UNIVERSITÀ DEGLI STUDI DI SASSARI

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE
*Direttore della Scuola: Prof. Franca Deriu*

INDIRIZZO DI GENETICA MEDICA, NUTRIGENOMICA E MALATTIE METABOLICHE
*Responsabile di Indirizzo: Prof. Francesco Cucca*
**XXVI Ciclo**

# A low-pass sequencing approach to phylogenetic analysis: reconstructing Sardinian and European demographic history with a panel of 1200 Y-chromosome samples

Direttore:
**Prof. Franca Deriu**
Tutor:
**Prof. Francesco Cucca**

Tesi di dottorato di:
**Dott. Riccardo Berutti**

Anno Accademico 2012/2013

# Contents

# Introduction

Phylogenetic analyses have constituted in the last years an important interdisciplinary topic between genetics, evolutionary studies and historical-antropological sciences.

Such studies can clarify the origins of contemporary populations and the demographic dynamics underneath. Studying the single nucleotide polymorphisms of the male specific portion of Y chromosome is particularly informative for these evolutionary analyses because of its lack of recombination and low mutation, reversion and recurrence rates.

All previous studies were hampered by partial genetic information or by a too small cohort of samples and by the lack of calibration points for the phylogenetic trees, which could have allowed a more precise estimate of the molecular change over time. The latest DNA sequencing technologies brought a great enhancement, since they can provide huge genomic datasets with an unprecedented speed and at constantly ecreasing costs.

A recent study was published considering 36 high coverage Y chromosome samples from different worldwide populations which was able to identify 6,662 phylogenetically informative variants [1] and was able to provide an estimation of the putative coalescence time for modern humans of between 101 and 115 thousands years ago. Such a study had a great benefit of the sequencing technologies and improved the number of the previously ($\sim$ 900) known phylogenetic markers significantly. Indeed it is still considering a small cohort of samples and has no informative calibration points for the study of coalescence.

Studying Sardinian population has been seen as particularly favourable for a number of reasons. Firstly, two major sequencing studies are under way which provided a huge number of both related and unrelated samples. Secondly, Sardinians, belonging to an isolate population, carry with themselves a set of private variants which can be related to archaeological findings, providing good calibration points for the phylogenetic tree and to the calculation of mutation rates.

In this work we will present a phylogenetic analysis on the low-pass sequences of 1,204 Sardinian male samples along with timing calculations performed over the tree. We will illustrate our calcolated putative coalescence age for our dataset, and compare it to other relevant historical events [46].

We will present the architecture developed with original contribution of the author during the last years for the analysis of the two big Sardinian sequencing projects aforementioned, whose data and analysis tools were used and furtherly improved for this work.

# Chapter 1

# Phylogenetics and Y chromosome

## 1.1 What is phylogenetics

Phylogenetics studies the origin and the evolutionary relationships between organisms. These can be a set of individuals of the same species, or belong to different ones. Phylogenetics comes from the Greek words *phylon* ( *φῦλον*, meaning clan) and *genesis* ( *γένεσις*, meaning origin).

These studies analyze observable parameters to reconstruct the evolutionary steps linking the organisms under examination and their most recent common ancestor (MRCA), shaping the achieved information into evolutionary trees.

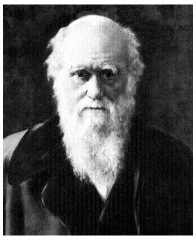
Tree branches can express time-associated quantities, in that case, it is possible to infer the time interval separating the analyzed samples from their most recent common ancestor, the time-to-MRCA (TMRCA).

The result of a phylogenetic analysis takes the name of phylogeny.

### 1.1.1 Some history

**Evolutionary theory**

Scientists and philosophers have always perceived, by observing common traits or features among species, that some relationship could be established among them. In the last three centuries science started realizing that those traits seemingly followed an evolutionary path, carrying species to fit best their own living environment and that evolution was still an active process. Building back the path to the roots of life became a thrilling challenge.



*Charles Darwin (1809-1882)*

In 1859 Darwin, into his "Origin of Species", explained his evolutionary theory and firstly introduced the concept of common ancestor and, thus, of coalescence: at a certain point in the past evolutionary lines of the various species, merge to a common ancestor.

In 1866, Haeckel proposed the first-ever phylogenetic tree (figure 1.1), which was based on similarities and speculative observations. His idea of evolution was wrong. By his "recapitulation theory" he proposed that every stage of embryonic development corresponded to evolutionary stages. Despite that, his work was fundamental, since it introduced the concept of evolutionary tree as representation of a logical sequence of events.

In 1893 Dollo enunciated his law of character state irreversibility, which states that once a complex trait is loss,it cannot re-evolve back to its previous state, or vice versa. This statement implies that a tree-like model is a realistic representation of evolutionary processes.

### Introduction of mathematical approach

*Ronald Fisher (1890-1962)*

With the introduction into biological and life sciences of new statistical concepts, (such as maximum likelihood (ML), analysis of variance, fiducial inference) by Ronald Fisher in the first decades of the twentieth century, strong foundations were laid for a systematic approach to population genetics.

In the 1940's new and old concepts were systematized with the foundation of modern cladistics by German scientist Hennig. Trees were intended, before, as tentative representations of a pathway. With cladistics quantitative information was embedded into the trees, which could be used for time calculations.

The first possible mathematical approach to infer evolutionary relationships among organisms was based on phenotypical and/or morphological data matrices. It was thus fundamental to translate observations into quantitative information, even binary, but numerable: e.g. number of eyes, has legs (yes=1, no=0), but even size measurements (height, weight) blood parame-

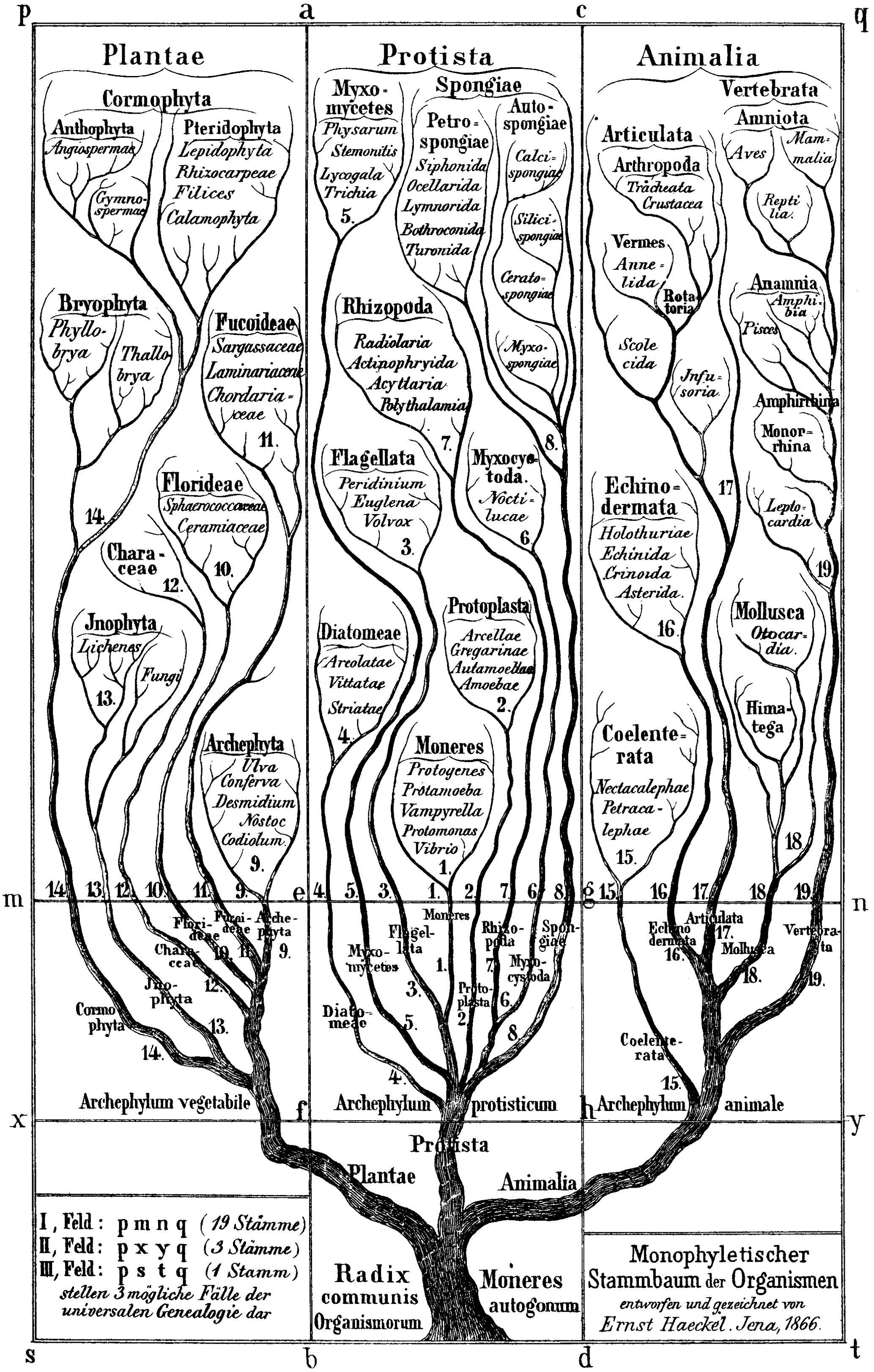


Figure 1.1: Haeckel's tree

ters, etc.

In the 1960's, notably, the first attempt to apply ML to phylogenetics was made by Cavalli-Sforza and Edwards, a former Fisher's student [3]. In the same years maximum parsimony and other similar criteria were introduced. All of them integrated a fundamental principle, the best tree will be the one explaining the evolutionary variation with the minimal variational paths between samples.

### Molecular approach

In the most recent years molecular techniques for genomic analysis gave a strong improvement to the possibility to map human genome variation. Molecular data reading capabilities evolved from microsatellite measurements to detection of point mutations (SNP) and fine structure variation (insertions, deletions, copy number variation).

**SNP**: single nucleotide polymorphism, a single nucleotide changes its type at a genomic locus

**insertion**: sequences of one or more bases are added to the genome

**deletion**: sequences of one or more based are removed from the genome

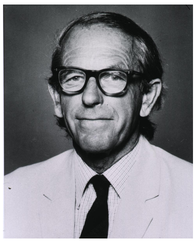
**CNV**: copy number variation, a gene or a region can be present in multiple copies, the number of copies is a genomic structural variation

Phylogenetic analyses could take amazing advantage from molecular data: with respect to phenotypical data matrices they are not obfuscated by non-genetic effects. Such a technical framework allowed to directly use statistical methods to relate measurable quantities which are clearly inherited with respect to some of the phenotypical traits. With the advent of the most recent genotyping and sequencing technologies it has become possible to descend to the mutation level, giving unprecedented information on the genetic bases of evolution and laying the bases to explain the differences that stay behind taxonomical classifications.

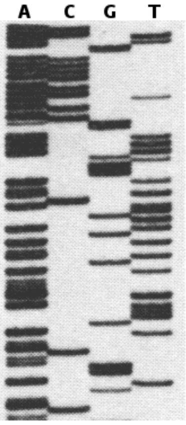
## 1.2 Technical background to our study

Molecular phylogenetics encompasses two main approaches. Structural variation based with microsatellites and SNP based with either Sanger sequencing, genotyping with DNA microarrays and genotyping with next generation sequencing.

Historically genotyping and sequencing technology evolved from Sanger sequencing, to microarray DNA genotyping, to next-generation sequencing.



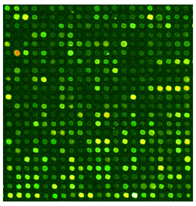
*Frederick Sanger (1918-2013)*



*Sanger gel example (UCSF Sch.of Medicine)*

- **Sanger sequencing**: a sequencing technique developed by Frederick Sanger in 1977 [5]. It is based on incorporation of small quantities of chain-terminating dideoxynucleotides during in-vitro DNA replication of a DNA template started by a specific primer. Dideoxynucleotides are deoxynucleotides lacking the 3'-OH group necessary to form a bond between two nucleotides. Once incorporated by DNA polymerase they stop the extension. Four separate reactions are performed adding A, G, C, T marked dideoxynucleotides. Due to the stochastic nature of the process fragments of different sizes will be obtained, each one blocked at an A, C, G or T depending on the reaction. The fragments are size selected, generally made run on an agarose gel, and the sequence could be read as depicted in figure. Automated techniques were later developed with proper Sanger sequencing machines. This technique allows the sequencing of fragments up to 700 bp. The human genome project was carried out with it. It is a slow method, only one fragment can be sequenced at a time, it is necessary to know what to sequence and has very high costs per base pair sequenced. On the other side it is extremely reliable, thus is still widely in use e.g. for SNP discovery confirmation.

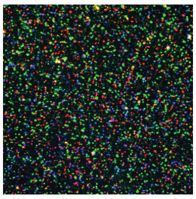
- **Genotyping with DNA Microarrays**: it is a genotyping technique allowing to query a number of known SNPs. A number of DNA known sequences, named probes, are attached into known positions to a solid surface, generally glass or silicon. DNA target sequences, from the sample containing the SNPs to query, are fluorescently labeled and made hybridize with the probes. The higher the target and the probe sequence match, the tighter the chemical bond will be. A wash-out is performed, which will wash almost all totally mismatching sequences, a negligible amount of the totally matching and a variable amount of the sequences containing one or more SNPs. An optical acquisition is performed, the intensity measured will be proportional to the allelic status, for a diploid genome it will be maximum for homozygous match, low for homozygous mismatch and halfway for heterozygous



*Example of a microarray matrix (Yale Univ.)*

SNPs. Generally for each known SNPs two probes (one carrying reference and the other carrying derivate allele) are present. This technique is fast and low cost, as a drawback it can read only a limited number of known markers that can be queried, thus is completely blind to new variation.

- **Next generation sequencing**: it is the most recent, reliable technique [7] to read an entire human genome by sequencing its fragments. DNA fragments to be sequenced are attached to a solid surface. Then consecutive cycles of 1)incorporation of A,C,G,T nucleotides with a reversible terminator and marked each one with a different fluorescence 2)optical acquisition and 3)cleavage of the terminal blockers performed. Sequences up to 300 nucleotides (latest instruments) can thus be acquired. We will describe extensively the method in the dedicated chapter 2. Next generation sequencing is the preferable method to sequence entire genomes. It has a low cost per base, high throughput. On the other side has the higher error rate among the genotyping approaches.



*Example of NGS clusters photo (Illumina)*

Next generation sequencing approach is the fastest and the most complete for our purposes, it will be treated in more detail in the following chapter. A crucial point about the difference between microarray genotyping and sequencing is that the first one, includes only known markers, thus it cannot be used as a tool for discovery of new phylogenetically informative variation. This ascertainment bias would also impair the possibility of attaching reliable dates to demographic events. It will also be shown how genotyped samples proved useful as a control step for the assesment of sequencing data quality.

## 1.3 SNP-based phylogeny and molecular clock

Our approach to phylogeny is SNP based. Given the genotypes of the samples in the dataset, we aim at finding a sequential hierarchy of mutations which, going backwards in time, could coalesce to a unique root.

Since we are analyzing samples of the same species *homo sapiens*, this introduces a simplification: we can directly compare genotypes at known (or newly spotted) loci among samples. In fact chromosomal structure does not change among individuals. Otherwise it would have been necessary to compare genes and to seek for homologous structures.

On sexually reproducing organisms each offspring contains a mix of the genetic material of the two parents, that means that we generally do not have a straight line connecting out samples to their ancestors. Very complicated mixing may happen as a consequence of migrations. Even if temporary, inbreeding traces can be heavy. With an elementary observation, to explain how intricated the scenario could be, we can just note that each of us has 2 parents, 4 grandparents, 8 greatgrandparents and so on. This exponential increase, with a supposed generation time of 30 years leads to an apparently paradoxical number of ancestors, just 1,000 years ago, larger than actual world population. It is straightforward that the same people appear more than once in the lines.

In a human genome two lines make an exception to recombination. They are Y-chromosome and mitochondrial DNA. These are respectively paternally and maternally inherited. Using mtDNA and Y chromosome it is possible to provide, with technology enhancements always more accurate estimates on the origins of paternal and maternal lineages.

Considering mutation rate and generational steps as constants (on average) over time, we can use genome data to map the timing of evolutionary events. This is the *molecular clock* hypothesis, which implicitly needs to be valid that every branch accumulated the same number of mutations over time.

In our study we analyzed the male specific line using the Y chromosome. We will briefly illustrate mtDNA and more extensively Y chromosome in the following paragraphs.

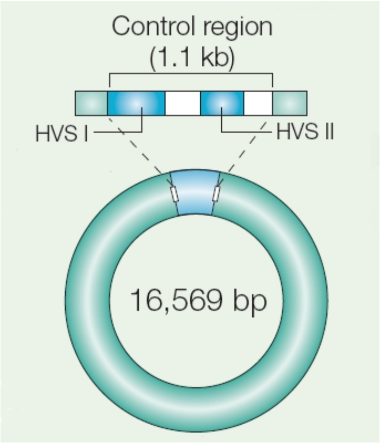


Figure 1.2: mtDNA molecule schematics, HVS1 and HVS2 regions are evidenced

## 1.4 Mitochondrial DNA

### 1.4.1 Structure

Human mitochondrial DNA (mtDNA) is a DNA molecule carried by the mitochondria, organelles carried by the cells whose role is generating most of the adenosyne triphosphate (ATP) available to the cell. Mitochondrial DNA is a small, circular, double stranded molecule. It measures slightly less than 16.6 thousands base pairs and codes for 37 genes. Its expression has a central role in ATP synthesis, thus mtDNA is an essential component for sustaining life. Due to its small size, historically this DNA has been the first complete portion of human genome to be completely sequenced

Mitochondrial DNA is inherited on a maternal line, since mitochondria carried by sperm generally do not enter the egg and, in case they do, their genome is marked with ubiquitin for degradation, thus they do not contribute to the offspring's mtDNA. Every single human cell can carry thousands of mitochondria, each one with its genome. Thus it may happen that a single individual carries mitochondria with different (inherited) genotypes: this phenomenon is named heteroplasmy. In figure 1.2 it is shown, in blue, the non-coding portion of the mtDNA, named control region. Inside this region lie two small portions named Hyper-Variable-Sections (HVS) 1 and 2. These sections show a high mutation rate, thus they have been the first to be subject to phylogenetic analysis.

The most recent common ancestor of the mtDNA of living humans is named Mitochondrial Eve. While this name derives from the Bible's Eve, it should not mislead us. She was not the unique living woman, but she is the unique woman whose mtDNA maternal line was not interrupted until today.

Tentative TMRCAs on mtDNA, date Mitochondrial Eve in a wide range among 120-200 thousands years BP [8][9][10]. It is a still opened debate.

## 1.5 Y chromosome

Human Y chromosome is one of the two gender determining chromosomes, X and Y. The Y is carried in one copy along with one copy of X in male individuals, while female individuals carry two X with no Y. The Y chromosome is the only one in the nuclear DNA which is not recombining, since it is supplied entirely by the father. Only small known regions can recombine with the X, thus it is mostly inherited unchanged on paternal lines.

The absence of recombination is an important point which sets us free from the confounding effects that arise not only from recombination between members of the same population but from inbreeding/migration effects that may alter the genetic heritage of a single population: i.e. we can find something out-of-place, but it would be an entire Y chromosome and not parts of it.

About the Y, it is worth to be noted that in case of syndromes causing more copies (XYY, XYYY syndromes, the first one being frequent $\sim 0.1\%$), these copies will theoretically recombine, but the recombination would happen between identical copies and will have no effect on the sequence.

The advantages of using Y chromosome for phylogenetic analysis are quite clear. With respect to mtDNA it is carried in just one copy, thus no we don't have heteroplasmy and not even heterozygosity. The drawback of that is that, being just one copy, it is more difficult to sequence (it will be furtherly explained in chapter 2) and, as it is longer, the analyses are more complicated

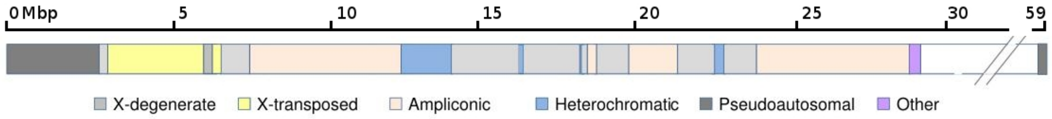


Figure 1.3: Y Chromosome structure and its regions

and need a computing infrastructure.

The lack of confounding effects is a major advantage: no recombination and low mutation, reversion and recurrence rates make Y chromosome suitable for phylogenetic analyses [2].

ISOGG: www.isogg.org

At the time we started our project (2011) just around one thousand variants defining Y chromosome evolutionary history were annotated at the ISOGG (International SOciety for Genetic Genealogy) [13].

### 1.5.1 Structure

Chromosome Y is among the shortest in human genome, spanning 59.4 Mbp. Its structure is made of discrete sequence classes [11]. The main constituting regions are (see fig.1.3):

- X-degenerate:
  a set of sub-regions which has no analogy with the other sex chromosome, the X.

- X-transposed:
  a set of regions with very high (99%) homology with the X

- Pseudo autosomal (PAR):
  a set of two small regions (PAR1 and PAR2), located into the telomeres,which actually recombine with the X, thus the pseudo-autosomal name.

