UNIVERSITÀ DEGLI STUDI DI SASSARI

Scuola di Dottorato in Scienze Biomediche

XXV CICLO DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE
INDIRIZZO DI GENETICA MEDICA, MALATTIE METABOLICHE E NUTRIGENOMICA

Direttore: **Prof. Eusebio Tolu**

Relatore: **Prof. Francesco Cucca**

Detecting the Sardinian Specific Variability Trough Next Generation Sequencing of 2120 Individuals

Relatore

Prof. Francesco Cucca

Dottorando:

Carlo Sidore

Anno Accademico 2012-2013

Index

1. Introduction

Genome Wide Association Studies (GWAS) have increasingly furthered our understanding of the molecular basis of many complex traits by finding, through genotyping and imputation, loci associated with many different traits. However, studies based on variants present in common genotyping arrays and imputation panels may not capture the fraction of human genome variation that is rare or geographically restricted and unique to specific populations.
Here we propose the analysis of the genome of 2120 Sardinian individuals generated with next generation sequence approach and we will show the benefits, in terms of resolution and investigation, to the analysis of the Sardinian specific variability.

1.1. From Genome Wide Association Studies to Next Generation Sequencing, motivations

Genome Wide Association Studies (GWAS) are methods to investigate associations with complex traits or diseases by scanning Single Nucleotide Polymorphisms (SNPs). SNPs are genotyped using dense arrays having hundreds of thousands of probes distributed along the genome. GWAS are based on two

main assumptions: the first one is that common variants with low effects are responsible of common diseases. Therefore, to detect such variants, we need a large number of individuals (typically greater than 6000) to achieve enough statistical power. The second assumption is that, by genotyping a reduced amount of SNPs (*tag* SNPs) along the genome, it is possible to access other markers by linkage disequilibrium (LD). Human genome, in fact, constitutes of haplotypes blocks and by typing a single SNP we can reconstruct the genotypes of other SNPs in the same block of LD by using statistical methods like imputation.

Since 2005, GWAS studies reported more than 1000 SNPs associated with more than 300 complex traits or diseases[1], testifying the validity and the worldwide popularity of this approach. However, the approach has two important limitations that may reduce the statistical power in detecting associations. The first limitation comes from the genotyping arrays, the SNPs present in the arrays are selected from populations included in large international projects like HapMap[2]. When investigating isolated population, like Sardinians, it is possible that the assumption of LD disequilibrium between tags and the remaining SNPs is violated in some regions. In this case, the accuracy of imputation will be affected as well as the statistical power of detecting the associations. The second limitation of GWAS comes from the analysis of rare variants. Rare variants have lower LD with the surrounding SNPs and it is much harder to access them by using tag SNPs. Furthermore rare variants are, in general, population specific and it is much more likely that they are completely missed from GWAS analysis, since they are not included in commercial arrays or reference panel.

In opposition to GWAS, Next Generation Sequencing (NGS) approaches aim to overcome these limitations by sequencing the complete genome. The approach is based on sequencing short read fragments, aligning them to the human reference genome and then analyzing every position to detect the

positions that differ from the reference. The variants discovery is not dependent on the technology and potentially it is possible to genotype all the variants in the dataset without recurring to tag SNPs or imputation. We will use this approach to sequence a total of 2120 Sardinian individuals and generate a better map of the variability present in the Sardinians. The sequenced samples are selected from two larger cohorts of Sardinian individuals.

1.2. The SardiNIA project

The SardiNIA[3] project aims to investigating the aging in the Sardinians by analyzing more than 300 quantitative traits to evaluate how they correlate with age. The phenotypes encompass a large set of traits, hematological (blood cells count, lipid levels etc), anthropometric (waist circumference, height etc) but also personality and are collected every 3 years to understand their variation during lifetime. More recently, the study focused its attention to the count of different cell types by means of fluorescence-activated cell sorting (FACS).

