking, Mark, and Johannes Krause. 2011. "Learning about Human Population History from Ancient and Modern Genomes." *Nature Reviews. Genetics* 12 (9) (September): 603–14.

Stringer, C, and C Gamble. 1994. *In Search of the Neanderthals (London: Thames & Hudson)*.

Swisher III, CC, GH Curtis, T Jacob, A Getty, and AS Widiasmoro. 1994. "Age of the Earliest Known Hominds in Java, Indonesia." *Science* 263 (September 1993): 23–26.

Talbot, SL, and GF Shields. 1996. "Phylogeography of Brown Bears (Ursus Arctos) of Alaska and Paraphyly within the Ursidae." *Molecular Phylogenetics and Evolution* 5 (3): 477–494.

Tattersall, Ian. 1996. *The Last Neanderthal: The Rise, Success, and Mysterious Extinction of Our Closest Human Relatives*. Macmillan Publishing Company.

The International HapMap Consortium. 2003. "[International HapMap Project]." *Nihon Rinsho. Japanese Journal of Clinical Medicine* 426 (December): 29–34.

Trapnell, Cole, and Steven L Salzberg. 2009. "How to Map Billions of Short Reads onto Genomes" 27 (5): 455–457.

Trinkaus, E, and P Shipman. 1992. *The Neandertals: Of Skel- Etons, Scientists, and Scandal*. Random House, New York.

Vernot, B., and J. M. Akey. 2014. "Resurrecting Surviving Neandertal Lineages from Modern Human Genomes." *Science* 343 (6174) (January 29): 1017–21.

Virchow, R. 1872. *Untersuchung Des Neanderthal- Schädels*. Z. Ethnol.

Wagner, Günther A, Matthias Krbetschek, Detlev Degering, Jean-jacques Bahain, and Qingfeng Shao. 2010. "Radiometric Dating of the Type-Site for Homo." *Proc. Natl. Acad. Sci.* 107 (46): 19726–19730.

Wakeley, J. 2008. *Coalescent Theory*. Roberts & Company.

Wall, Jeffrey D, and Sung K Kim. 2007. "Inconsistencies in Neanderthal Genomic DNA Sequences." *PLoS Genetics* 3 (10) (October): 1862–6.

Wall, Jeffrey D, Melinda a Yang, Flora Jay, Sung K Kim, Eric Y Durand, Laurie S Stevison, Christopher Gignoux, August Woerner, Michael F Hammer, and Montgomery Slatkin. 2013. "Higher Levels of Neanderthal Ancestry in East Asians than in Europeans." *Genetics* 194 (1) (May): 199–209.

Wang, Shuoguo, Joseph Lachance, Sarah a Tishkoff, Jody Hey, and Jinchuan Xing. 2013. "Apparent Variation in Neanderthal Admixture among African Populations Is Consistent with Gene Flow from Non-African Populations." *Genome Biology and Evolution* 5 (11) (January): 2075–81.

Weaver, Timothy D. 2009. "The Meaning of Neandertal Skeletal Morphology." *Proc. Natl. Acad. Sci.* 106 (38): 16028–16033.

Weaver, Timothy D, Charles C Roseman, and Chris B Stringer. 2007. "Were Neandertal and Modern Human Cranial Differences Produced by Natural Selection or Genetic Drift ?" 53: 135–145.

White, TD, B Asfaw, Y Beyene, Y Haile-Selassie, CO Lovejoy, G Suwa, and LG WoldeGabrie. 2014. "Ardipithecus Ramidus and the Paleobiology of Early Hominids Tim." *Science* 64 (2009): 75–86.

Williams, Amy L, Suzanne B R Jacobs, Hortensia Moreno-Macías, Alicia Huerta-Chagoya, Claire Churchhouse, Carla Márquez-Luna, Humberto García-Ortíz, et al. 2014. "Sequence Variants in SLC16A11 Are a Common Risk Factor for Type 2 Diabetes in Mexico." *Nature* 506 (7486) (February 6): 97–101.

Wood B & Collard M. 1999. "The Human Genus." *Science* 284: 65–71.

Wood, Bernard. 1996. "Human Evolution." *BioEssays* 18: 945–954.

Wu, Hao, Rafael a Irizarry, and Héctor Corrada Bravo. 2010. "Intensity Normalization Improves Color Calling in SOLiD Sequencing." *Nature Methods* 7 (5) (May): 336–7.

Zilhão, João, Diego E Angelucci, Ernestina Badal-García, Francesco d'Errico, Floréal Daniel, Laure Dayet, Katerina Douka, et al. 2010. "Symbolic Use of Marine Shells and Mineral Pigments by Iberian Neandertals." *Proceedings of the National Academy of Sciences of the United States of America* 107 (3) (January 19): 1023–8.

Zilhão, João, Francesco D'Errico, JG Bordes, A Lenoble, JP Texier, and JP Rigaud. 2006. "Analysis of Aurignacian Interstratification at the Cha ˆ Telperronian-Type Site and Implications for the Behavioral Modernity of Neandertals EVOLUTION." *Proceedings of the National Academy of Sciences* 103 (33): 12643–12648.