- Ampliconic:
  a set of regions containing palindrome and repeated sequences

- Heterochromatic:
  regions which are depleted of genes

The non-recombining, non-ambiguous regions of the Y chromosome as a group are named male-specific-Y (MSY) / non recombining Y (NRY).

### 1.5.2 Mutation accumulation

Since MSY regions are transmitted from father to son with no recombination, their variants tend to accumulate and differentiate over time as generations branch, as shown in figure 1.4.a (page 14).

A logical hierarchy links the terminal branches to a common ancestor. Y lines are monophiletic: each offspring is linked through a straight line to his ancestors. Only de-novo mutations can change the genotype from father to son.

Some mutations define the branches from the main trunk and some others differentiate sub-branches from the main branches. It is important to understand nomenclature and classification. This information will be introduced in the following paragraph.

### 1.5.3 Haplogroups and nomenclature

Phylogenetics classifies individuals assigning them to their haplogroups. An haplogroup is a group of haplotypes which are defined by shared variants, whose evolutionary history is defined, and are inherited from a common ancestor. Each variant which marks the difference between two haplogroups is named marker.

In figure 1.5 and in the following text it is briefly explained how classification works according to the nomenclature promoted by the Y-Chromosome-Consortium (YCC) [12].

The main branches of the tree, and recursively all the sub-branches, are the haplogroups (and sub-haplogroups). Main branches are indicated with capital letters, second level branches with progressive numbers (1,2,3,...), third level with progressive lowercase letters (a,b,c..) and so on alternatively numbers and letters.

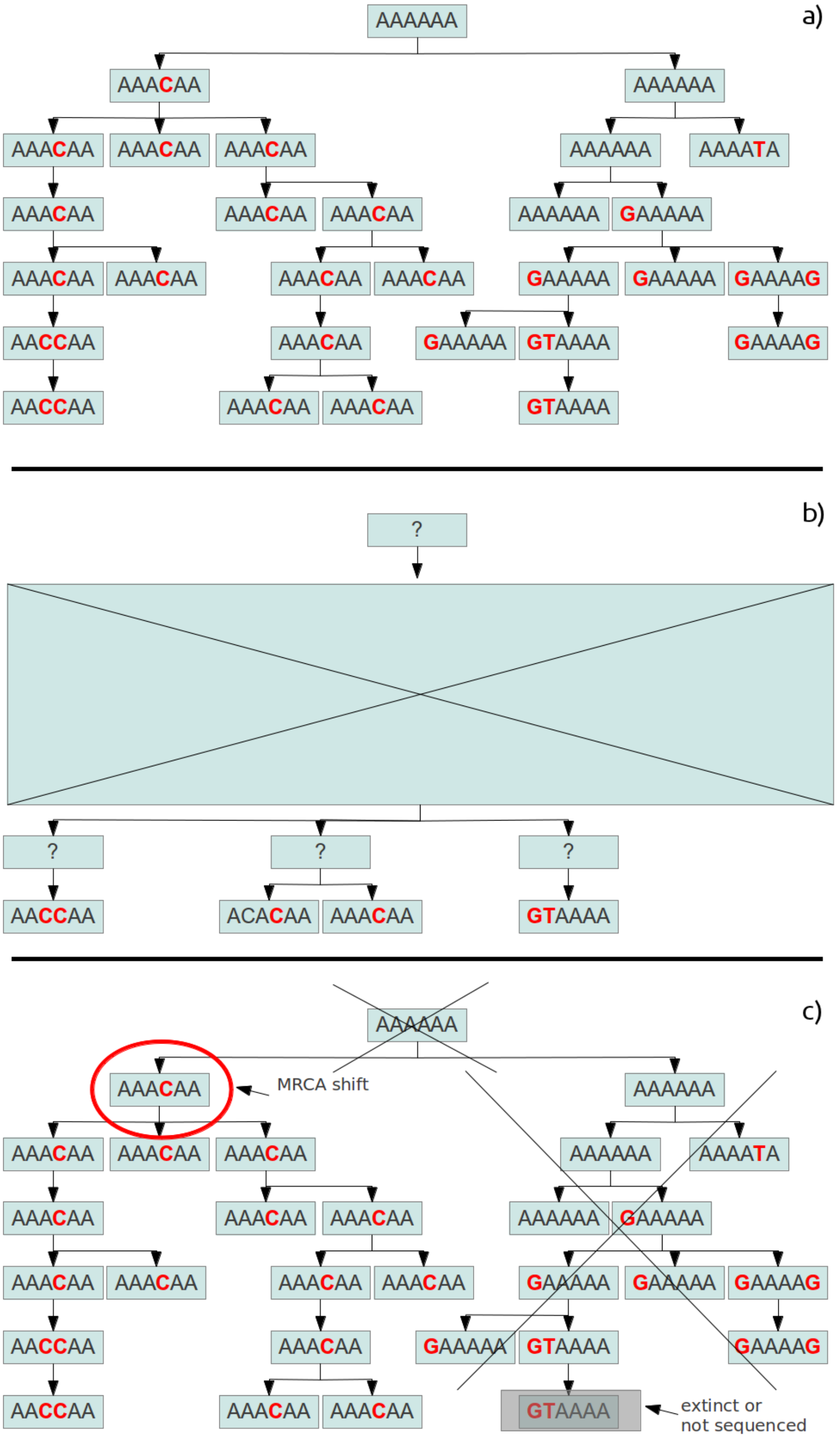


Figure 1.4: **a)** Hypothetical variant accumulation on a Y chromosome genotype *AAAAAA* over time; **b)** Real data, only living samples available and, occasionally some parental information; **c)** A sampling bias or an extinct member of the tree may impair our ability to reach the true most recent common ancestor of all the lines. Thus we may only refer to the most recent common ancestor of a dataset

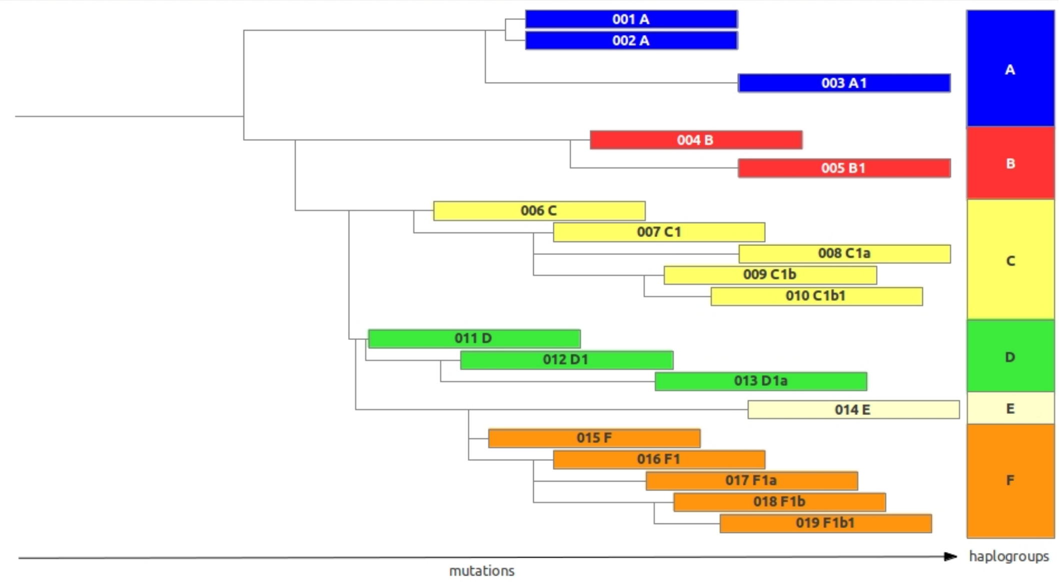


Figure 1.5: Haplogroup nomenclature. Each box represents a sample, with an arbitrary sample number and an explicative haplogroup attribution. Nomenclature is hierarchical and constituted by a capital letter defining main haplogroup and an alternative succession of numbers and lowercase letters to indicate further sub-branchings.

As an example, the individual “I” is part of the haplogroup Z. Thus he carries the mutations that define Z. If he also carries the mutation defining the sub-branch 2 and the sub-sub-branch b will be classified as Z2b.

Markers are usually named with a capital letter and a following number (points or underscores may be into the name).

## 1.6 Phylogenetic analysis results

The resulting information from a phylogenetic analysis is generally a tree in the form of a dendrogram which reconstruct the evolutive/mutational history of a group of people. Supposing as the real population tree the one shown in fig. 1.4.a in the real case we only have access to the individuals belonging to the living generation. Occasionally we can also have some accessory information about the past, but generally we cannot access ancestors DNA. Thus the input data available has the form shown in fig. 1.4.b. Several lines developed and then lost, and on the present generation, no hint remains of their existence.

While losing these lines doesn't alter the route to the most recent common ancestor (MRCA), if we lose the sample marked in fig. 1.4.c by sampling bias or by extinction of the line, the common ancestor will switch from the one with the dummy genotype $AAAAAA$ to the one with the dummy genotype $AAACAA$.

In this case we have lost all information about the first generation, and we have identified as first generation what is, in reality, the second.

Constructing the phylogenetic dendrogram means reconstructing within how many mutational steps two samples converge to their common ancestor (coalescence), and so on to obtain the common ancestor of the entire dataset that we are examining (the coalescence time for the samples).

As for the mitochondrial case, the MRCA of all living humans is named Y-chromosomal Adam. Also in this case it doesn't mean that he was the first, unique man, but he was the only one whose paternal line survived without being interrupted up to today.

A mutation rate can be calculated, examining de-novo mutation occurrence through pedigrees or by exploiting our knowledge of archaeological or demographic events that can be placed on the tree (this will be discussed in detail), associating genetic distance with evolutionary time-span. The time in the past when the coalescence point is located is named time to most recent common ancestor (TMRCA).

From what considered above, any missing line may cause a too recent estimate of the ancestor of the samples. This problem could be mitigated by enlarging the dataset as much as possible and using appropriate outgroups that may carry more ancient alleles.

It is essential to clarify that any conclusion that may be taken out of the tree will be more in accordance with the true evolutionary tree the bigger the dataset is. But, rigorously it can only be stated that conclusions are valid within the scope of the samples included in the dataset.

It is also fundamental to keep into account that the reference genome, against which we evaluate mutations is not representative of the roots of evolution but is built from recent samples which belong to their respective haplogroups.

Thus for many loci the original alleles for *homo sapiens*, named ancestral alleles, could be different. This makes necessary to use an additional ancestor outgroup to clarify which allelic states are ancestral and which are derivate: the chimpanzee sequence will has been used in our work.

# Chapter 2

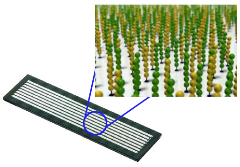
# Next Generation Sequencing

Next Generation Sequencing (NGS) is a technique allowing to read the content of genomic material. Its main features are high throughput at a relatively low price per base sequenced. As a drawback, there's a high initial cost for the machine, a non-negligible error rate and the need of a specialized IT infrastructure and bio-informatics know-how to analyze data. In this chapter it will be made reference to the Illumina[14] sequencing technology we used in our study.

**Illumina**: illumina.com

## 2.1 How does it work

After DNA is extracted, it is sheared and fragmented by means of sonication. According to Illumina sample prep protocol, fragments are ligated to adapters ($\sim$ 130bp long) that will bind the flow-cell, the medium which holds DNA molecules for sequencing (fig. 2.1.1).



*Flow-cell and clusters (Illumina)*

Fragmented DNA is size selected with electrophoresis: DNA is made run into an agarose-gel under an electric field which will move molecules at different speeds according to their mass. By means of proper markers the gel is cut to select fragments of an average length of 500 bp (Gaussian distribution) which means around 370 bp long DNA reads. Then a PCR enrichment step is performed. The fragment pools are then loaded into the flow-cell with an instrument called cBot. Each flow-cell is a glass support containing 8 channels named lanes which are filled with a matrix of adapters

**PCR**: polymerase chain reaction, a molecular technology to amplify (exponential multiplication) DNA fragments to ligate the fragments. Lanes are readable independently. Thus up to 8 independent samples (or mixes) can be poured for sequencing.

Once fragments ligate the adapters on the flowcell they bridge, extend and thus duplicate original DNA fragment, which on the other side contains another adapter which is recognized by other adapters on the flowcell. This process goes on several times, and multiple copies are locally created around the original fragment. This takes the name of bridge PCR. It allows to create spots of identical copies of the DNA fragments that we want to sequence. Before starting sequencing DNA is denaturated obtaining clusters of single stranded DNA fragments. All this process is named cluster generation and is illustrated in figure 2.1.2

The fragments are then ready to be sequenced. The chemical process is called sequencing-by-synthesis (SBS), see figure 2.1.3.

The process is performed in cycles. At the $n^{th}$ cycle every fragment of every cluster will be attached one fluorescently labeled nucleotide complementary to the $n^{th}$ base starting from the adapter. The fluorescent nucleotide has a reversible terminator which blocks further extension of the DNA fragment.

The flowcell is then illuminated with lasers with different wavelengths corresponding to the excitation energies of fluorescent labels carried by the incorporated bases. Each of the four nucleotides has its own different fluorescence wavelength.

For each of the four wavelengths a photo is taken and software processed. For every cycle the software calculates the coordinates of the clusters and their base content.

Fluorescence is then cleaved and also the terminator is unblocked. The $(n+1)^{th}$ cycle starts, with a new layer of fluorescent nucleotides that comes to read the $(n+1)^{th}$ base in the sequences contained into the clusters.

After 100 cycles (in our settings) a stack of base-reads for each cluster is collected, which linearized, gives the 3'-5' sequence of the first 100 bp of the fragments.

After that step we did the so-called "pair-end", another bridge PCR is performed and the fragments are reversed and another 100 cycles or sequencing are performed.

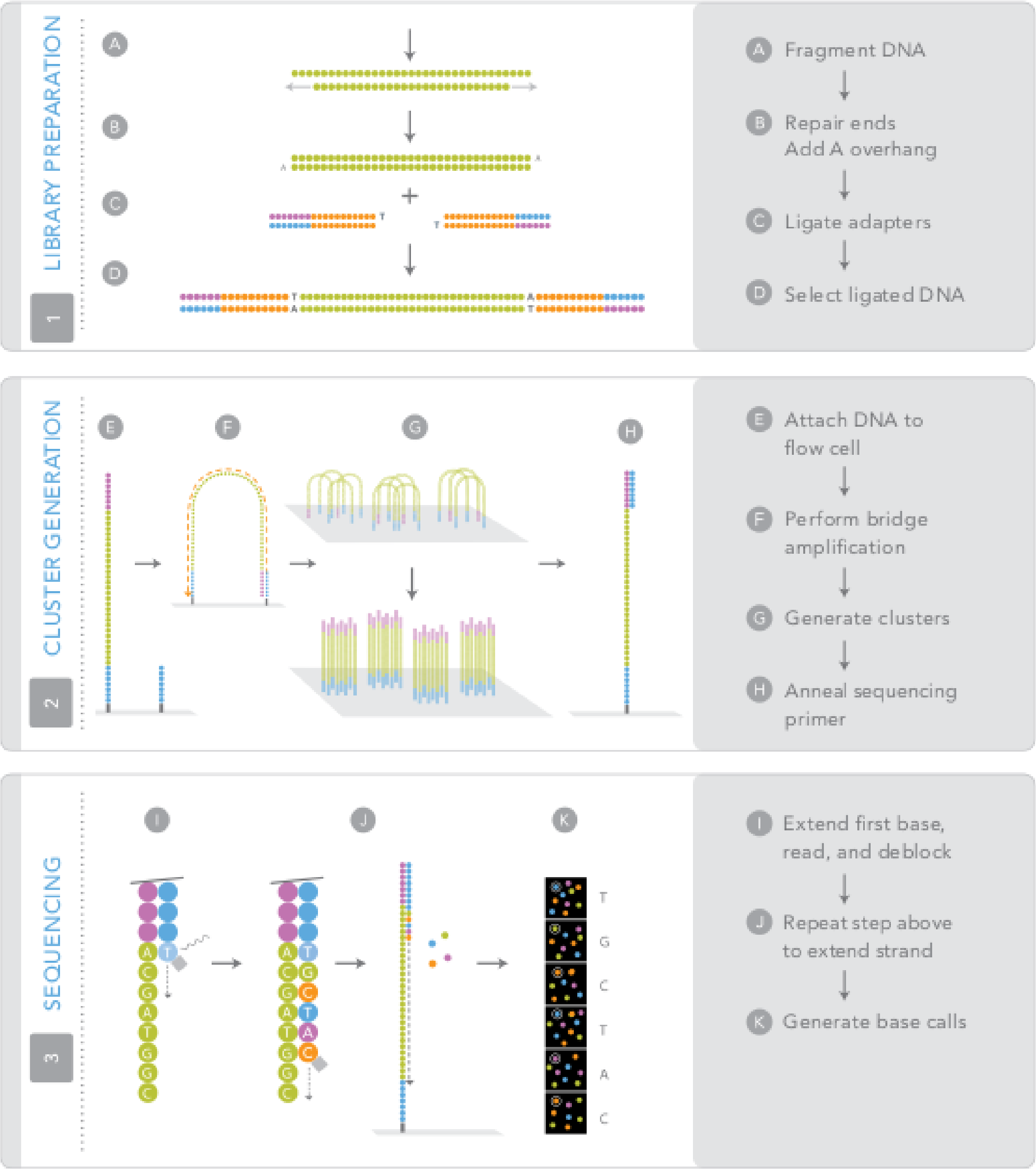


Figure 2.1: Illumina sequencing steps, 1) Library preparation, 2) Cluster Generation, 3) Sequencing (Source Illumina)

The sequence of the first 100bp 3'-5' of each fragment is named read1, the sequence of the 100bp 5'-3' is named read2.

## 2.2 Machines and special features

Our approach has used the Genome Analyzer IIx and the HiSeq 2000 sequencing machines from Illumina.

**Genome Analyzer IIx**: illumina.com/systems/ genome_analyzer_iix.ilmn

**HiSeq2000**: illumina.com/systems/ hiseq_2000_1000.ilmn

These are consecutive advancements of the same technology that has been described before, with improvements in chemistry efficiency, optical quality.

HiSeq 2000 machines can sequence up to two flow-cells at once, with 8 lanes each in around 11 days. With the latest chemistry each of the flowcells can read up to 300 Gbp (that's 100 human genomes), that's 38 Gbp of data per lane.

Each lane can collect one single sample or a mix of samples. For this second purpose Illumina provide some kits which are able to attach a 7 nucleotide sequence (namely 'barcode') to the fragments to make them recognizable. The barcode can be found on one side or on both sides, allowing the user to mix up to 96 samples onto a single lane.

## 2.3 Output

Resulting data is supplied in homologous couples of read 1 and read 2, the sequences of the fragments' extremales. The number of reads is very high. For a 1X human genome, with its $3 \cdot 10^9$ bp length we get something like $3 \cdot 10^7$ reads of 100 bp length each Due to the fragmentation process reads are mixed and the original genomic position they belonged to is not known. Thus further processing, named alignment is necessary to identify sequences, assigning to the short reads a position with respect to a genomic reference. In the analysis chapter 4 we will illustrate how is it possible to deal with alignment and, in general with such a "big-data" scenario.

## 2.4 Quality and coverage

### Coverage

A complete human genome measures around ∼3 Gbp. In human resequencing it is common to measure the coverage as the number of bases sequenced divided by these ∼3 Gbp. Thus if we get ∼6 Gbp out of our machine, we can tell that we have sequenced with a 2X (times) coverage. These 2X represent the average read depth for each genomic locus. For a diploid genome like human's we theoretically need a 2X coverage to see all the variants and their heterozygous allelic states.

This is not really true. An average coverage of 2X does not guarantee that depth is uniform. Regions which are most difficult to amplify, will surely be missed and, indeed, the read depth is not uniform at all. More, with low coverage, the effects of alignment (see 4), will perform badly with highly mutated regions.

In figure 2.2 is illustrated the difference in coverage depth for two different sequencing runs, of the same sample, for the same region, with a 15x coverage and a 2x. Even with a 15x coverage the depth of genome covered is not uniform, but with a 2x, depending on our scope for sequencing, we have a huge number of holes caused by non-uniform coverage. The price of increasing the coverage is proportional to the gain that is obtained.

### Quality

Quality of bases and interpretation of quality is of fundamental importance. An incorrect interpretation or an underestimation of their effects may lead to nonsense results.

A modern instrument can reach a precision, on the single base pair better than $10^{-3}$. This means less than one wrong base call every thousand. Such a low error rate seems quite impressive, and it is. But depending on the application it can still lead to huge error rates. We'll examine it in detail in the analysis chapter 4.