The study started in 2001 and currently involves almost 7000 individuals from the Lanusei valley in the most isolated part of eastern Sardinia. The samples have been collected in families, so far almost 1000 pedigrees are available, they are up to 5 generations deep and the largest family has more than 625 genotyped individuals. The idea is to exploit the familial relationships to ease the analysis of rare variants shared between individuals of the same family having similar phenotype.

So far, the study genotyped the individuals using whole genome GWAS arrays (Affymetrix 6.0, Affymetrix 500k, OmniExpress 750k) and fine mapping

arrays (Metabochip, Exomechip, Immunochip). Imputation was performed using the previously available reference panel like HapMap or 1000 Genomes.

The study has given a large contribution to the GWAS literature by publishing more than 60 papers since 2006, confirming the quality and validity of the analysis and data produced. After the start of the sequencing project, the project is focusing on the analysis of rare and Sardinian specific variants responsible for variations in the available phenotypes.

1.3. Autoimmunity in Sardinia

The autoimmunity project is a large genotyping project on Sardinia individuals motivated by the high incidence of autoimmune diseases in Sardinia. The study focuses its attention to Type-1 Diabetes (T1D) and Multiple Sclerosis (MS).

Different studies of Multiple Sclerosis showed that, in general, northern European populations have higher incidence of MS. However the Sardinia is a complete outlier in the European distribution, being one of the regions with the highest incidence and with an extremely large difference when compared to other regions in the Mediterranean area[4] (see Figure 1 in the next page). Similar results can be found when analyzing the incidence of T1D.

The study involves 8000 individuals from Cagliari and Sassari, enrolled in a case control study with ~2000 T1D cases, ~3000 MS cases and ~3000 controls. A strong requirement is that individuals must have 3 of the 4 grandparents born in Sardinia, to include in the study only individuals having Sardinian origins.