Another quality to be kept under control is the mapping quality which accounts of probability that the recognition of the read is successful. This

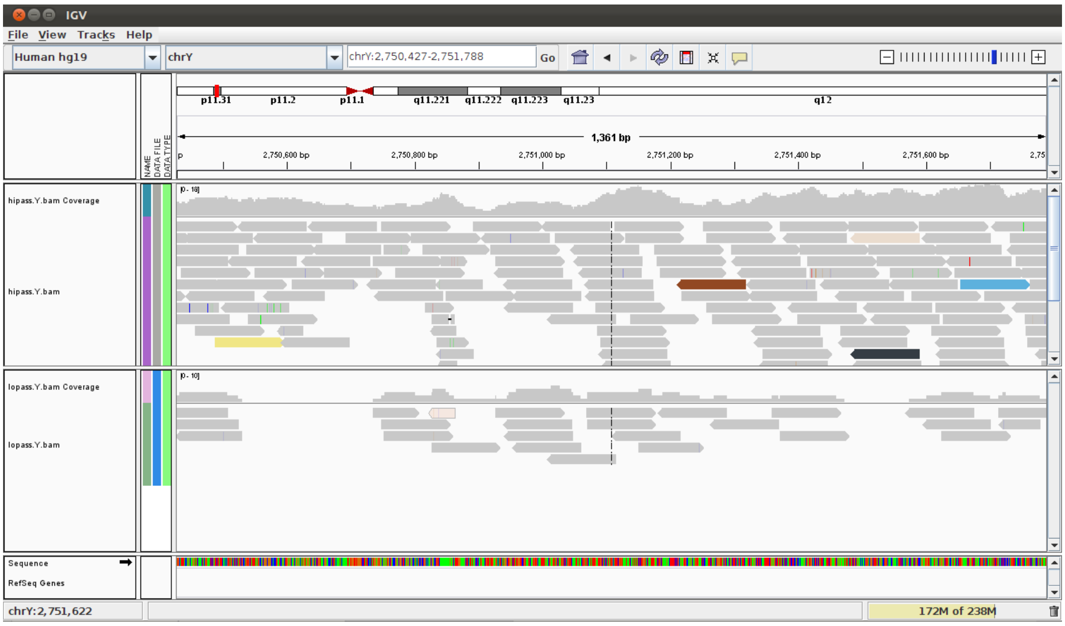


Figure 2.2: Difference in genome coverage between a 2x and a 15x experiment. We refer to the same sample, in the same genomic loci of Y chromosome. In the upper band Y chromosome coordinates are reported, in the middle band (marked with hipass) we see a continuous plot which represents read depth for each genomic locus in the interval, and below, small stripes that represent reads aligned to the reference. The bottom band is referred to a 2x low pass sequencing

quality parameter is strictly dependent on the tool chain that is applied.

### Phred scale

Quality scores in sequencing, both for base quality and mapping quality are usually expressed using the Phred scale: a logarithmic quality score giving the probability that a measure is an error. 2.4

$$Q_{Phred} = -10 \cdot log_{10}(P_{err}).$$

So while comparing error rate with Phred score we obtain:

| Phred Score | Error rate | Err% |
|---|---:|---:|
| 10 | 0.1 | 10% |
| 20 | 0.01 | 1% |
| 30 | 0.001 | 0.1% |
| $n \cdot 10$ | $10^{-n}$ | $10^{2-n}$ % |

Thus, a modern instrument with error rate E$< 10^{-3}$ we would expect a Phred base quality score Q$>$ 30.

## 2.5 Infrastructure

Doing sequencing needs a big infrastructure with a huge number of different professional figures. Alongside with biologists, geneticists and clinicians, computer scientists and engineers must work to ensure that the whole infrastructure works properly. An HiSeq 2000 can output over 6 terabytes of data in just 11 days. A computer center is the optimal solution to store and analyze such high data flow, which can not be viably processed on common desktop machines.

One crucial point, especially when dealing with an increasing number of samples is to keep all the workflow and data organized. Database with both web interfaces for biologists and command line interface for computer scientists and statistical geneticists are generally implemented. Within the collaboration we built a set of software tools, named LIGA (Lab Interface for Genomic Analysis), to keep track of the analysis processes, to show quality and result data to the Lab technicians and, on the command line side, to fastly select samples and studies. LIGA is illustrated in appendix A.

**command line**: a Linux system terminal which enables, by text typing, to launch software commands, doing searches and complex queries

# Chapter 3

# Sardinian cohort, demographic context and outgroups

In this chapter the Sardinian dataset is examined. Some detail on the sequencing experiment that provided Sardinian samples to our study is given. It will be examined Sardinian demographic and archaeological context, whose information will be useful in calculating mutation rate parameters. Non Sardinian samples used to contextualize Sardinian variability into a wider scenario are briefly described.

## 3.1 Sardinian cohort

Sardinian data encompasses 1,204 Sardinian samples from two major studies carried out to investigate different aspects of Sardinian population. Out of the 1,204 samples, 873 are unrelated individuals and 331 paternally related pedigrees: 154 father-son pairs, 5 families of 3 individuals, 2 families of 4 individuals.

| 1,204 | total |
|---|---|
| 873 | unrelated |
| 331 | related |
| $154 \cdot 2$ | father son |
| $5 \cdot 3$ | 3 ind.families |
| $2 \cdot 4$ | 4 ind.families |

### 3.1.1 SardiNIA

SardiNIA is a joint project between IRGB-CNR, the University of Michigan and the NIA Institute of the NIH [35].

It is a study on aging, aiming at understanding the exceptionally high average age that is reached in several Sardinian villages from the Lanusei Valley in the Central-Eastern region named Ogliastra, which was historically one of the most isolated parts of Sardinia. This study analyzes around 300 quantitative traits on a cohort of 7,000 samples, who were genotyped. Around two thousands of them were also whole genome sequenced with a low pass approach (4x-6x coverage). Data was collected in families, thus more than 1000 pedigrees are available, the largest family is 5 generation deep and encompasses 625 individuals.

**SardiNIA**: web sardinia.nia.nih.gov

**IRGB-CNR**: Institute for Genetics and Biomedical Research - National Research Council, Cagliari, Italy www.irgb.cnr.it

**UMICH**: University of Michigan (US) - Dept. of Statistical genetics sph.umich.edu

**NIA/NIH**: National Institute of Aging - National Institutes of Health (US) nia.nih.gov

Map of Sardinia. In red, marked with OG the Ogliastra region, origin of SardiNIA samples. Marked with black circles Cagliari (CA) and Sassari (SS), the two main cities of Sardinia where the autoimmunity study samples originated.

### 3.1.2 Type 1 Diabetes / Multiple Sclerosis Autoimmunity study

Another major study carried out in Sardinia, as a cooperation between IRGB-CNR, University of Michigan and CRS4, is a case/control study on autoimmunity. It focuses on two pathologies Type 1 Diabetes (T1D) and Multiple Sclerosis (MS) which have a high incidence in Sardinia. It has been observed that Northern European countries have a higher incidence of MS and T1D with respect to Southern European ones, with the exception of Sardinia which has the highest [36] incidence and shows as an outlier with respect to other Mediterranean populations.

**CRS4**: Center for Research, Development and Advanced Studies in Sardinia, Italy www.crs4.it

This study involves 8,000 individuals collected from the hospitals of Cagliari and Sassari, distributed among 2,000 T1D cases, 3,000 MS cases and 3,000 healthy controls. For enrollment it was requested that at least 3 over 4 grandparents were born in Sardinia, to maximize the investigation on Sardinian specific variants for autoimmune diseases. Around one thousand individuals from this project have been sequenced whole genome with a low-pass approach (4x-6x coverage).

For our study we ignored the case/control status but take into proper account pedigree information.

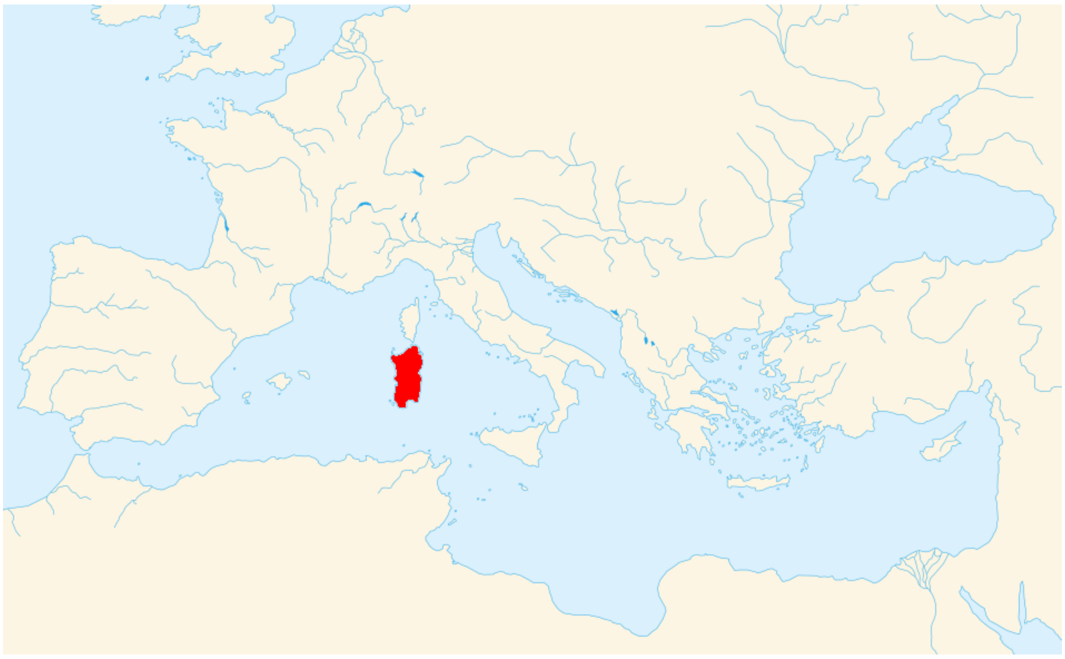


Figure 3.1: Sardinia (red) in the European and Mediterranean geographic context

## 3.2 Social and geographic context

Sardinia is the second largest island in the Mediterranean Sea. Its strategic position in the middle of the Basin has for long time being contended. While the coastal population may have been contaminated by the invasions and the continuous exchange of people during commerce and wars, a significant part of the population lived in the interior and in the mountains in a great number of remote villages and sparse settlements.

Terrain asperity and the scarce inclination of ancient and relatively modern Sardinians to move from their lands or to mix with population from other villages led to a great number of genetic isolates which are still clearly detectable.

All this led, on the Y chromosome, to a great diversity and to the establishment of a big amount of private Sardinian variability (genotypes which cannot be observed elsewhere).

This is the context that led to establish the aforementioned two major sequencing studies, SardiNIA and Autoimmunity: to investigate some special genetic features that are present only in Sardinia.

## 3.3 Archaeological context

To better contextualize the genetic tree a literature search was carried out, to check for dates of archaeological findings that testify population expansion in Sardinia.

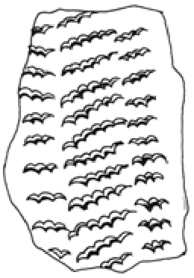
Sardinia was first populated by modern *homo sapiens* in the upper Paleolithic. The most ancient evidence is a ~20,000 years old,radiocarbon dated, human phalanx discovered at the cave named *Grotta Corbeddu* [37]. According to available evidence, these groups likely remained isolated and didn't contribute to the modern population of Sardinia.

Sardinia started becoming regularly inhabited during the Mesolithic period (dating 10,500-8,000 yr), as shown by six known settlements [38], see figure 3.2.1. First inhabitants were hunters and gatherers, therefore limited food resources, mainly constituted by the *prolagus sardus* and mollusks, didn't allow for population increase and the main settlements remained located on the coasts [39].

**prolagus sardus**: a pika-like rabbit, now extinct, but probably present in Sardinian fauna until 2/300 yr BP

Early Neolithic saw the switch to farming and animal breeding ( 7,700 yr BP, see figure 3.3 ), thus resulting in a population expansion [40][41] at that point. Archaeological evidence support this theory with 73 known sites. Among these, 45 are open-air and 28 are found in caves or inside rock shelters [41] [39]. Site distribution is shown in figure 3.2.2, it is also possible to note that the site distribution has changed and they are located both on the coasts and inland.

These times were characterized by appearance of *cardium pottery* and the beginning of the obsidian trade which led to contacts with the Northwestern Mediterranean coastal regions [43].



**cardium pottery**: pottery whose decoration is made imprinting the clay with the shell of a marine mollusk named *cardium edulis*
(G.Sobin - Un.Calif.)

This population evolved into the *culture of Bonu Ighinu*, named after a small location in Northwestern Sardinia near Sassari. This culture dates into the Middle Neolithic and it is attested by 76 archaeological sites whose dis-

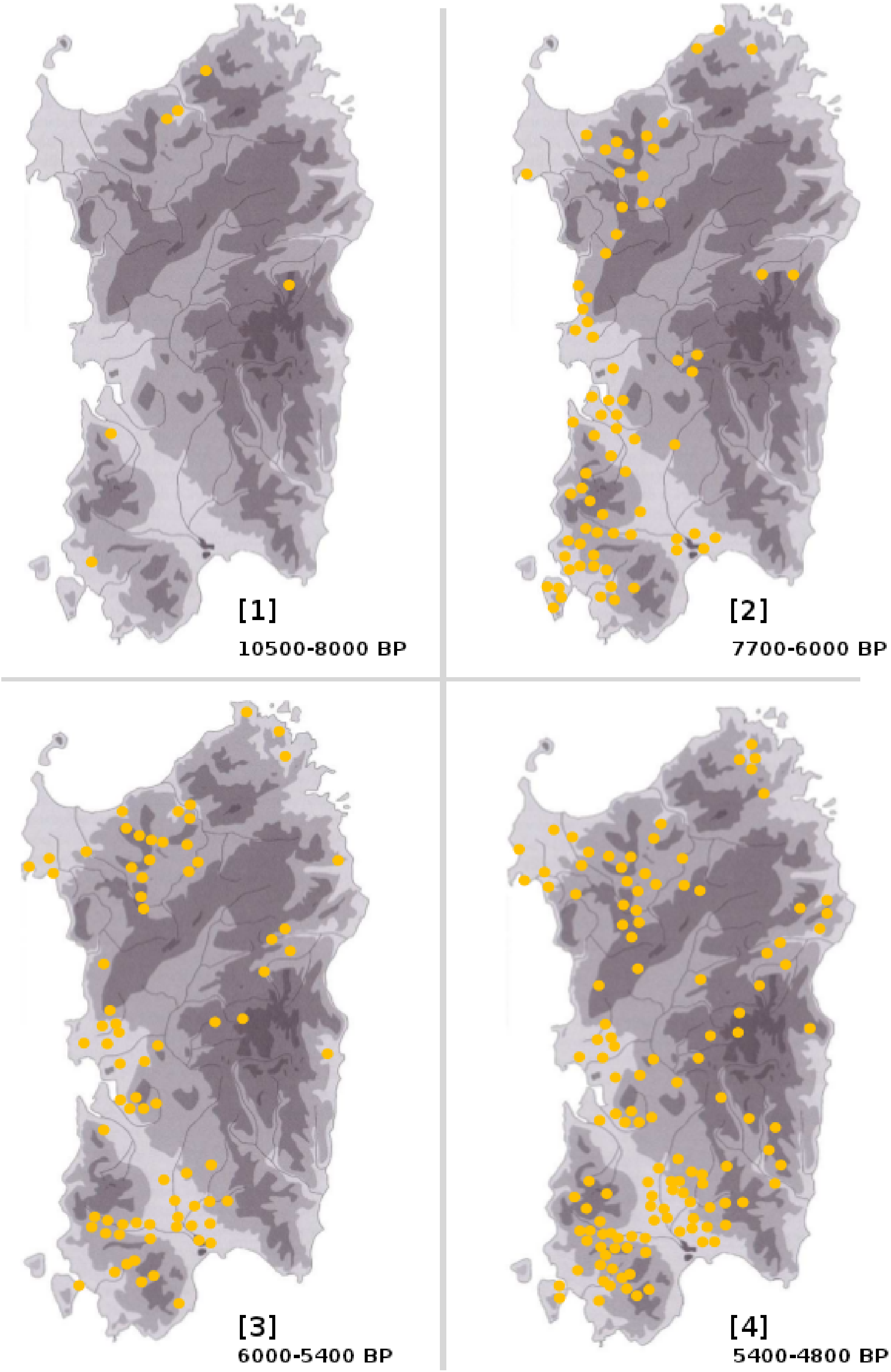


Figure 3.2: Neolithic settlements distribution in Sardinia [46]

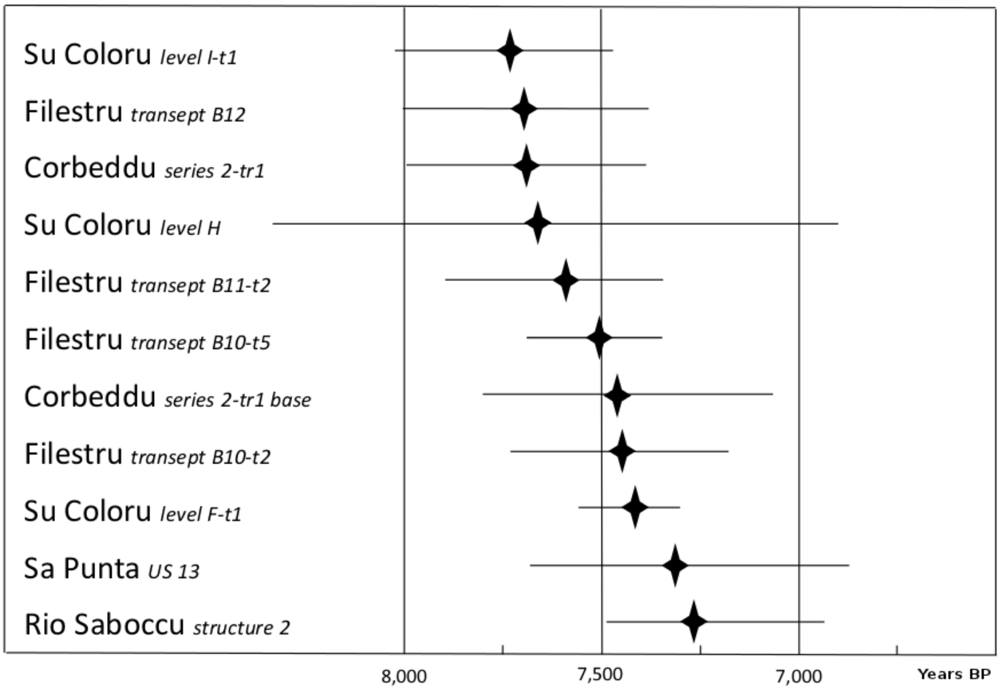


Figure 3.3: Early-Neolithic settlements radiocarbon dating estimates [42]

tribution is shown in figure 3.2.3. It is characterized by a greater abundance of open-air sites with respect to the past.

Another increase in population is recorded during the Late Neolithic (around 5,200 yr BP), with the *culture of Ozieri* (named after a Northern Sardinian town). In figure 3.2.4 are shown the 127 discovered sites related to that age: new territories were settled and internal movements are recorded [44]. Settlements took the form of new villages, and burial sites where characterized by the so called *domus de janas*, which are hypogeal tombs.

Eneolithic and Bronze age (4,800 - 2,900 yr BP) had as their highest point the *nuragic* culture, whose name derives from the big, conic, stone buildings whose name is *nuraghe*, that still mark Sardinian landscapes.

From the archaeological context it can be derived that the first and most massive expansion of Sardinian population can be dated to 7,700 yr BP.

This dating estimate is also supported by mitochondrial DNA estimates [10][45].

# 3.4 Outgroups

Building a tree with only Sardinian samples would make impossible to distinguish Sardinian private variability from variability that is common to the main haplogroups for the Y chromosome at continental (or global) level. Thus outgroups are fundamental not only to obtain results of interest for other populations, but also necessary to discriminate and understand important branchings that are relevant to Sardinian demographic history. Straightforwardly when trying to draw conclusions on *homo sapiens* an outgroup "outer" than homo sapiens is needed. We chose the Chimpanzee, we'll see why, in the following paragraphs. First we'll examine /textithomo sapiens outgroup samples.

## 3.4.1 1000 Genomes project

The 1000 Genomes project has been the first broad sequencing project, aiming at reaching at least one thousand samples sequenced on the whole genome with low coverage. The consortium sequenced (2013) more than 2700 samples, from throughout the globe. As long as the data are analyzed they are made publicly available.

**1000 Genomes Consortium**: 1000genomes.org

We used 133 male samples from their group named *phase 1* of European ancestry:

- 40 British
- 35 Finnish
- 8 Iberians
- 50 Italians from Tuscany

we ignored the CEU samples because they were from European ancestry but from groups living in Northern America which may not be consistent at the Y chromosome level.

These samples were used to clarify the position of variants within the main haplogroups hierarchy, discriminating European variants from private Sardinian mutations.

### 3.4.2 Additionally sequenced outgroup samples

To better improve the distinction between Sardinian and European samples we have chosen four non-Sardinian samples carrying known SNP markers, to

The samples belong from the following ethnicities:

- Corsican
- Basque
- Tuscan
- North-italian

The contribution of these samples has been fundamental to enhance the resolution of our study.

### 3.4.3 Ötzi

*Ötzi*, or the *Mummy of Similaun*, is a 5,300 years old mummy that was discovered in 1991 in the Tyrolean Alps. After his, likely violent, death the glacier ice mummified his corpse which arrived in an excellent status nowadays. A research published in 2012 [17] carried out a whole genome sequencing of *Ötsi* and released public data that we used in our work. We decided to include both to ascertain recent and older variability and to check whether, after TMRCA calculations the mummy's distance from branching point is coherent with its radio-carbon date.

### 3.4.4 Chimpanzee

Chimpanzee, as an ancestor of *homo sapiens*, carries the Y-chromosome from which, at a certain point in time, the human Y diverged. While it is obvious that several regions evolved differently over time, some other are still rich in analogies each other. We used homologous regions of *homo sapiens* and chimpanzee Y chromosomes to check the ancestral alleles of the basal haplogroup A. We used the panTro4 reference publicly available from UCSC[34].

## 3.5 Deep sequencing

Sardinian dataset encompasses 1,204 low-pass samples. The low-pass approach for the two aforementioned projects was chosen to allow sequencing a great number of samples while an high pass approach would have been limited by costs.

When descending to rare variants, or sample-level private variability, the downside of this approach comes.

As illustrated in paragraph , the low coverage leaves a lot of holes. They can be filled by imputation methods while haplotypes are shared with other samples. Rare variants while shared among only few samples are more likely to be missed, or can turn to singletons (seen in just one sample). And singletons, on a low pass approach cannot be trusted: due to the high error rate isolated variants have a 20% chance of being wrong (1X). If we retain them all the analysis could be impaired.

### 3.5.1 Why adding deep sequencing samples

Deep sequencing helps into resolving these problems. Providing a more uniform coverage the majority of chromosomal loci are likely to be covered. Real singletons with an adequate depth can be confirmed and rare variants have more chance of being spotted.

### 3.5.2 Improvements

We chose strategically a small set of 7 samples, 4 on the most diffused Sardinian haplogroup, one with an outer branch of the I haplogroup and 2 from external outgroups as A and J to improve the resolution among tree branches, to isolate Sardinian variability in a more consistent way and to evaluate the variability that was missing due to the low-pass approach.

# Chapter 4

# Analysis pipeline, from sequences to genotypes

In this chapter an overview of the analysis process is given. Detailed information on the steps that have been performed are reported, from DNA sequencing to genotype calls through strict quality assessment and filtering.

Software workflows are usually called pipelines when some input data flows through a variable number of software tools which manipulate, combine, filter them to obtain desired output. The name pipeline properly refers to the scripts or higher level software taking care of the execution of the underlying software tools.

As a margin note, the alignment workflow described here has been developed partly by and partly with original contribution of the author while following the sequencing data alignment, quality control and genotype calling for the two major sequencing experiments in Sardinia whose male samples are shared with our study (chapter 3).

Data analyses were carried on at the CRS4 High Performance Computing center, in Pula (Cagliari, Italy). This datacenter features a Hewlett Packard computing cluster with a processing capacity of 34.6 Tflops, that is $34.6 \cdot 10^{12}$ operations per second. The cluster is constituted of 400 nodes, each with 8 cores and 16 gigabytes of RAM (memory available for calculations), for a total of 3,200 computing units with 2 gigabytes of memory available each. A fast GPFS storage from DDN with a capacity of 4 Pb ( $4 \cdot 10^{15}$ four million gigabytes, equivalent to store more than one million complete human

**CRS4**: www.crs4.it

genomes), stores the sequences and serves them to the analysis tools.

## 4.1 Workflow overview

Alignment workflow aims at identifying the chromosomal coordinates of the sequencing reads and to call genotypes from them trying to lower the initially high error rate intrinsic to the sequencing method used. The main steps of the workflow (see figure 4.1) are described in the next paragraphs.

Analysis workflow is made of a custom pipeline developed by the author and colleagues within a collaboration on the analysis of Sardinian genomes. The pipeline took advantage of Hadoop MapReduce [19] scalable architecture to speed up calculations on the parallel environment that was available at the CRS4 computing center, using the Seal software package [21]A, which was developed within the collaboration and the Distributed Computing Group at CRS4. Seal implemented several tools of common use on sequencing analysis in a highly parallel environment.

**Hadoop**: hadoop.apache.org

**Seal**: biodoop-seal.sf.net

## 4.2 Sequence demultiplexing

Raw sequence reads must be extracted from a sample mix in which each sample's fragments were attached an unambiguous sequence (still unique with 2 sequencing errors), which is named barcode. This strategy allows to mix more samples on a single flow-cell lane (the minimum unit which can contain DNA). Once read, the barcode is stripped and the reads are assigned to the respective samples. This step made use of the `demux` package of the Seal software.

## 4.3 Sequencing data alignment

Once reads are properly assigned to their original samples, the alignment step can take place. To each read is assigned, where found, a chromosomal coordinate and a number of quality parameter calculated by the aligner, the software deputed to carry out this step. To perform this operation a reference

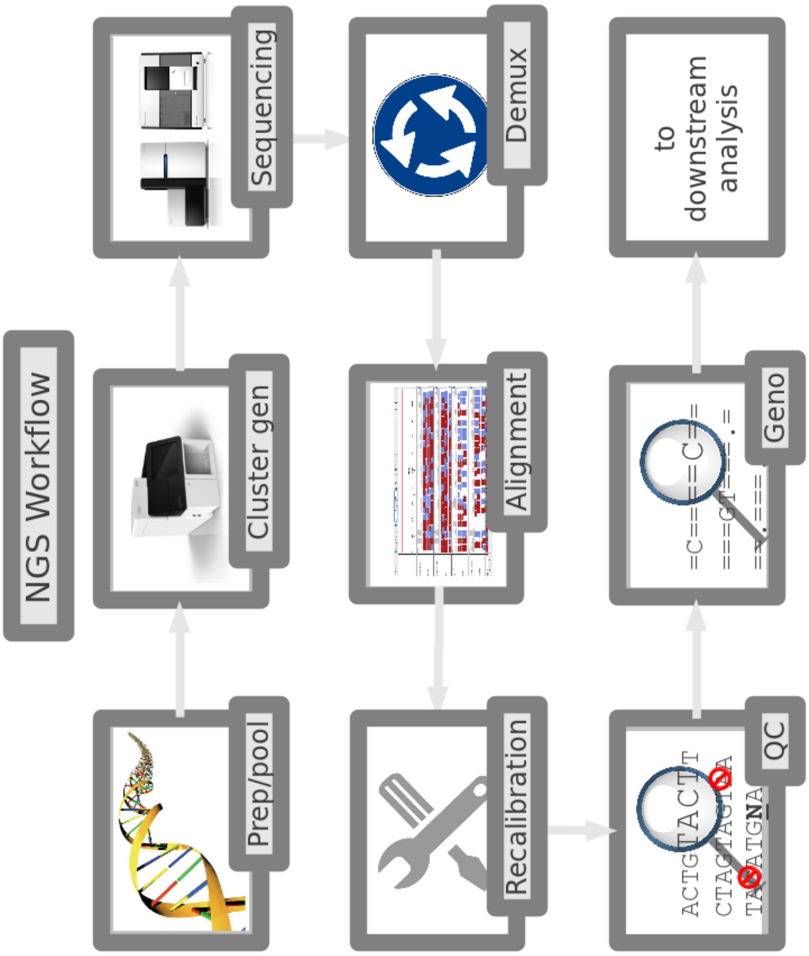


Figure 4.1: Next generation sequencing alignment workflow

genome is needed. We used the GRCh37/hg19 human genome reference [18] which is shared by NCBI and UCSC. Aligner takes advantage of the paired end sequencing: since the two reads must be aligned to coordinates within some hundred bases from each other. Generally read1 and read2 are aligned separately and matched at a second stage with the unique ID that Illumina machines assigns to the clusters at this purpose. Alignment (mapping) qualities are dependent on the probability that an alignment is wrong which is calculated by the alignment software, they are expressed in the Phred scale and they are 0 when no mapping or ambiguous mapping is detected. The rest of the scale is strictly dependent on the used alignment tool.

Each couple of paired end reads is treated as an individual entity for the alignment. Very small blocks of reads are sent automatically to node machines for the analysis. Thus we achieved a great scalability on the computing cluster we worked on.

This step was performed using Seal packages `PairReadsQseq`, `seqal` and `readsort`.

**BWA**: bio-bwa.sf.net

The `seqal` module integrates the BWA[20] tool which is the most widely used and reliable alignment software for whole genome resequencing.

Alignment is represented in figure 4.2.

## 4.4 Recalibration

Recalibration is a very important step into the alignment process. It takes the aligned reads and rescales the base quality scores according to the amount of computed sequencing errors, to resolve possible bias in score assignment. Polymorphisms databases plus reference genome are needed for this computation since they can exclude known SNPs from the calculation of errors.

**GATK**: broadinstitute.org/gatk

We used Gatk Recalibrator package [23] to perform this task. The calculation of the recalibration table has also been implemented into the Seal package.

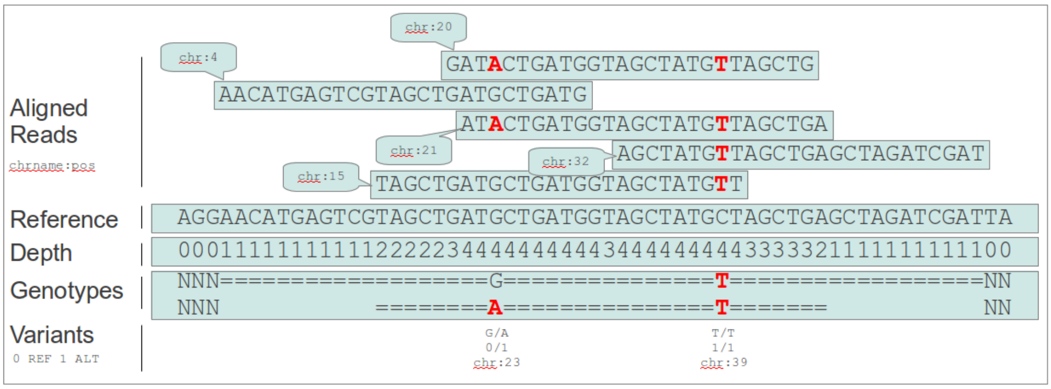

chr:20
chr:4
GATACTGATGGTAGCTATGTTAGCTG
AACATGAGTCGTAGCTGATGCTGATG
Aligned Reads
chrname:pos
ATACTGATGGTAGCTATGTTAGCTGA
chr:21
chr:32
AGCTATGTTAGCTGAGCTAGATCGAT
chr:15
TAGCTGATGCTGATGGTAGCTATGTT
Reference
AGGAACATGAGTCGTAGCTGATGCTGATGGTAGCTATGCTAGCTGAGCTAGATCGATTA
Depth
00011111111111222223444444444434444444443333321111111111100
Genotypes
NNN===================G===============T==================NN
NNN           ========A===============T=======           NN
Variants
0 REF 1 ALT
G/A
0/1
chr:23
T/T
1/1
chr:39

Figure 4.2: Representation of alignment. Reads are compared to a reference genome. When matching they are assigned a positional label. A locus depth is how many reads do actually support its genotype call. A naive genotype call is shown, in the real case complex quality considerations are performed before assigning a genotype. When read depth is zero, no genotype could be called. When just one, no diploid genotype could be confirmed.

## 4.5 Quality control

Quality control (QC) is a primary source of validation after the sequencing process.

There is a huge number of parameters that can indicate that something was wrong during sequencing or throughout the steps of the alignment pipeline.

Base and mapping quality are among the first ones to check. They express the probabilities that a single base is wrong or that the entire read has been mapped incorrectly to the reference genome.

Among the others, most important ones,

- incorrect distribution of the insert size which should be Gaussian, is directly related to the size selection on gel mentioned at page 19
- percentage of variants over the total amount of bases read,
- percentage of incoherence between the direction of alignment of read 1 and read 2 (should be $R(1/2) \rightharpoonup ... \leftharpoondown R(2/1)$ ),

**insert size**: the size of the original fragment sequenced, as inferred by relative alignment positions of read 1 and read 2

**QCTool**: berutti.net/public/qctool

To check and display all these parameters we used the QCTool software tool [22] (see appendix B) which enabled reporting both to the sequencing lab and for the analysis pipeline. Its output data were used for sample quality selection.

## 4.6 Identity control

All Sardinian samples belonging to our study, as they were enrolled for two genotyping and sequencing projects, have also been genotyped with Illumina and Affymetrix chips. While dealing with thousands of samples, especially in the laboratory part where manual work is necessary, it is relatively easy to swap accidentally or to misinterpret handwriting notes. This could be potentially destructive for studies. In our case we had to avoid to insert paternally related samples or duplicates without being acknowledged, since they would bias the results. More, no female sample should enter the cohort for a Y chromosome study and the check of concordance helps into this task. Sequencing and genotyping were carried out at two independent times, thus it is considerable as a reliable check.

**Chip**: Properly silicon based microarrays, colloquially DNA microarrays in general

**VerfyBamId**: genome.sph.umich.edu/wiki/VerifyBamID

We performed this step with the verifyBamId [24] package. This would check sequencing data against a genotype panel built from genotyping chips data. Firstly it checks the concordance of genotypes extracted from sequences with the data relative to the same sample into genotype panel. In case of discordance it will look for matchings between the sample's genotype and any of the genotyped samples. Thus it can find who is the sequenced sample and spot if there are duplicates whose name was not reported correctly.

## 4.7 Gender control

Even after a correct identity match between chip data and sequencing data not all resulting samples were males as expected. A simple wrong gender flag on the initial sample record could have propagated to the two databases creating an error. To test the gender of our samples we used the appropriate QCTool feature, which on a whole genome basis, compares the number of reads aligned on chromosomes X and on Y and will, normalizing the counts with autosome coverage. From them the gender is inferred. It is straightforward that unmatching samples were reported, corrected, and in case of female gender stripped from our study.

## 4.8 Sample selection and extraction

We developed a database system, the LIGA (see Appendix A, holding path, qualities and capable of storing, where available, several other metadata. This ensure the most precise and fast possible selection for samples. We used it to select samples for gender, minimum coverage, minimum quality (base and mapping), along with several other parameters.

The system extracts automatically the genotype out of the requested samples in a format named GLF (genotype likelihood format)[25]. This format encodes the marginal likelihoods for loci extracted from NGS data. For the ten possible allelic configurations into a diploid genomic locus, (AA AC AG AT CC CG CT GG GT TT) it is computed and stored the likelihood value (probability value that it is the true genotype).

**GLF**: genome.sph.umich.edu/wiki/GLF

## 4.9 Variant Calling step

The variant calling step is the first analysis involving all the sample cohort together.

At this step it is possible to apply filters both at a sample and at a panel level. We used here a modified version of GLF Multiples [26] (see for the custom version Appendix C) .

**glfMultiples**: genome.sph.umich.edu/wiki/GlfMultiples

The filtering criteria applied at this stage are reported in the following table:

| | |
|---|---|
| Positional filter | |
| Regions | 8.97 Mbp mask - MSY - X-degenerate |
| Average quality filters | |
| Base Quality avg. min (sample) | Q20 |
| Map Quality avg. min (sample) | Q50 |
| Genotype quality filters | |
| Base Quality min (genotype) | 20 |
| Map Quality min (genotype) | 60 (paired end), 37 (single end) |
| Genotype penalty filters | |
| Genotype penalty | $10^{-3}$ prob. genotype $\neq$ reference |
| Heterozygous genotype call | suppressed |
| Panel based filters | |
| Singleton minimum depth | 4 |
| Max het samples at locus | 1, if more ignore locus |

Where by singleton is defined a locus that shows an alternative allele in only one sample throughout a panel, by doubleton in only two. By max heterozygous samples it is meant the number of samples that, at a locus, show an heterozygous call: for the male specific part of Y chromosome, which is haploid, having many samples with heterozygous calls at a locus it is a signature of spurious alignments in place.

The meanings and the evolution of such filters will be explained in the following section.

After this step an error rate of 1% is present into the panel. On the next steps, described in the following chapters phylogenetic criteria are applied and we'll show how the error rate could decrease of an order of magnitude.

## 4.10 Choice of filtering criteria

The choice of the filters previously described has been performed using a great number of criteria. A recursive improvement has allowed to refine quality filtering parameters and region selection on the Y chromosome. In the following paragraphs we will illustrate briefly the process.

### 4.10.1 Positional filtration

The structure of the Y chromosome (see fig. 1.3) and the mechanisms of the alignment process needed after sequencing require that a selection on Y chromosome regions is made. We had to choose a subset of regions where the alignment and the resulting calls could be reliable and philogenetically consistent.

In the following table a list of the regions of the Y chromosome and of their inclusion status for our study is shown:

| Region | Selection | Excluded for | Explanation |
|---|---|---|---|
| X-degenerate | kept | - | no homology with X |
| X-transposed | discarded | alignment fail | homology with X |
| Pseudo autosomal | discarded | phylog. inconsistent | recombines |
| Ampliconic | discarded | alignment fail | palindromes and repeats |
| Heterochromatic | discarded | phylog. inconsistent | non-coding |

We indeed wanted to validate our choice and we examined the distribution of some quality parameters, after the panel was built, along the Y chromosome. The quantity we chose to examine are the following:

- Read depth
- Zero map quality rate
- Percentage of reads passing basic filters

Every quantity has been calculated with a bin size of 10kbp, see fig. 4.3 and its legend for details.

The samples were sequenced low-pass with an average whole-genome coverage of 4x, that means that every locus (whole genome) has been sequenced on average 4 times. On the haploid Y chromosome we expect a 2x coverage . In the first plot the read depth distribution is shown. The regions outside the X-degenerated mask show odd coverage, with peaks and drops (which are evidence of bad alignment).

The second plot reports distribution of map-zero quality genotypes, i.e. the percentage of genotypes into each chromosomal bin whose mapping quality is 0. They usually are badly aligned read like ambiguous alignments or

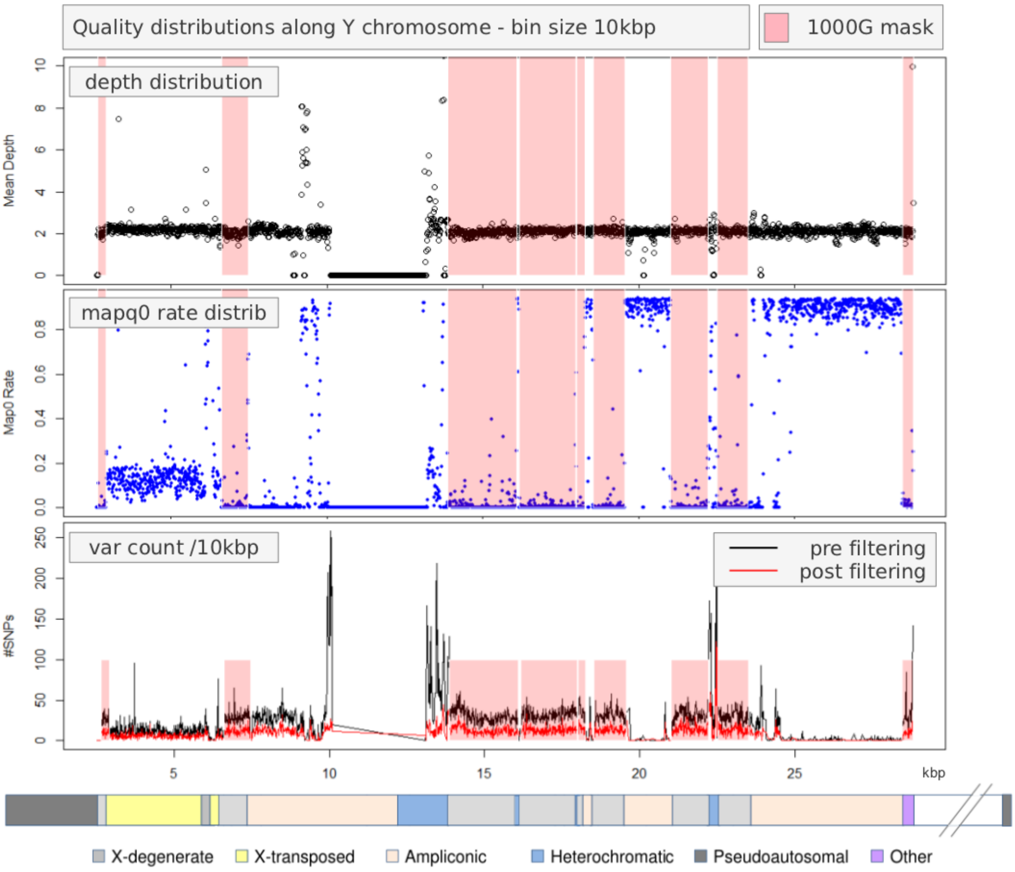


Figure 4.3: Distribution of depth, mapping quality, and number of variants pre and post filtering along Y chromosome. The bin size is 10 kbp. The first two plots are referred to unfiltered data. Pink color represents the 1000 genomes mask which was also applied in our work. **plot 1)** Read depth should be constant, peaks or drops indicate bad alignments. **plot 2)** Mapping quality zero rate. X-transposed regions having lots of homology with X have bad mapping quality, ampliconic for their structure have almost overall zero mapq (mapq0rate∼1), part of the X transposed has no reference, thus no mapped reads on it. **plot 3)** Raw and filtered called variants number (per 10kbp) are reported. Significant drops means bad reads (mapping with too many blank bases which act as a wildcard) or bad mappings leading to low mapQ (filtered) or to heterozygosity which for Y is spurious

the best possible alignments but with an unrealistically high mismatch rate. Again, it is quite evident that several regions outside the X-degenerated mask have high Map0 rates.