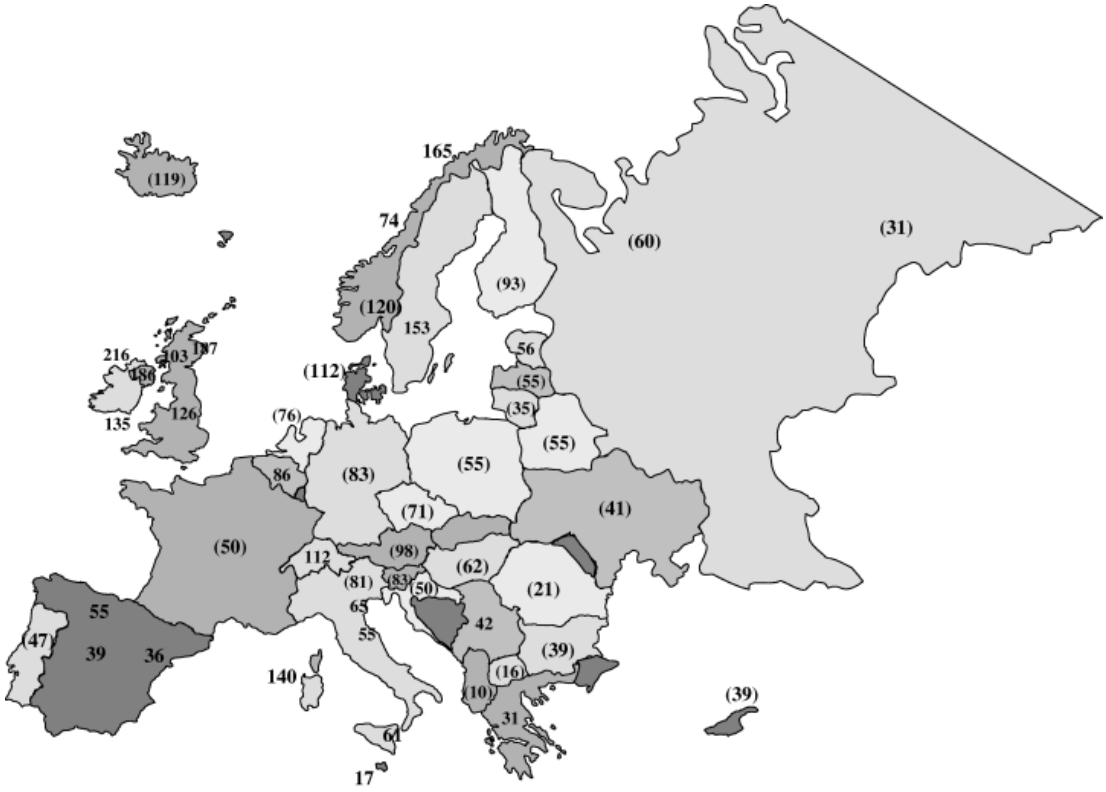

Figure 1: Prevalence (per 100,000 individuals) of MS in Europe[4]

Samples, so far, have been genotyped using GWAS array Affymetrix 6.0 or Immunochip and have been imputed using HapMap and 1000 Genomes reference panel. Results detected the HLA signals, as expected, detected association signals with higher effect size in Sardinians compared to other populations (i.e. CBLB[5]) and were able to replicate some of the other loci found by other group studying the same disease. The study will benefit of the sequencing project in the investigation of rare variants and in the better genotyping of variants having different LD and frequencies between Sardinians and the 1000G populations. Furthermore, sequencing the HLA region, will give better insights of the HLA structure in the Sardinian population.

2. Next Generation Sequencing: data generation

In this chapter we will describe the data generation and the sequencing of our samples, highlighting the complex procedures needed to generate the complete dataset of sequenced individuals. Sequencing of our samples has been performed in parallel at the sequencing core of the University of Michigan and at the sequencing center of the CRS4 in Pula, Sardinia, using, in both cases, the Illumina technology.

2.1. How sequencers work: Illumina technology

We sequenced our samples using Illumina GAII and HIseq sequencers, based on a technology that reads pair-end of short DNA fragments.

DNA strands are sheared by means of sonication and adapters are ligated to each DNA fragments (Figure 2a).