The last of the three distributions that was examined is the number of called variants per bin before and after filters. Some regions outside the mask have a post-filter higher drop in the number of variants called. This can be related to bad alignments leading to spurious heterozygosity, which is filtered, or to erratic alignments of reads with a too high number of blank base calls.

The quality controls illustrated above, plus the a-priori considerations that we made, enforced the idea that a selection of regions is mandatory to build a solid analysis. We thus chose the 8.97 Mbp mask based on retaining the chunks that make up the X-degenerated, male specific portion of the Y chromosome (MSY), and no outer regions.

**1000 Genomes Project**: 1000genomes.org

The same mask has been proposed by the 1000 Genomes Consortium [27] for their pilot study on human variation including Y chromosome [28].

### Quality filters

The basic filters we applied were aimed at assessing a lower limit to the mapping quality and the base quality that can be accepted on our data. In Phred score scale we required a mapping quality of:

| Map Quality min | read type |
|---|---|
| $MQ60$ | paired-end reads |
| $MQ37$ | single-end reads |

these numbers come from the top qualities assigned by our reference aligner BWA. [20] In the same scale we required a minimum base quality

$$Q20 \text{ (1\% error)}$$

to take into account the lower qualities of older data.

### Singleton filter

Illumina method's advantage lies mainly on its really high throughput. As a drawback it has a raw error rate which is very high on the single variant.

The average error rate for newer machines, is $Q30$, that means less than one wrong base call every thousand. Thus might let us think misleadingly that the error rate on variants is $10^-3$. Let's see it by the low-pass point of view. When we select variants, supposing to allow a minimum Q of 30, without any further recalibration we would select the 0.1% loci carrying an erratic base that would be identified as a variant. Considering that every human carries around 10 million SNPs, over 3 Gbp, they constitute 0.3% of the genome. Thus on a minimal low pass approach (1X) we select 0.1% bases that are errors and 0.3% that are real variants. One over four will certainly be wrong. With a mere 2X the noise can even be more, since we do not expect that random errors occur two times into different bases, in different reads that will map to the same locus.

From strict to relaxed filters it is generally expected for a single sample that the number of called variants will saturate to the total number of real variants plus errors. Since, inside a panel, the real variants will be partly shared they can be a small number, compared to a flood of errors, which will not, and will linearly increase with the number of samples we add to the panel.

Provided that duplicates are removed, a great part of sequencing and PCR errors mentioned above could be removed by rejecting singletons with read depth $< 2$, i.e. rejecting that loci that are mutated on a single sample within the panel. This selection will lead to lose a significant percentage of de-novo or private variability on a low-pass approach, thus it will be necessary to estimate the lost variability, which can be done by reprocessing selected samples with an high-pass approach.

### Heterozygous loci filter

In the MSY case it is important to remember that no heterozygous calls are possible. Therefore any call of that type is to be considered spurious, it might be caused by misalignment or by sequencing errors. If a locus for a certain sample is called heterozygous its genotype is discarded, thus avoiding sequencing errors. If more than one sample for the same locus is called heterozygous the entire locus is discarded, thus avoiding alignment errors which could be systematic, e.g. caused by homologies of the adjacent sequence.

### SNP filtering

When filtering single nucleotide polymorphisms we should take into account that the probability that a locus is mutated is of the order of magnitude of $10^{-3}$, that comes from the ratio between mutations and the total number of loci in the human genome ($\sim 0.003$). Thus every mutated locus into a sample is charged with a penalty corresponding to that probability. Only genotypes withstanding the penalty without changing are kept.

## 4.10.2 Panel Quality Control

Several parameters could help tuning the filters to obtain the best conditions to get clean and proper genotypes. These are observed a-posteriori, so a filter hypothesis must be done. Here we list some of them.

### Transitions to transversions ratio

DNA bases, according to their chemical form, are divided into purines and pyrimidines.

| | | |
|---|---|---|
| Purines | A | G |
| Pyrimidines | C | T |

such that A,G and C,T are respectively similar in structure among themselves. Mutations can be divided into two classes:

- Transitions are mutations in which a purine changes into the other purine or a pyrimidine changes into the other pyrimidine.
- Transversions are instead more radical changes in which a purine changes into a pyrimidine or vice-versa.

Figure 4.4 graphically illustrates what described above.

Since transitions imply smaller structural changes than transversions they are favoured. Almost two over three nucleotide changes are transitions, that means that average transitions / transversions ratio is $\sim 2$.

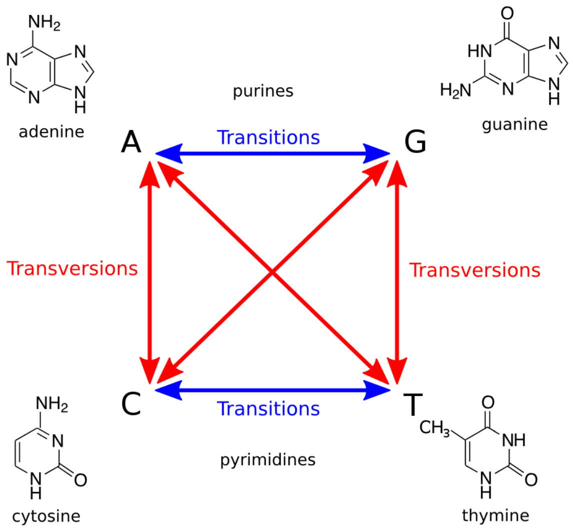


Figure 4.4: Chemical structure of purines and pyrimidines and the possible mutation classes

**dbSNP**: ncbi.nlm.nih.gov/SNP

For the Y chromosome, we measured that ratio among known mutations on MSY reported into dbSNP (version 136) [15] [16] this ratio is $ts/tv_{dbSNP}$ 1.32.

We thus set 1.32 as a minimum threshold for ts/tv ratio on our panel, for

1. Whole variants
2. dbSNP known variants
3. New variants
4. Singletons

This proved to be a major criteria contributing to filter tuning and to the optimization of panel quality.

### Loci with multiple alternative alleles

Since it is extremely unlikely that a single locus has encountered different mutations in different lines the presence of 2 or 3 alternative alleles at a locus throughout the panel can be a signature of bad quality reads that are still passing the filters, or of ambiguous calls that where not recognized by the aligner (which would have assigned a zero score). Since we aimed to keep our error rate on pre-phylogenetic analysis panel lower than 1% we tuned our filters to require that multi allelic positions are less than 1% of total loci, we achieved 0.6%.

### Chip concordance check

The two sequencing experiments that supplied Sardinian data to our projects have genotyped far more samples than sequenced, and all sequenced samples have been genotyped on Illumina or Affymetrix chip. This, on a first step helped in identity check and, a-posteriori, helped checking genotype quality. Concordance tests have been carried out and subsequent filter tuning was performed to ensure a discordance rate lower than 1%. The target was accomplished with a discordance rate of

$$E_{chip} = 0.9 \ \pm 0.5\%.$$

Note that this error rate is contributed not only by sequencing errors but also by genotyping errors.

### Pedigree concordance

Another concordance check can be done with Pedigree data. Considering that father and son will have identical genotypes, they can be compared. Over the 172 father son couples it has been measured discordance. With optimal filtering it was achieved an excellent error rate of:

$$E_{pedigrees} = 0.12 \pm 0.05\%$$

with a maximum discordance of 0.5% (and a minimum 0.02%). This before phylogenetic filtering.

### Deep sequencing check

In our panel we analyzed publicly available low coverage samples from the 1000 Genomes consortium. Eight of the Europeans were also sequenced at

**Complete Genomics:** completegenomics.com

high coverage by a company named Complete Genomics and made publicly available. While sequencing has a high error rate at low coverage, this effect is mitigated if not eliminated while raising coverage since random errors are usually overcome by several other reads reporting the correct genotype. To validate our filters we tested for concordance the genotypes called with our pipeline on low-pass data against the genotypes called by Complete-Genomics pipeline. After filter tuning we reached a surprisingly low discordance rate:

$$E_{CG} = 0.05 \pm 0.03\%$$

with a maximum peak of 0.1% discordance in one sample and two of them perfectly matching.

### Filter tuning and results

We tried several different configurations for the filters to incorporate the biggest number of informative variants as possible, without loading too much false positives. The final filtering is the one reported into previous paragraph 4.9.

# Chapter 5

# Phylogenetic analysis, mutation rate and TMRCA

In this chapter we will describe the phylogenetic analysis process, from the building of the tree skeleton, to the selection of informative variants and phylogenetic imputation. We will describe the kind of analysis that we performed to control the ancestral states of the alleles reported in the panel and why it is important to do so. As a final point, we will evidence the importance of the coalescent points into the tree and describe how to use them to calculate a mutation rate based on archaeological data. Another approach to calculate the mutation rate will also be discussed. In the end the calculation of TMRCA is illustrated.

## 5.1 Phylogenetic analysis process

A phylogenetic analysis on MSY aims at finding a hierarchy between the variants that are found in a group of individuals, isolating the common ones and descending into the various branching isolating time-by-time the variants that are distinguishing a branch from the other.

### 5.1.1 Premise: switching to ancestral reference

A genotype panel is the starting point for phylogenetic analysis. It is formed by the genotypes of all samples in the cohort at every loci of the Y chromosome that have been read and validated in at least one of the samples. The variant panel provided by the alignment and filtering process is positionally referred to the “hg19” [18] genome reference.

The human Y chromosome reference contained into hg19 is a mosaic of at least two individuals. The major portion belongs to the R haplogroup, while 1 Mbp from 14.3 to 15.3 Mbp (hg19) belong to the G haplogroup (according to the YCC nomenclature[12]).

When examining variants on haplogroups older than R it is essential to discriminate whether these are private of the single haplogroups or if the reference contains the more recently mutated allele with respect to its ancestor haplogroups.

Thus a core feature of the process is to switch to an ancestral reference: i.e. a reference built using the alleles carried by the older haplogroups

It is relatively easy to distinguish between private variability among recent haplogroups. When a genotype at a locus on an haplogroup and all newer ones is reported as reference while older haplogroups report the alternative allele this is a clear signature that reference itself reports for that locus the derived allele while the ancestral is the one detected as alternative.

Thus it is necessary to swap. Our reference will thus be made of ancestral alleles and our variants will not be called alternative, but derivate.

When variants are reported on the A haplogroup (which is the oldest one, of African origin) it is impossible, referring only to human samples, to discriminate whether reported variants are private for haplogroup A, or if they are ancestral alleles.

An external outgroup is the only way to solve this issue. This is why we used a primate with an high level of homology with *homo sapiens*, the chimpanzee.

Failing to switch to ancestral reference would lead to a puzzled hierarchy and inconsistent haplogroups and dating estimates.

This process is not straightforward: a preprocessing of the tree is needed and ancestral correction is performed through the various steps. Once the last corrections are performed the tree should be re-generated.

### 5.1.2 Selecting informative variants

The ISOGG (International Society for Genetic Genealogy) defines a list of the known markers  for haplotype attribution.

**marker**: in this case a derivate allele which marks the subdivision of Y chromosome branches

As a first step, the main haplogroups were identified with the known markers. This was used to built the skeleton of the tree and assign samples to their respective main haplogroups. Then the process is repeated with known sub-haplogroups and with variants that could be unequivocally associated to branches. For any locus a maximum tolerance of  1% genotypes not respecting hierarchy was accepted to classify derivate alleles as phylogenetically informative.

### 5.1.3 Hierarchical inference of the variation

Due to the low-pass sequencing approach every sample in the panel has a lot of missing base calls and, even after genotype filtering, the error rate is not negligible.

After the tree skeleton was built, allowing for tolerance, it became possible to identify incoherent or missing alleles which become trapped into a hierarchy.

Thus it is possible to infer their ancestral or derivate status or to correct the detected one.

The process is shown in the small scheme reported in figure **??**.

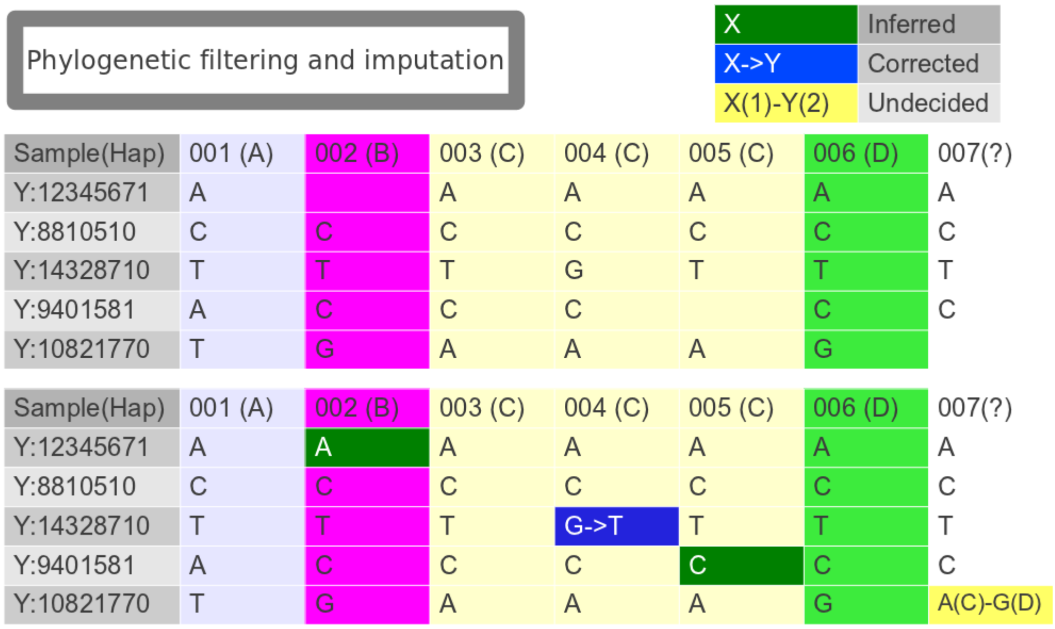
This approach, as noted into paragrah 3.5 is intrinsically less powerful for rare variants, that, in our case, are both the rare lineages and the terminal branchings of the lineages. Deep sequencing allows far more detail over these regions.

### 5.1.4 Building the tree

The phylogenetic panel is complete while every locus has an assigned allelic status for all the samples.

We built the tree using the Phylip [29]  software tool. We used the Pars package, implementing a discrete character parsimony algorithm[30].

**Phylip**: evolution.genetics. washington.edu/phylip.html



Phylogenetic filtering and imputation

| | |
|---|---|
| X | Inferred |
| X->Y | Corrected |
| X(1)-Y(2) | Undecided |

| Sample(Hap) | 001 (A) | 002 (B) | 003 (C) | 004 (C) | 005 (C) | 006 (D) | 007(?) |
|---|---|---|---|---|---|---|---|
| Y:12345671 | A | | A | A | A | A | A |
| Y:8810510 | C | C | C | C | C | C | C |
| Y:14328710 | T | T | T | G | T | T | T |
| Y:9401581 | A | C | C | C | | C | C |
| Y:10821770 | T | G | A | A | A | G | |

| Sample(Hap) | 001 (A) | 002 (B) | 003 (C) | 004 (C) | 005 (C) | 006 (D) | 007(?) |
|---|---|---|---|---|---|---|---|
| Y:12345671 | A | A | A | A | A | A | A |
| Y:8810510 | C | C | C | C | C | C | C |
| Y:14328710 | T | T | T | G->T | T | T | T |
| Y:9401581 | A | C | C | C | C | C | C |
| Y:10821770 | T | G | A | A | A | G | A(C)-G(D) |

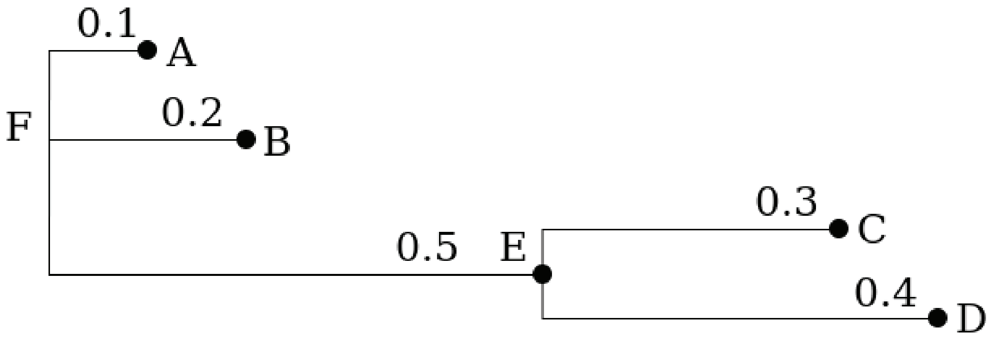
Figure 5.1: Phylogenetic filtering and imputation. A simplified group of samples is given (001-007). All of them except 007 have been identified as belonging to the main (fictitious) haplogroups between A and D. Mutations (Y:position) are ordered hierarchically (oldest to newest). In the second table sample B position Y:12345671 can be inferred as A, since older and newer haplogroups carry an A. Similarly for sample 005 locus Y:9401581. The ancient A haplogroup carries a A while the others a C, thus even this position can be inferred. Similarly, looking at sample 004, locus Y:14328710 we read a G while the same locus in other samples which share downstream genotypes (in the hierarchy) carry a T. Thus the genotype which likely comes from a sequencing error can be corrected highly improving the error rate.

Simplifying, this algorithm cycles through possible tree configurations and finds the simplest genetic path (in terms of mutations) that may produce, from a common root, the genotypes reported in our panel.

These criteria comply with the following:

1. There's a unique path connecting a sample to his MRCA. And this is consistent with Y chromosome inheritance mode.
2. The probability that the same locus has more than one alternative allele through the panel is negligible.

The resulting tree is produced in the Newick [31] format which consists of a hierarchy of nested parentheses representing the tree branching. As an example, the simple tree in the following figure:



has the following representation in the Newick format:

(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F

The resulting tree in Newick format has then been plotted with the FigTree 1.4.0 software tool.

**FigTree**: tree.bio.ed.ac.uk/software/figtree

## 5.2 Inferring ancestral status for A

For recent haplogroups it is always possible to get information about the ancestral status by comparing their alleles with both the reference and the other haplogroups. It cannot be done with the oldest haplogroup, the A, of African origin. To identify whether the supposed variability (compared to the reference) is ancestral status or variability internal to A we used the LASTZ [32] software tool as implemented in the Ensembl Compara pipeline [33] for comparative genomics, according to the approach of Wei et.al [1].

**Ensembl Compara**: the Ensembl comparative genomics pipeline ensembl.org/info/genome/ /compara

Basically we compared the two reference genomes *homo sapiens* (hg19)[18] and chimpanzee (panTro4)[34] and obtained a correspondence table between the two, annotating in the hg19 coordinates the observed differences. Simplifying the pipeline performed a sort of alignment of the chimpanzee reference against *homo sapiens* reference, then the genotypes were extracted, expressing chimpanzee genotypic status at homologous loci in humans.

Exploiting chimpanzee allelic status we were able to root the tree. We clarified almost all allelic uncertainties, only 26 positions were uncertain and were discarded.

## 5.3 Mutation rate

One of the aims of this work is to give estimates on the putative time in the past when relevant events occurred such as branching, migrations, expansions. By a phylogenetic point of view this means: relate time intervals to the number of mutations between branching points in the tree.

To set up such relationship we need to estimate the *mutation rate*, a quantity expressing the accumulated de-novo variability over time. Mutation rate can be defined by generation, or by year, in the following ways:

$$\text{per yr: } m_{rate} = \mu \cdot bp^{-1} \cdot yr^{-1}$$
$$\text{per gen: } m_{rate} = \mu \cdot bp^{-1} \cdot gen^{-1}$$

where $\mu$ stands per number of mutations and *gen* per generation count. Depending on the available data a molecular approach or a phylogenetic one can be used. Here, both are illustrated, focusing on the second one, which is the method we used in our work.

### 5.3.1 Molecular mutation rate

Molecular approach to mutation rate supposes a complete knowledge of the variability inside the tree. It can be evaluated with a de-novo study on pedigrees and must take into account that the probability that a de-novo rises on Y chromosome is dependent on father's age at conception, but this dependence is not linear. Provided that an average father's age at conception can be established mutation rate can be established as a multiplicative coefficient to that. Thus it is necessary or to model dependence or to evaluate the average reproductive age for males in the population under examination, and in the general population if we aim to take widely applicable conclusions.

The mutation rate will be expressed by the following form:

$$mol_{rate} = \frac{\overline{\mu}_{gen}}{\overline{\Delta T_{gen}} \cdot MSY_{length}} \tag{5.1}$$

Where $\overline{\mu_{gen}}$ stands per average mutation number per generation, $\overline{\Delta T_{gen}}$ for the average generational step (age of father at conception), and $MSY_{length}$ is the length in base pairs of the portion of the Y-chromosome that we are want to consider (MSY).

With a low pass sequencing approach it is not feasible to get complete discovery of the variability: chances that some common variant could be loss are small, but for the terminal branches, carrying increasingly private variability (among unrelated individuals) we lost almost all the information.

We thus needed to switch to a phylogenetic approach, that we describe.

### 5.3.2 Phylogenetic rate and calibration

The shape of the tree can point out some events that have happened in the past.

When a branching point from which a lot of ramification departs is found, it indicates a population expansion. Few individuals carrying few (or one) haplogroups generate lots of separate family lines in which variation can accumulate independently.

A branching point generally separate common variability (upstream) from more private variability (downstream).

In Sardinian case final branchings will be deeply ramified, indicating the expansion of the ancestors of the present Sardinian population.

Whenever it is possible to date somehow precisely the expansion points it is also possible to give estimates on the per/year variability without a prior knowledge of the generational step. The branching point used for mutation rate calculation is named *calibration point*.

The mutational rate per year, in this case will be more properly named phylogenetic rate, since it is calculated on the tree clades. The following formula is used:

$$phyl_{rate} = \frac{\overline{\mu}_{now,calibrationpoint}}{\overline{\Delta T}_{now,calibrationpoint} \cdot MSY_{length}} \tag{5.2}$$

where, $\overline{\mu}_{now,calibrationpoint}$ stands for average number of mutations in the branches encompassed between the terminal lines and the calibration point, $\overline{\Delta T}_{now,calibrationpoint}$ for the average time separating terminal lines living time (now, in our case) from the the date in the past of the expansion event indicated by calibration point.

The low pass approach will still generate some problems, but could be solved, as we did, by deep sequencing, which helped us to refine the real length of the branches, especially terminal lines.

Generally this approach introduces relevant errors, due to relatively large variance on archaeological sites age estimation plus the measurement errors, and to the association of this average age to the calibration point whose mutational distance from the present is, again, mediated among the downstream branches. Thus, a molecular approach for the mutation rate is always preferable when possible.

## 5.4 TMRCA Calculation

As an inverse process, with respect to the calculation of the mutation rate, TMRCA applies the mutation or phylogenetic rate to the whole tree, giving estimates of past events. While the origin of the main haplogroups is known, it can be used to date in the past some fundamental demographic events. It is also possible to compute a putative date for the living time of the most recent common ancestor to the individuals in the dataset. The wider (in terms of haplogroups) the dataset, the most general the conclusions are. Theoretically other expansion events whose date is unknown or branchings between well known haplogroups can be tentatively dated, rigorously these estimates could be precise only when the dataset is encompassing all contemporary living population.

The TMRCA for the branch $i$ will be calculated thus as:

$$TMRCA_i = \frac{avg(\sum_j^{subbranches_i} \mu_j)}{phyl_{rate} \cdot MSY_{length}} \tag{5.3}$$

Where $\mu_j$ is the number of mutations in each segment of the sub branches of $i$ from the terminal branches to the coalescent point of the branch $i$, averaged for the evolutionary lines pointing at the point.

# Chapter 6

# Results

In this chapter the results of our phylogenetic analysis will be discussed in detail. The tree built out of the 1,204 samples will be reported and the results of tree calibration and phylogenetic rate calculation will be explained. It will be shown how adding high coverage samples improves the knowledge about rare and private variability and consequently refines the calculation of the phylogenetic rate. An overview of the main haplogroups will be given and a TMRCA results, while not rigorously applicable, will be associated with relevant population events. A brief statement of perspective work will be given.

## 6.1 Some numbers

The analysis produced a catalogue of 11,763 phylogenetically informative markers defining Y-chromosome haplogroups and sub-haplogroups. Among them 5,012 were known and catalogued in the ISOGG [13] database or in dbSNP [16] as of 2013, while the other 6,751 are new. Among them 4,872 build up the root of the tree.

| Total markers | 11,763 |
|---|---|
| New | 6,751 |
| Reported* | 5,012 |

* on ISOGG or dbSNP

The newly found mutations are now being included into the ISOGG database.

## 6.2 Phylogenetic tree

We built the phylogenetic tree out of the 1204 low-coverage Sardinian samples. We used the out-grouping samples from 1000 genomes consortium to define better Sardinian private variability but we chose not to insert them into the tree for homogeneity.

The resulting tree is shown in Figure 6.1 [46].

Tree clearly shows how Sardinian samples encompass most of European variability. As explained in paragraph 5.3.2 most main haplogroups show signs of expansion with dense branching in the terminal lines.

The tree is uniform. The average length of branches (from the terminal part to the MRCA) has a very small standard deviation:

$$1002.6 \pm 21.2 \text{ SNP}$$

every sample has on average the same number of variants in its lineage with respect to ancestral reference.

This supports the hypothesis of an independent and neutral evolution for the different branches. Neutral evolution of the branches is required for the *molecular clock* hypothesis to be valid (see page 8).

A better description of the haplogroups will be given in paragraph 6.7.

Haplogroup frequencies that were found are consistent with previous estimates on Sardinian and European population as shown in figure 6.2.

## 6.3 Phylogenetic rate calculation on the tree

To evaluate the TMRCA for the main branches and the entire tree we defined a calibration point and calculated the phylogenetic rate on it using the archaeological information reported in paragraph 3.3.

### 6.3.1 Calibration with haplogroup I

We chose as a calibration point the root of the I2a1a-$\delta$ branch of I, which encompasses $\sim$ 35.7% of our dataset (I $\sim$ 40.7%). An expanded view of I2a1a-$\delta$ is shown in figure 6.3.

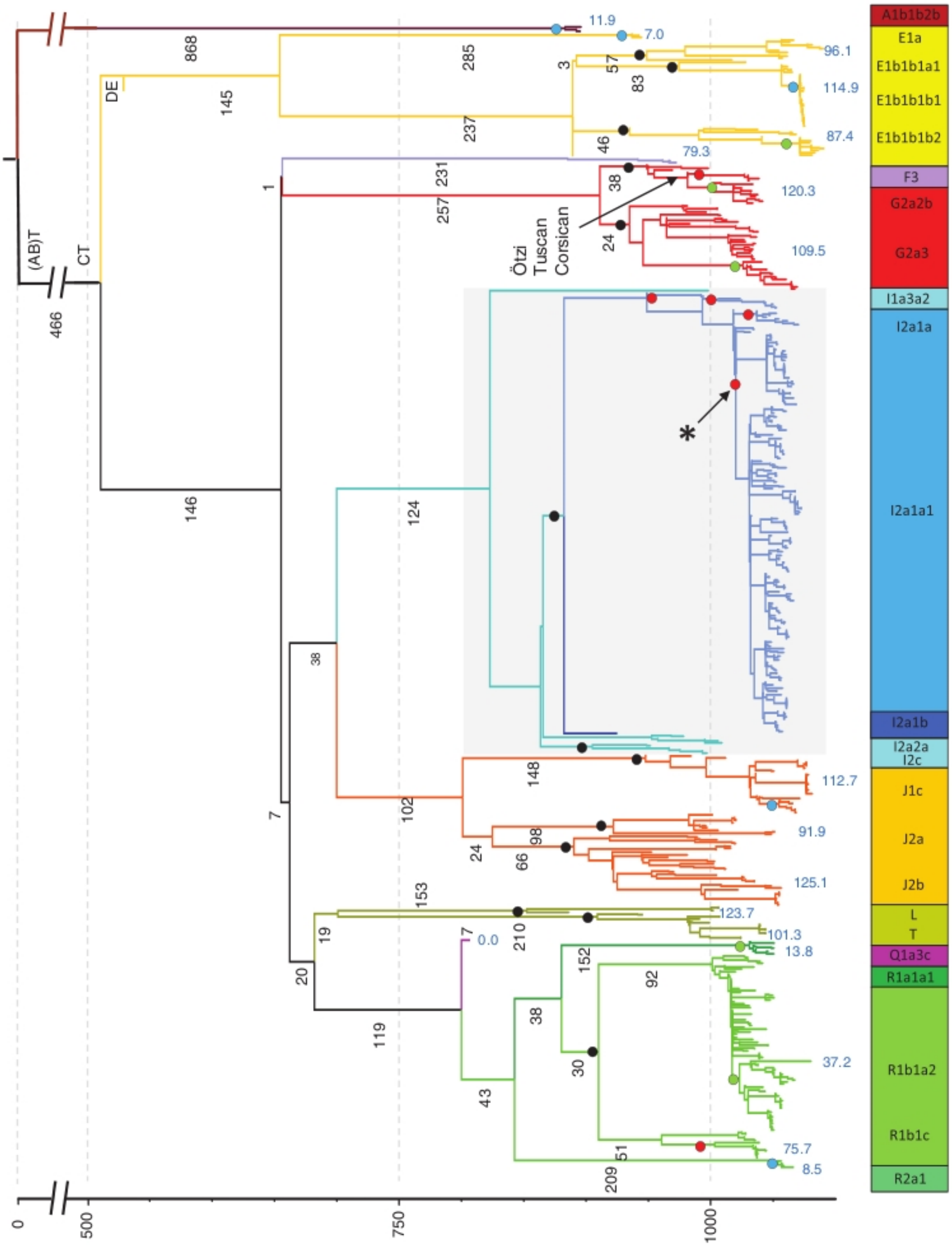


Figure 6.1: Phylogenetic tree [46]

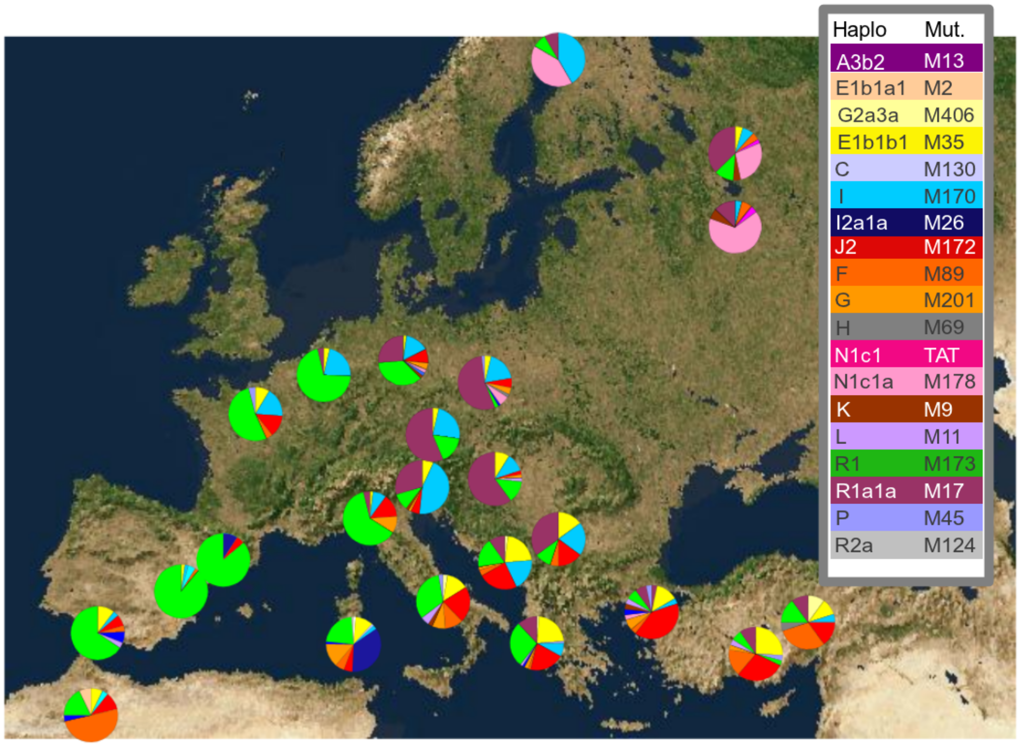


Figure 6.2: Haplogroup distribution in European areas. Data combined from different sources [2][47].

The variation downstream the branching point is private Sardinian variability, thus indicating a population expansion. This has been confirmed with the analysis of four non Sardinian samples with M26 mutation, which define I2a1 haplogroup. In Figure 6.1 the 4 samples are marked with B: Basque, I: North Italian, T: Tuscan, C: Corsican. It is evident how their branches are separated from the main Sardinian specific variability.

The calibration point, in our case, could be dated with archaeological data. As shown in figure 3.2.(1.2.3) and in plot 3.3 there was a huge increase of population sites between pre-Neolithic and early-Neolithic. The average age for the new sites date back to 7,700 years ago, which we chose as calibration date (par. 3.3).

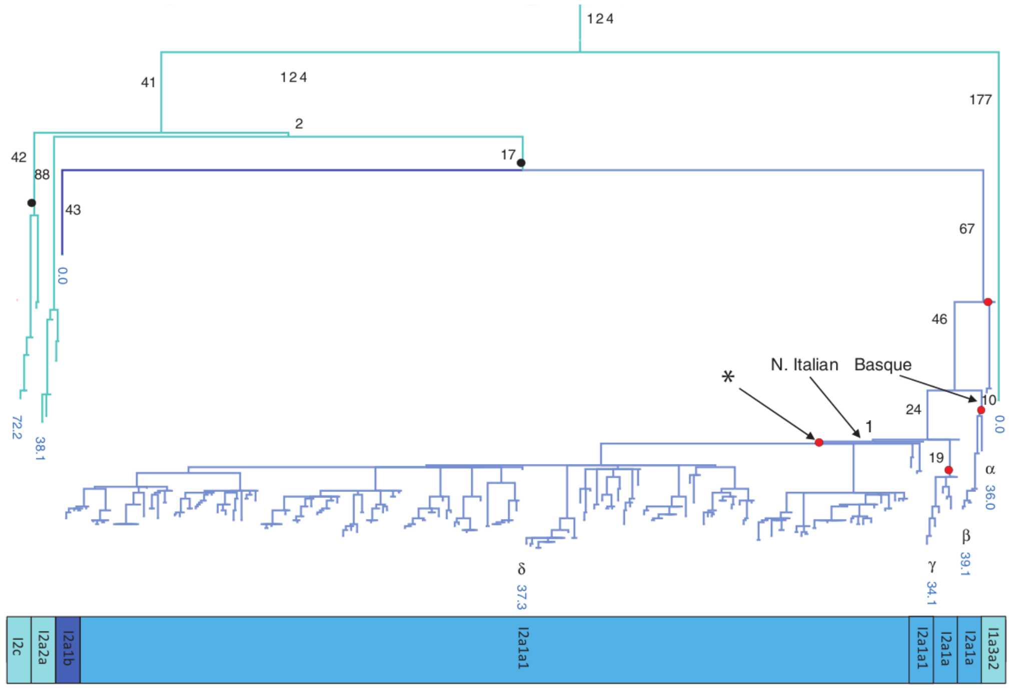


Figure 6.3: Phylogenetic tree [46]

### 6.3.2 Provisional rate

Once the calibration point was chosen and an estimate date was appointed on it, phylogenetic rate was calculated.

The most diffused Sardinian haplogroup is ideal for its very high resolution throughout common and most rare variants. Indeed, due to the low pass approach there are still:

- sample private variants (singletons) which cannot be validated,
- variants with frequencies lower than 0.5% whose discovery is limited both by the sample size and the coverage unevenness.

We calculated the average mutation length of the I2a1a-$\delta$ Sardinian private branch. The branch is shared by 430 individuals, it has an average variability of

$$\overline{\mu} = (37.3 \pm 7.8)\ \text{SNP}$$

Took as calibration point:

$$\Delta T = (7,700 \pm 300)yr$$

On the MSY length of:

$$\text{MSY}_{length} = 8.97 \cdot 10^6 \text{ bp}$$

Thus followed, using the formula 5.2:

$$phyl_{rate} = (0.59 \pm 0.11) \cdot 10^{-9} bp^{-1} yr^{-1} \tag{6.1}$$

This rate is expected to be slightly lower than the real one, since terminal lines, are actually longer than estimated. In fact low-pass approach leads to missing rare and private variability.

Thus the error rate while correctly calculated on available data it does not encompass bias contributions.

Deep sequencing helped to improve information on missing variability. In paragraph (6.4.2) it will be calculated the corrected rate.

## 6.4 High coverage samples

The tree built out of the 1,204 samples evidenced Sardinian private variability and unequivocally associated new variants to haplogroups. Still, some questions were unsatisfactorily answered:

- Amount of missing private variability on the tree, which is the real branch length? Only singletons with high coverage could be considered validated
- I2a1a-*delta*, Sardinian specific haplogroup boundaries could be better defined
- Define better I haplogroup boundaries
- Define better separation between A/ER and refine ancestral information

Seven samples were strategically selected and sequenced at high coverage:

- 4 within the Sardinian I2a1a-$\delta$ haplogroup

- 1 on the I2a1a-$\beta$ to maximize discovery of markers differencing $\delta$ and $\beta$

- 1 from J haplogroup as close outgroup to separate non-I variability

- 1 from A haplogroup (the African most ancient one) to assess the most of ancestral variation by means of comparisons with the chimpanzee.

### 6.4.1 Results and improvements

For these seven samples, with the low pass approach 6,268 polymorphic sites were identified: 2,818 of them were directly sequenced, the other 3,450 were imputed with phylogenetic criteria. All these were detected with the high pass, only four among them resulted wrong.

More, high-pass approach not only validated inferred genotypes but allowed the discovery of 304 new SNPs, and included into the tree 58 new markers which were discarded as singletons in the low-pass approach, so a total of 362 new markers entered into the phylogeny.

The new samples had an average coverage of (17.3 ± 4.4)X, ranging 14-27X. The average error rate for low pass calls is (0.1 ± 0.2)%, the left range going below zero because four out of seven samples had no genotyping errors.

A reduced tree made out of the 7 samples was built and is shown in figure 6.4. As expected, it is evident that the major relative gain in number of variants lies into the terminal branches when the 4 I2a1a-$\delta$ samples had an average increase of 52% more discovered variants. Increase is far less on shared ramifications. An apparent major increase for the samples related to the other haplogroups is due to the lower number of samples present in the original panel and, thus, to a larger amount of apparently private variability.

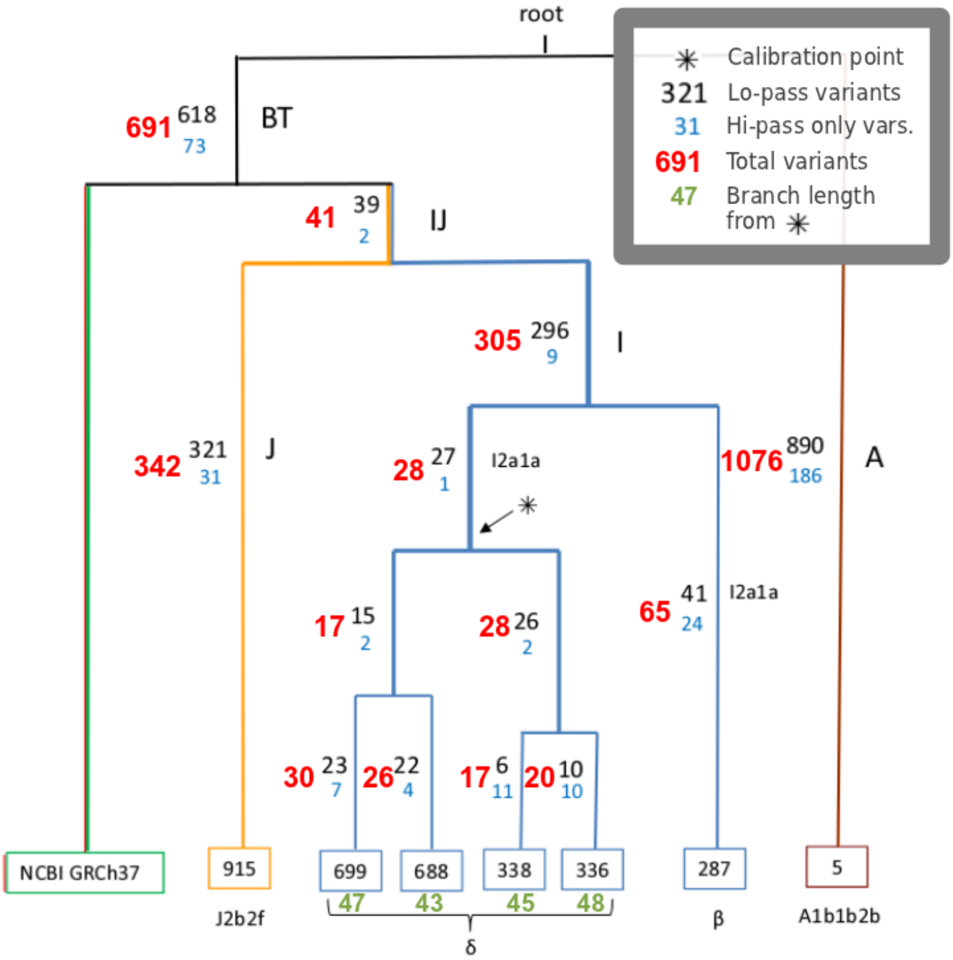


Figure 6.4: Phylogenetic tree for the seven high pass samples, variant count increase and terminal branches length. On the boxes the samples ID (1-1204 in hierarchical order) and on the left how hg19 human reference is positioned when referred to ancestral alleles. [46]

### 6.4.2 Phylogenetic rate

The new average mutation count in the terminal branches downstream the calibration point is:

$$\overline{\mu} = (45.7 \pm 2.2) \text{ SNPs}$$

It follows, using the parameters in 6.3.2 that the revised mutation rate is:

$$phyl_{rate} = (0.66 \pm 0.04) \cdot 10^{-9} bp^{-1} yr^{-1} \tag{6.2}$$

An increased number of individuals sequenced at high coverage would definitely improve the estimate, and, in case of pedigrees, would allow to calculate a molecular mutation rate, as discussed in paragraph 5.3.1. This would lead to a significant decrease in error rate which is now around 6%.

On Y chromosome this rate means one mutation every 169 years.

## 6.5 TMRCA

TMRCA was both calculated for the total low-pass approach and with the deep sequencing correction. Such approaches gave significantly different results in term of time for the whole dataset's MRCA.

### 6.5.1 Without deep-sequencing

Without deep sequencing samples, the phylogenetic rate calculated, in the result numbered 6.1 is $(0.59 \pm 0.11) \cdot 10^{-9}\text{bp}^{-1}\text{yr}^{-1}$, the length of the MSY portion 8.97 Mbp, the average length SNP amount per sample of $(1002 \pm 21)$ SNPs. We used formula 5.3 to calculate the TMRCA of the whole dataset.

$$TMRCA_{low\_pass} = \frac{(1002 \pm 21)}{(0.59 \pm 0.11) \cdot 10^{-9} bp^{-1} yr^{-1} \cdot 8.97 \cdot 10^{6} bp} = \\ = (189 \pm 35) \cdot 10^{3} yr \tag{6.3}$$

The standard deviation over the length of terminal branches for haplogroup I, due to the low-pass approach is relatively high and affects the final error on the measurement ( $\frac{\Delta TMRCA}{TMRCA} = 0.18$)

### 6.5.2 TMRCA refined

We calculated, then, the TMRCA using the refined phylogenetic rate. With a mutation rate of $(0.66 \pm 0.04) \cdot 10^{-9}\text{bp}^{-1}\text{yr}^{-1}$, we obtained:

$$TMRCA = \frac{(1002 \pm 21)}{(0.66 \pm 0.04) \cdot 10^{-9} bp^{-1} yr^{-1} \cdot 8.97 \cdot 10^{6} bp} = \\ = (169 \pm 11) \cdot 10^{3} yr \quad (6.4)$$

Thus, should any variability still be missing, this will lower more the coalescence estimate.

### 6.5.3 Tentative historical interpretation

We empirically correlated dating estimates on branching and expansion events evidenced by the tree shape, with real events in the past and found interesting concordance.

*As stated in chapter 1.6 the estimates reported below are only tentative and the associations, while concordant, are reported for the sake of completeness and as an interesting curiosity but, rigorously, estimates can be done only when complete sample ascertainment is done: full variant discovery and all the living samples examined.*

| Tree characteristic | Date est.(yr) | Related historical event |
|---|---:|---|
| TMRCA whole dataset | 170,000 | – |
| Separation A/ rest | 110,000 | Out-of-Africa |
| European clade | 14-24,000 | Post-glacial peopling of Europe |
| Calibration point | **7,700** | Early Neolithic expansion into Sardinia. Date taken as calibration point |
| Hap. G expansion | 5,500 | Late Neolithic expansion of a new migrated group into Sardinia |
| Hap. A private var. | $<$ 2,000 | Entrance of African origin slaves in Sardinia during Roman domination and Vandal invasion |

## 6.6 Ötsi in the tree

The ancient DNA from Ötsi was included into the tree. It fits into G2a2b haplogroup, which encompasses a Tuscan, a Corsican and 8 Sardinian sam-

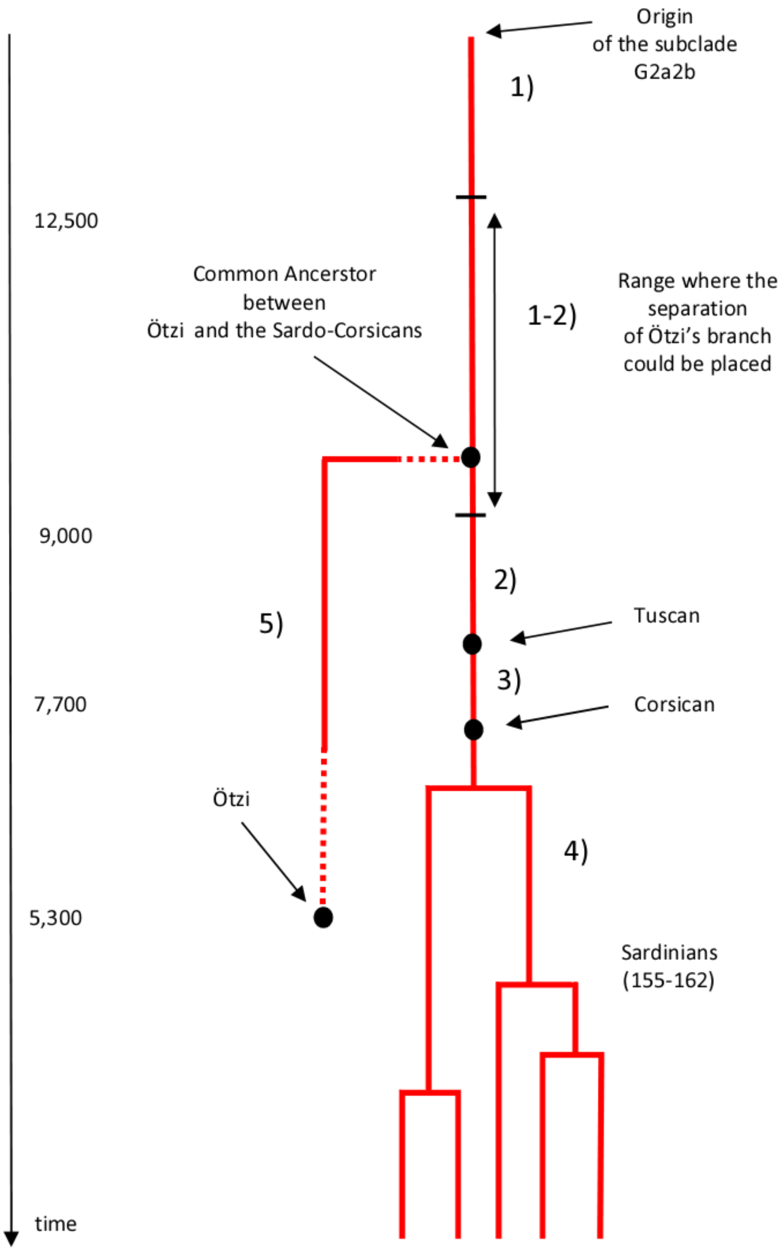


Figure 6.5: Ötsi's phylogenetic tree into G2a2b clade. On the left a tentative time-scale of the branching events. The numbers have the following meanings: 1) root of the subclade (common to all samples), 1-2) uncertainty area where Ötsi lineage may have branched (their genotype in Ötsi is unknown) 2) SNPs common to Tuscan, Corsican and downstream Sardinian Samples 3) SNPs private to the Corsican/Sardinian clade 4) SNPs private to Sardinian samples 5) Ötsi private variability.

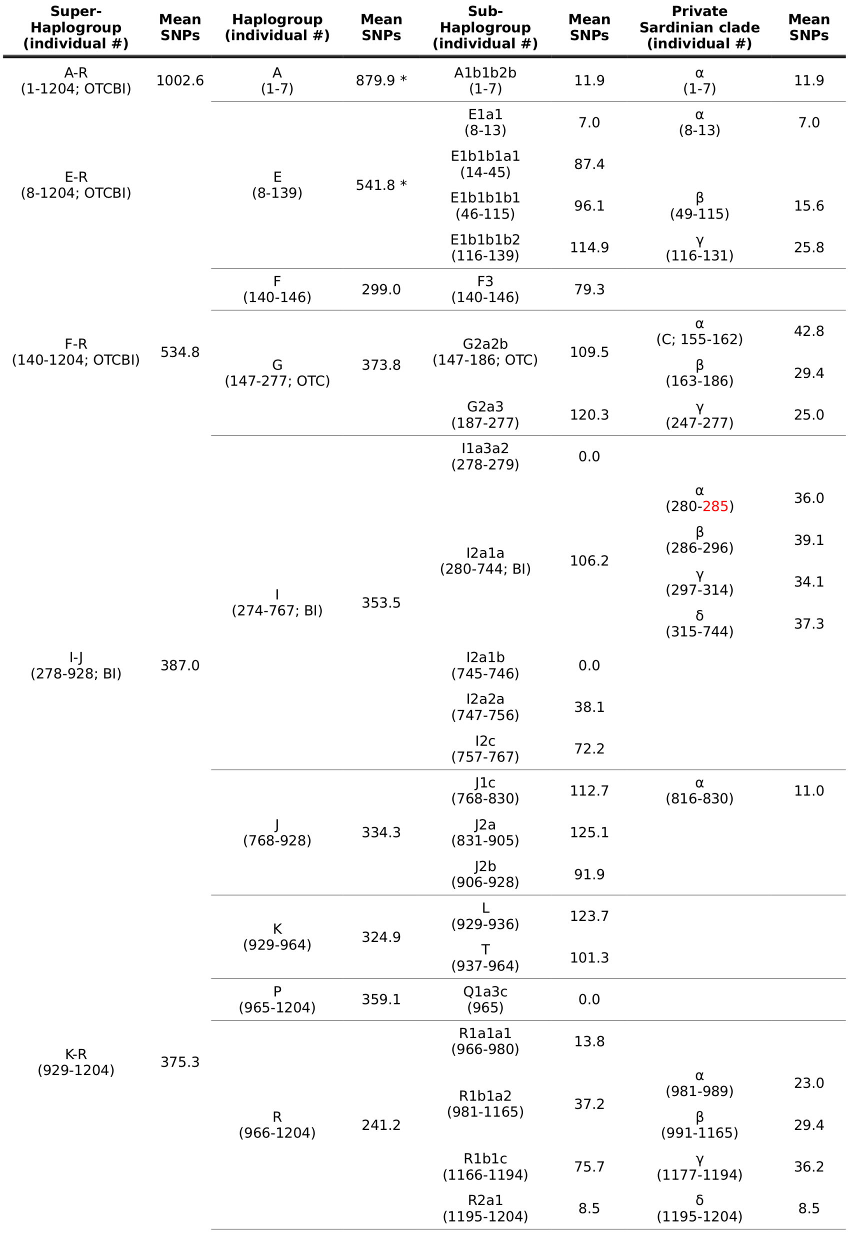
ples. A tree of the subclade is represented in figure 6.5, where the ascertained and unknown allelic states are evidenced. Ötsi goes through the G2a2b clade with 10 derived SNPs, while for the next 17 in the hierarchy its allelic status is unknown. The following 8 SNPs in the hierarchy, instead, separate Ötsi from the Tuscan-Corsican-Sardinian branch. On his own branch the *iceman* has its own private variability spotting 15 singletons, but, as stated before, it is expected that they are much more and errors could be on detected ones. In the tree in the figure aforementioned, it' also possible to see how Tuscan branch separates from Sardo/Corsican clade and where Sardinian private variability begins. Applying the precedently calculated phylogenetic rate, the MRCA between Ötsi and the Sardo/Corsican branch of G2a2b could have lived around 9,000-12,000 years BP. The private variability of Ötsi (15 SNP) would calculate, with the aforementioned mutation rate, converted into a mutation over 169 years, to 2,535 years independent evolution, which tentatively repositions Ötsi around its real era ($\sim$ 6,450 yr BP), which is realistic considering that due to missing variability the number of private variants is underestimated.

## 6.7 Haplogroups discussion

Sardinian samples show a high degree of inter-individual variability. All most common Y-chromosome haplogroups present in Europe are detected in our dataset, with the exception of the N, which is from northern Urals. The distribution of the haplogroups in our dataset is shown in the table reported in figure 6.6. A brief description of each of the detected haplogroups is given in the following subsections. Make reference to the schematic tree reported in figure 6.1.

### 6.7.1 A

The first bifurcation point in the tree separates haplogroup A samples from the others. Haplogroup A is the most basal group for *homo sapiens* of African origin. In our tree it encompasses 7 samples (0.6% of the dataset), all part of the haplogroup A1b1b2b, a sub-Saharan cluster of Y-chromosome lineages,



| Super-Haplogroup (individual #) | Mean SNPs | Haplogroup (individual #) | Mean SNPs | Sub-Haplogroup (individual #) | Mean SNPs | Private Sardinian clade (individual #) | Mean SNPs |
|---|---|---|---|---|---|---|---|
| A-R<br>(1-1204; OTCBI) | 1002.6 | A<br>(1-7) | 879.9 * | A1b1b2b<br>(1-7) | 11.9 | α<br>(1-7) | 11.9 |
| E-R<br>(8-1204; OTCBI) | | E<br>(8-139) | 541.8 * | E1a1<br>(8-13) | 7.0 | α<br>(8-13) | 7.0 |
| | | | | E1b1b1a1<br>(14-45) | 87.4 | | |
| | | | | E1b1b1b1<br>(46-115) | 96.1 | β<br>(49-115) | 15.6 |
| | | | | E1b1b1b2<br>(116-139) | 114.9 | γ<br>(116-131) | 25.8 |
| F-R<br>(140-1204; OTCBI) | 534.8 | F<br>(140-146) | 299.0 | F3<br>(140-146) | 79.3 | | |
| | | G<br>(147-277; OTC) | 373.8 | G2a2b<br>(147-186; OTC) | 109.5 | α<br>(C; 155-162) | 42.8 |
| | | | | | | β<br>(163-186) | 29.4 |
| | | | | G2a3<br>(187-277) | 120.3 | γ<br>(247-277) | 25.0 |
| I-J<br>(278-928; BI) | 387.0 | I<br>(274-767; BI) | 353.5 | I1a3a2<br>(278-279) | 0.0 | | |
| | | | | I2a1a<br>(280-744; BI) | 106.2 | α<br>(280-285) | 36.0 |
| | | | | | | β<br>(286-296) | 39.1 |
| | | | | | | γ<br>(297-314) | 34.1 |
| | | | | | | δ<br>(315-744) | 37.3 |
| | | | | I2a1b<br>(745-746) | 0.0 | | |
| | | | | I2a2a<br>(747-756) | 38.1 | | |
| | | | | I2c<br>(757-767) | 72.2 | | |
| | | J<br>(768-928) | 334.3 | J1c<br>(768-830) | 112.7 | α<br>(816-830) | 11.0 |
| | | | | J2a<br>(831-905) | 125.1 | | |
| | | | | J2b<br>(906-928) | 91.9 | | |
| K-R<br>(929-1204) | 375.3 | K<br>(929-964) | 324.9 | L<br>(929-936) | 123.7 | | |
| | | | | T<br>(937-964) | 101.3 | | |
| | | P<br>(965-1204) | 359.1 | Q1a3c<br>(965) | 0.0 | | |
| | | R<br>(966-1204) | 241.2 | R1a1a1<br>(966-980) | 13.8 | | |
| | | | | R1b1a2<br>(981-1165) | 37.2 | α<br>(981-989) | 23.0 |
| | | | | | | β<br>(991-1165) | 29.4 |
| | | | | R1b1c<br>(1166-1194) | 75.7 | γ<br>(1177-1194) | 36.2 |
| | | | | R2a1<br>(1195-1204) | 8.5 | δ<br>(1195-1204) | 8.5 |

Figure 6.6: Haplogroup description

known to be found sporadically in Mediterranean Europe, but constantly in Sardinia, although at very low frequencies. It shows Sardinian private variability which could be dated less than 2kyr, which is consistent with the entrance of African origin slaves during late Roman domination or Vandalic invasions.

### 6.7.2 E

Haplogroup E is believed to have originated in Eastern Africa, then spread to Mediterranean area and Southern Europe with a peak in the Balkans. In our dataset, E encompasses 132 individuals (11% of the dataset). It is mainly present with the European clade E1b1b1 (E1b1b1a1, E1b1b1b1, E1b1b1b2) characterized by the marker M35 and with a small number of samples, 9, within the African clade E1a (E1a1).

### 6.7.3 F

F is a rare haplogroup. In our tree F is shared by 7 individuals (0.6%). Newly discovered SNPs evidenced a relationship with haplogroup G which shares a brief common history with F.

### 6.7.4 G

Haplogroup G is mostly diffused in the Caucasus area. It is rather common in Sardinia, 131 samples (10,9%), it shows similar frequencies in Corsica [47] and Southern Italy [48]. It shows two subclades of the G2a clade (G2a2b, G2a3). Private Sardinian/Corsican variability is present in subclade G2a2b which is shared with a Corsican sample and with the ancient eneolithic sample for South-Tyrolean Italian Alps named Ötsi [17] (5,300 yr).

### 6.7.5 I

Haplogroup I is the most diffused in Sardinia. It encompasses 490 individuals (40.7% of the dataset). It is mostly represented with the I2a1a clade, denoted by the marker M26. This marker is highly diffused in Sardinia [47] and rare elsewhere [49].

The other clades present are I1a3a2, I2a1b, I2a2a, I2c. Clade I2a1a has 4 lines of clearly private Sardinian variability identified with the names $\alpha$, $\beta$, $\gamma$, $\delta$. I2a1a-$\delta$ is by far the most diffused with 430 individuals, it expanded in Sardinia around 7,700 years BP and its MRCA has been taken as calibration point to calculate the mutation rate.

### 6.7.6 J

Haplogroup J is of putative south-west Asian origin, relatively present in Europe especially in Mediterranean area. In our dataset encompasses 161 individuals (13.4%). It is present with the two main subhaplogroups J1 and J2 (J1c, J2a and J2b), with clade J1c showing some private Sardinian variability (marked as $\alpha$) within 15 individuals.

### 6.7.7 K

Super-haplogroup K (an haplogroup encompassing two other ones) has its putative origins in Southwestern Asia. It is present with 36 samples (3%) in the two branches L (8 samples) and T (28 samples).

### 6.7.8 P

P is a subhaplogroup of K and contains the haplogroups Q and R. It is of putative central/Southern Asia origin and it is commonly distributed worldwide except in African and Oceanian native populations. While R deserves its own paragraph, Q (Q1a3c) is present in just one sample (0.08%).

### 6.7.9 R

Haplogroup R is the second more common haplogroup in Sardinia, it encompasses 239 samples (19.9 % of the dataset). It occurs mostly with the Western European M173-M269 branch R1b (R1b1a2, R1b1c) encompassing 214 individuals. Remaining samples are splitted within two other subclades, R1a1a1 marked by the M17 mutation of Eastern European origin (thought to be linked to the Indoeuropean linguistic expansion) and R2a1. R1b clades show private Sardinian variability with clear expansion figures that together with E and G clades may indicate a Late Neolithic expansion (see table in paragraph 6.5.2).

## 6.8 Future work

The lowering prices of sequencing is now making possible to sequence at high coverage a larger cohort of samples, with a better quality. This will make possible to ascertain almost all missing variability on rare variants and singletons.

By deep sequencing the trios of the two Sardinian sequencing cohorts will also allow a precise evaluation of molecular mutation rate within the huge number of pedigrees collected by the two Sardinian studies and to shape its dependence on paternal age.

On the archaeological side a lot of improvements can be done using the rich heritage of Sardinian prehistorical sites. Most of them were already radiocarbon dated. With forensic techniques is already possible to extract their DNA and sequence it and to separate contaminant contributions from real ancient DNA, exploiting *post mortem* conversion of cytosine into uracil, which is not present into contaminants, obtaining really clear results. Combining dates and sequences will enable refinements of the mutation rate, enable a phylogeny on an ancient snapshot and could refine hierarchical position of variants into the tree.

On this data further investigations on mitochondrial line and whole genome are already on the way and the previously described future work is already at an advanced planning status.

# Conclusions

In this work we performed a phylogenetic analysis, making use of 1,204 male samples of Sardinian ancestry. We extracted the male specific portion of Y chromosome for all the samples and built a phylogenetic tree on top of them. We calculated MSY mutation rate using the rich archaeological information about Sardinia which testifies a population expansion 7,700 years before present. We calculated the time to most recent common ancestor of the dataset, finding it around 170,000 years BP. We found that Sardinian population encompasses most of European genetic variability on Y chromosome, but also that each of the main branches shows sign of expansion with, as a consequence, a great number of private Sardinian variants. Future work is planned that will enhance even more our knowledge of past events, involving also mitochondrial DNA and whole genome analysis. Our results were published on a peer reviewed journal in August 2013 [46].

As a closing note, we care to point out the flexibility of sequencing data. Sardinian samples were meant for completely different analyses with respect to ours. Analyses which are still under way and giving successful results. And the same cohort of data will take part, in the future, to new and maybe even yet unplanned analyses, that will make use of sequences and of the abundance of metadata that were collected with them. Sequencing has huge costs, in terms of money and required work. But, once collected, data will be a goldmine for several years turning the huge costs into a smart looking-forward investment.

# Appendix A

# Data management and Workflow details

## A.1 SEAL 0.3.1

Seal[21] is an alignment tool built at CRS4 to implement a fast, parallel and scalable strategy for DNA sequencing alignment.

The tool is specifically designed for Illumina data. It implements the BWA (Burrows-Wheeler-Aligner)[20] tool in its 0.5.9 version and uses the Apache Hadoop parallel environment.

**bwa**: bio-bwa.sf.net

**hadoop**: hadoop.apache.org

It implements the following tools:

- Demux: demultiplexing, splits samples from a multiplexed run
- PairReadsQSeq: takes the qseq files where reads for read 1 and read 2 are stored separately and merge R1 and R2 to integrate the alignment process
- Seqal: performs the Burrows Wheeler alignment directly on the read pairs generated by the PairReadsQSeq tool.
  It emulates the behaviour of BWA and of Picard MarkDuplicates tool, which removes PCR duplicate reads.
- ReadSort: Sorts reads by chromosomal coordinates, generally needed for all tools downstream the alignment.

**Picard tools**: picard.sf.net

- RecabTable: Emulates the behaviour of GATK CountCovariates [23] which creates the recalibration table for bam recalibration.

**GATK**: broadinstitute.org/gatk

The advantage of this tool is located into its scalability, since it splits the jobs on its own without any need of intervention. Processing machines can be added or removed to the cluster seamlessly and the parallel environment manages to re-allocate failing processes, which is usually a major problem when scaling to parallel architectures.

This tool has been used for the alignment of all paired-end sequencing samples included in this work. Single end experiments are not supported and have been aligned with a standard version of bwa-0.5.9.

This tool has been developed with original contribution of the author who recently started working on porting the new bwa versions into Seal.

## A.2 LIGA

LIGA, meaning Laboratory Interface for Genomic Analysis is a software package, developed to manage sequencing raw data, analysis results, quality control and data access for both the laboratory operator and for the bioinformaticians.

It is constituted by:

- Database - mySQL based
- Web Interface - PHP based
- Terminal Interface - bash based with a PHP-based REST service dumping database information

**mySQL**: an open-source widely used database engine - mysql.com

**PHP**: a scripting language for web pages - php.net

**REST**: a data transfer system - drupal.org/project/rest_server

### A.2.1 LIGA usage Workflow

- Laboratory operators load their own samplesheets (text files defining samples and their multiplexing configuration) while initiating a sequencing run on an Illumina machine. This step is performed via a web-based interface.

- For each sample more information can be optionally stored in the database such as gender, affection status, family code (useful when sequencing pedigrees)

- LIGA records the run and with the terminal interface provides automated facilities for bio-informatics analysis, to launch the alignment pipeline.

- Once the pipeline has terminated it launches the QCTool ([22] App. B), collects its quality control results and automatically uploads them via the REST interface to the database.

Quality data are both viewable with the web-interface and available as parameters for search and filtering samples.

Processed files are viewable on the web-interface with an embedded IGV (Interactive Genomic Viewer) [50] which enables to see with a graphical interface individual sequence files aligned to the reference genome.

IGV: broadinstitute.org/igv

Sample and experiments paths for both raw data and processed files are available via the terminal interface. Data can be selected by project, by type (DNA,RNA,EXOME), by gender or other optional parameters and by quality.

This tool was used for the process management in the sequencing experiments described in this work and as a basic substrate for all operations requiring access to sample information.

As it is an unpublished tool, major credits for the work go to Atzeni R (main author), Berutti R, Reinier F.

# Appendix B

# QCTool

QCTool software[22] is a quality control tool developed by the author and some of the collaborators to this thesis work.

It is a versatile quality control software for the analysis of sequence alignment results. It is able to assess a wide number of parameters which are fundamental for quality assesment of most common DNA/RNA/Exome sequencing applications.

Reporting features include XML/text output for LIMS and automated systems such as LIGA (Appendix B) and PDF reporting to deliver easily readable run diagnosis documents to non-bioinformatician users.

The software was originally developed to perform quality checking of the sequencing data of the two Sardinian sequencing projects described in chapter 3, whose sequence data constitute the Sardinian dataset of this work.

Its input is a BAM file (binary alignment format), which is a binary file containing reads along with their qualities and alignment information. On top of that it can generate quality metrics.

With reference information or target region information the tool can provide more precise coverage assessment and, in case, detailed statistics on probe sets.

QCTool is included as appendix in this document as it is an original contribution of the author and as it constituted a fundamental tool for the analysis pipeline.

## B.1 Features

### B.1.1 Scalar parameters

In the following table the parameters that can be calculated by the software are listed and briefly explained.

| Parameter group: | Description: |
|---|---|
| Experiment properties | Read Len (average if several lengths), Single End/Pair End status, Insert size, recalibration status |
| Quantitative evaluation | Coverage vs. reference (nominal length/real length), Depth, Percentage of reference covered at least once. |
| Counters | Number of reads in BAM, number of aligned reads, number of not-aligned reads, PCR duplicates, number of clipped bases / clipped reads, number of total bases. |
| FlagStat | Statistics on read flags (produces identical results as samtools flagstat) |
| Alignment statistics / Errors | Mapping rate, mismatch counter and rate (mismatching bases/bp), percentage of reads perfectly aligning to the reference (no mismatches), blank bases counter and rate, counter of reads sporting blanks. Counter of reads with bad insert size (not properly paired). Recognition and counting of badly oriented pairs. |
| Qualities | Average Mapping quality, percentage of reads with map quality zero, average base quality, count and percentages of bases Q< 10, Q>= 20, Q>= 30 |
| Base content | Relative base content %A, %C, %G, %T, %N is computed and compared to reference genome base content |
| Extra features | Detection of Human genome and, if so, gender recognition, ploidy estimation for each chromosome |

### B.1.2 Plots

Beyond scalar parameters, plots are produced out of input sequencing data to describe cycle dependent parameters (mostly quality) and distributions. In the following table the main plots produced by QCTool are listed:

| Plot: | Description: |
|---|---|
| Original base quality | Original base quality ( if present recalibrated BAM ) by cycle |
| Base quality | Reported base quality (recalibrated, if present) by cycle |
| Original base quality distribution | Distribution of original base quality scores ( if present recalibrated BAM ) |
| Base quality distribution | Distribution of reported base quality scores |
| Original / recalibrated base quality | Correlation between original and recalibrated quality scores |
| Recalibrated / empirical base quality | Correlation between recalibrated and empirical base qualities (empirical calculated accounting reference mismatches) |
| Mapping quality distribution | Distribution of assigned (alignment software) mapping qualities. |
| Insert size distribution | Distribution of the inferred size of the fragments |
| Blanks per cycle | Number of blanks by cycle |
| Blanks per read | Distribution of blank count in the reads |

# Appendix C

# Variant Calling Software

Variant calling is an essential part, not only in a phylogenetic analysis but as the final step for most sequencing experiments.

In this appendix we will present the variant calling mini-pipeline, the modifications made to existing tools to implement Y chromosome specific filtering and to achieve higher performance by means of parallelisation.

## C.1 GLF Multiples

The glfMultiples package is a variant calling tool [26]. It is an open source software developed at the University of Michigan.

It is tailored to handle GLF files. GLFs, for each sample, hold the genotype marginal probabilities, i.e., for each locus the probabilities that any possible genotype combination (on a diploid genome) is true are stored. Probabilities reported into GLFs are directly computed using alignment data without further processing.

This tool recalibrates posterior probabilities with priors based on the panel, such as transitions to transversions ratio and by maximizing the likelihood of the observed bases with respect to allele frequency.

## C.2 Customization

The glfMultiples tool, itself, was not meant for Y chromosome, thus we customized it adding the following:

- At genotype level:
  - Recalibrating mutation posterior with prior probability ($10^{-3}$) of a mutation (1 over 1000 bases). Get rid of all genotypes that are changed by the penalty.
  - Removing heterozygous base calls
  - Map quality and base quality filters
  - Require min depth
- At locus level:
  - Allow heterozygous call for max N samples, otherwise get rid of locus
  - Require min depth for singletons
- At panel level:
  - Map quality and base quality average filters
- Output level:
  - Output a panel format with explicit genotype calls

All the previous filters can be set and unset through the command line. This made simple filter tuning.

## C.3 Parallelization and Performance

The implementation of glfMultiples plus the filters runs on a single thread (single process, unbranched on a single core). It can easily get messed up with the 1,200 samples and took several hours to process the whole dataset with all filtering facilities activated.

A soft parallelization approach was taken. The main executable and the algorithm remained almost untouched, except for some optimization to drop unnecessary features and calculations, and to optimize for a big number of input files.

A wrapper pipeline was built around to split GLF files in small chunks to allow parallel processing and a RAM drive memory was used as a temporary storage for chunks (this allowed an average 10 times faster processing speed).

**RAM drive**: a virtual storage which is allocated on the RAM (random access memory) of the computing machine. This is generally ten time faster than a traditional hard drive.

The pipeline waits for the chunks processing to terminate and merges back the results applying whole-panel level filters which are deactivated in the single-chunk processing.

Quality controls are then performed on the genotypes, with a last step check which calculates statistical parameters such as number of calls per sample, number of variants, ts/tv, quality, etc.

Performance optimization reduced computational speed of around two orders of magnitude, around 10x contribution comes from ram drive optimization and, the other, linear, comes from chunking and it is limited only by the panel level filtering that must be done monolithically at the end.

# Bibliography

[1] Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C, "A calibrated human Y-chromosomal phylogeny based on resequencing", Genome Res, 23(2): 388395, (2013), doi: 10.1101/gr.143198.112

[2] Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiæ M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA, "The Genetic Legacy of Paleolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective", Science 290, 1155 (2000), doi:10.1126/science.290.5494.1155

[3] Cavalli-Sforza LL and Edwards AWF, "Phylogenetic analysis: models and estimation procedures", American Journal of Human Genetics 19:233257 (1967)

[4] Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL., "High resolution of human evolutionary trees with polymorphic microsatellites", Nature. 1994 Mar 31;368(6470):455-7

[5] Sanger F, Nicklen S, Coulson AR, "DNA sequencing with chain-terminating inhibitors", PNAS 74 (12): 54637 (1977)

[6] Schena M, Shalon D, Davis RW, Brown PO, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science, (1995) 270(5235):467-70

[7] Bentley DR et al, “Accurate whole human genome sequencing using reversible terminator chemistry”, Nature (2008);456(7218):53-9. doi: 10.1038/nature07517

[8] Cann RL, Stoneking M, Wilson AC (1987), “Mitochondrial DNA and human evolution”, Nature 325 (6099): 3136, doi:10.1038/325031a0

[9] Soares P, Ermini L, Thomson N, et al. “Correcting for purifying selection: an improved human mitochondrial molecular clock” AJHG 84 (6): 74059 (2009), doi:10.1016/j.ajhg.2009.05.001

[10] Olivieri A et al, “The mtDNA Legacy of the Levantine Early Upper Palaeolithic in Africa”, Science, Vol. 314 no. 5806 pp. 1767-1770 (2006), DOI:10.1126/science.1135566

[11] Skaletsky, H et al., “The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes”, Nature 423, 825837 (2003).

[12] The Y Chromosome Consortium, “A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups”, doi: 10.1101/gr.217602 Genome Res. 2002. 12: 339-348, Cold Spring Harbor Laboratory Press, `http://genome.cshlp.org/content/12/2/339.full`

[13] ISOGG - The International SOciety for Genetic Genealogy, `http://www.isogg.org/tree`

[14] Illumina `http://www.illumina.com`

[15] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. “dbSNP: the NCBI database of genetic variation”, Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

[16] The dbSNP database `http://www.ncbi.nlm.nih.gov/SNP`

[17] Keller A et al., “New insights into the Tyrolean Iceman’s origin and phenotype as inferred by whole-genome sequencing”, (2012) Nature Commun. 3 (2): 698. doi:10.1038/ncomms1701. PMID 22426219.

[18] Genome Reference Consortium,
`http://www.ncbi.nlm.nih.gov/projects/genome/assembly/ --→`
`--→ grc/human/`,
Human Reference GRCh37,
`ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/ --→`
`--→ vertebrates_mammals/Homo_sapiens/GRCh37`

[19] Dean J and Ghemawat (Google Inc.), “MapReduce: Simplified Data Processing on Large Clusters”, OSDI ’04, Pp. 137150 of the Proceedings,
`https://www.usenix.org/legacy/event/osdi04/tech/ --→`
`--→ full_papers/dean/dean.pdf`

[20] Li H and Durbin R, “Fast and accurate long-read alignment with Burrows-Wheeler Transform”, (2010) Bioinformatics, Epub. [PMID: 20080505], `http://bio-bwa.sourceforge.net`

[21] Pireddu L, Leo S, Reinier F, Berutti R, Atzeni R, Zanetti G, “Scaling with the flow: advantages of a MapReduce-based scalable and high-throughput sequencing workflow”, International Conference on Human Genetics (ICHG) 2012, Montreal, Canada, `http://biodoop-seal.sourceforge.net`

[22] Berutti R, Reinier F, Atzeni R, Angius A, Cusano R, Marcelli M, Oppo M, Pilu R, Urru MF, Valentini M, Zara I, Sanna S, Cucca F, Jones C, “QCTool: an efficient toolkit to automatically generate quality metrics of next-generation sequencing data”, ESHG Nurenberg (2012), `http://www.berutti.net/public/qctool`

[23] DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D and Daly, M, “A framework for variation discovery and genotyping

using next-generation DNA sequencing data", (2011) Nature Genetics. 43:491-498, http://www.broadinstitute.org/gatk

[24] Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis G, Boehnke G and Kang HM, "Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data", American journal of human genetics doi:10.1016/j.ajhg.2012.09.004 (volume 91 issue 5 pp.839 - 848), http://genome.sph.umich.edu/wiki/VerifyBamID

[25] GLF File format Appendix to SAM file format specification, http://samtools.sourceforge.net

[26] GLF Multiples software package, University of Michigan, Center for Statistical Genetics, http://genome.sph.umich.edu/wiki/GlfMultiples

[27] The 1000 Genomes Project Consortium, http://www.1000genomes.org

[28] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing", Nature 467, 10611073 (28 October 2010) doi:10.1038/nature09534, http://www.nature.com/nature/journal/v467/ --→
--→ n7319/full/nature09534.html

[29] Phylip software, University of Washington, http://evolution.genetics.washington.edu/phylip.html

[30] Fitch WM, "Toward defining the course of evolution: minimum change for a specified tree topology", Systematic Zoology 20 (4), 406-416 (1971)

[31] Newick format
http://evolution.genetics.washington.edu/phylip/ --→
--→ newicktree.html

[32] Harris RS, "Improved pairwise alignment of genomic DNA", Ph.D. Thesis, The Pennsylvania State University (2007) http://www.bx.psu.edu/~rsharris/lastz

[33] Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E, “EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates”, Genome Res. 19, 327335 (2009). doi:10.1101/gr.073585.107 Medline

[34] Chimpanzee reference panTro4, UCSC Genome Browser, `http://genome.ucsc.edu/cgi-bin/hgGateway?org=Chimp`

[35] Pilia et al, “Heritability of cardiovascular and personality traits in 6,148 Sardinians”, PLoS Genet. 2 8): e132. (2006) doi:10.1371/journal.pgen.0020132

[36] Pugliatti M, Rosati G, Carton H, Riise T, Drulovic J, Vcsei L, Milanov I, “The epidemiology of multiple sclerosis in Europe”, European Journal of Neurology 2006 Jul;13(7):700-22

[37] Sondaar P, et al., “The human colonization of Sardinia, a Late-Pleistocene human fossil from Corbeddu cave”, Comp. Rend. Acad. Sci. Paris 320, 145150 (1995)

[38] G. Tanda contribution into:
B. Ferrer et al., “Iberia e Sardegna, Legami Linguistici, Archeologici e Genetici dal Mesolitico allEt del Bronzo”, Le Monnier, Firenze, Italy, 2013, pp. 234249

[39] Aimar A, Giacobini G, Tozzi C, “Trinitá d’Agultu (Sassari). Localitá Porto Leccio”, Archaeological Bulletin, 45, 83-87 (1997), Ministry of Cultural Heritage, Italy

[40] Tykot RH contribution into:
Skeates R, Withehouse R (editors), “Radiocarbon Dating and Italian Prehistory”, Accordia Specialist Studies on Italy, London, 1994, pp. 115145

[41] Lugliè C contribution into:
Lugliè C, Cicilloni R (editors), “Atti della XLIV Riunione Scientifica

La Preistoria e la Prototoria della Sardegna, Vol. 1”, Istituto Italiano di Preistoria e Protostoria, Firenze, Italy, 2009, pp. 3747

[42] Tanda G contribution into:
Luglié C, Cicilloni R (editors), “Atti della XLIV Riunione Scientifica La Preistoria e la Prototoria della Sardegna, Vol. 1”, Istituto Italiano di Preistoria e Protostoria, Firenze, Italy, 2009, pp. 59-78

[43] Tykot H, “Characterization of the Monte Arci (Sardinia) obsidian sources”, Journal of Archaeological Science 24, 467479 (1997). doi:10.1006/jasc.1996.0130

[44] Dyson SL, Rowland RJ Jr, “Shepherds, Sailors & Conquerors. Archeology and History in Sardinia from the Stone Age to the Middle Ages”, University of Pennsylvania, Museum of Archaeology and Anthropology, Philadelphia, PA, 2007

[45] Pala M et al, “Mitochondrial Haplogroup U5b3: A Distant Echo of the Epipaleolithic in Italy and the Legacy of the Early Sardinians”, AJHG, Volume 84, Issue 6, 814-821, (2009), doi:10.1016/j.ajhg.2009.05.004

[46] Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, Sanna D, Useli A, Urru MF, Marcelli M, Cusano R, Oppo M, Zoledziewska M, Pitzalis M, Deidda F, Porcu E, Poddie F, Kang HM, Lyons R, Tarrier B, Bragg Gresham J, Li B, Tofanelli S, Alonso S, Dei M, Lai S, Mulas A, Whalen MB, Uzzau S, Jones C, Schlessinger D, Abecasis GR, Sanna S, Sidore C, Cucca F, “Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs European Y-Chromosome Phylogeny”, Science 341, 565 (2013) DOI: 10.1126/science.1237947

[47] Francalacci P, Morelli L, Underhill PA, et al, “Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability”, (Jul 2003) American Journal of Physical Anthropology 121 (3): 2709. doi:10.1002/ajpa.10265. PMID 12772214

[48] Capelli C, Brisighelli F, Scarnicci F, et al., "Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter", (Jul 2007) Molecular Phylogenetics and Evolution 44 (1): 22839. doi:10.1016/j.ympev.2006.11.030. PMID 17275346

[49] Rootsi S et al, "Phylogeography of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe", American Journal of Human Genetics 75, 128137 (2004)

[50] Robinson JT, Thorvaldsdttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP, "Integrative Genomics Viewer", Nature Biotechnology 29, 2426 (2011)
`http://www.broadinstitute.org/igv`

# Data access and notes

### Published material

The work described in this thesis has been subject of a paper published in August 2013 on Science. See reference [46] on Bibliography.

### Data access

Original sequencing data for the MSY portion of the Y chromosome for all the 1204 samples has been deposited at the EGA Europen Genome-phenome archive (`http://www.ebi.ac.uk/ega`) with accession number EGAS00001000532.

### Disclaimer notes

The present study was approved by the Ethical committees of the ASL 6 in Lanusei and ASL 1 in Sassari, and by the IRB of the National Institute on Aging, NIH. Each participant signed an informed consent form for the samples used.

### Funding

This research was supported in part by NIH contract NO1-AG-1-2109 from the National Institute of Aging to the IRGB institute; by Sardinian Autonomous Region (L.R. no. 7/2009) grants cRP3-154 to Cucca F and cRP2-597 to Francalacci P; by Fondazione Banco di Sardegna grants to Francalacci P and Morelli L; by Basque Government grant G.I.C. IT-542-10 to Alonso S; and by National Human Genome Research Institute grants HG005581, HG005552, HG006513, and HG007022 to Abecasis GR

# Acknowledgements

I would really like to thank all people who contributed materially or with their support to this thesis work.

Firstly I would like to thank my tutor, Prof. Francesco Cucca, my co-tutor Dr. Chris Jones, and Prof. Paolo Francalacci, for the precious discussions and arguings which contributed a lot to the quality of the final work, for your human qualities and the good time spent, and last but not least, for having taught me a lot: **thank you**.

A special thought goes to Laura Morelli, who was another protagonist of this work and sadly passed away early this year. For the precious scientific discussions, for you support and for your great human qualities: **thank you**.

Thanks to all samples donors and the personnel involved into the Sardinian sequencing experiments for your valuable work: **thank you**.

Special thanks goes to the CRS4 sequencing team for the hard work and the enjoyable free time, Rossano Atzeni, Roberto Cusano, Manuela Oppo, Alessandro Puddu, Ilenia Zara, Frederic Reinier, Rosella Pilu, Maria Valentini, Andrea Angius and any other I could have forgotten: **thank you**.

Thanks to the CRS4 IT team who supported a lot and shared lots of enjoyable, nice and painful moments “we lost data!”-“oh no!”, the latest always brilliantly and professionally solved. Michele, Marco M, Marco P, Antonio, Muriel, Carlo I, Carlo P, Franco, Luca C, Luca P, Luca P(2), Katia, Lidia: **thank you**.

Thanks to the CNR-IRGB team for the support, scientific discussions, for having taught me a lot and for the nice time spent together: Maristella, Carlo, Serena, Eleonora, Elisabetta, Michael, Sonia, Francesca, Luca, Laura, Manuela, Luiz and many others: **thank you**.

Thanks to Prof. Thomas Meitinger and Dr. Holger Prokisch of the Institute for Human Genetics of the Helmholtz Institute in Munich who supported and funded me during the writing of this work and for the nice way you welcomed me: **thank you**.

Thanks to all the other people that contributed scientifically or with their nice presence to make these last three years a valuable and enjoyable time: **thank you**.

**Thanks to my friends, some of them are already mentioned, to my girlfriend and to my family: without you all this time would have lost most of its taste. Thank you!**

And thanks to any reader who had read this work, but also to whom had just opened this thesis with the purpose of reading this page: please write down your name here, you deserve it! _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ : **thank you!**.