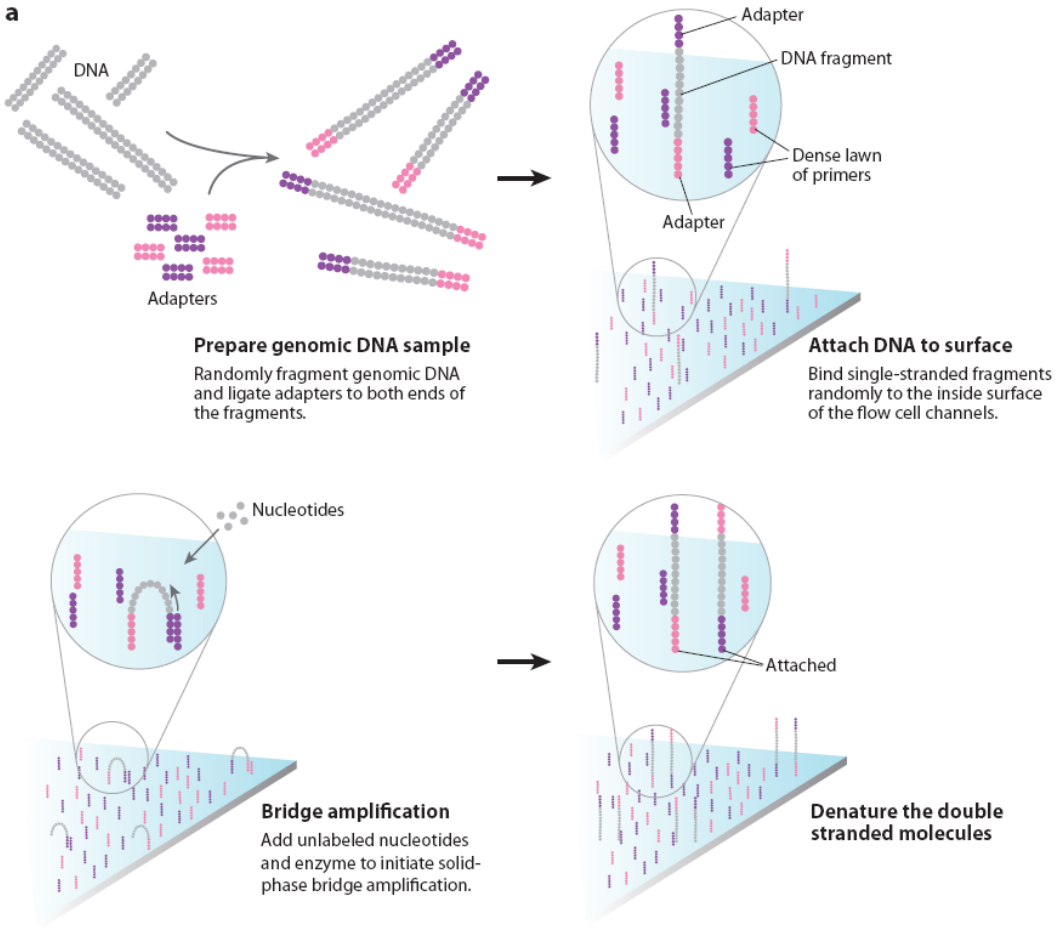

Figure 2: Schematic of Illumina sequencing technology.

Single stranded DNA fragments adapters bind to the flowcell (Figure 2b) and they are bridge-amplified creating, for each fragment, a copy of the same fragment on the opposite strand (Figure 2c).

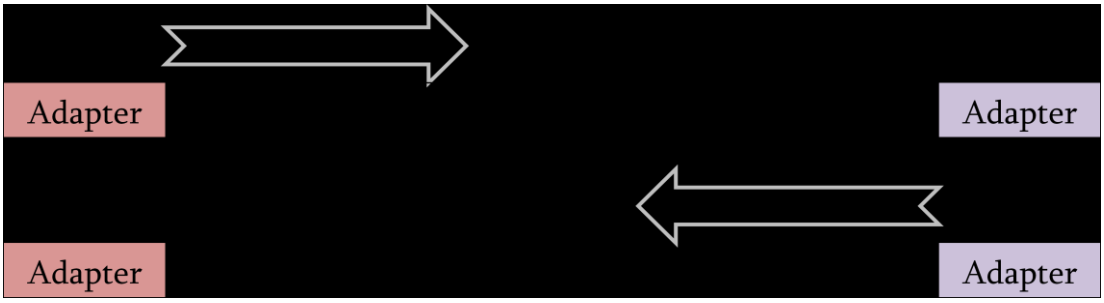

Figure 3: schematic of the two reads of each fragment in the flowcell

At this point each of the two copies of the fragment is read independently base by base (Figure 3). At the end of the process we will have two non overlapping reads for each fragment which will map in the same genomic region after being complemented to be in the same strand.

2.2. Alignment, recalibration, quality checks

Data produced by the Illumina raw data processing need to be further analyzed to be used for genetic analyses. Here we will describe the procedure to generate individual level Binary Alignment/Map(BAM) file starting from Illumina raw data.

2.2.1. Mapping to the human reference genome

Raw data are generated by reading small fragments of DNA without knowing their position in the genome. The first step of analysis is the mapping to the reference genome that we performed using the software BWA[7]. We used, as genomic reference, the same file used by 1000 Genomes project, it is based on genomic build 37 and contains decoy sequences to remove problems due to hard-to-map regions. The mapping procedure consists in indexing the reads and mapping them to the reference. The mapping exploits the pair-end sequencing to increase the mapping accuracy, by verifying that both ends of the fragment map in the same genomic region. Aligned data are stored in BAM files, the standard for storing and quick-accessing the aligned data for next generation sequencing.

2.2.2. Duplicate removal

The next step of the sample preparation is the PCR duplicate removal. Illumina sample preparation involves DNA amplification by PCR, which generates errors in DNA fragments that may be amplified resulting in multiple copies carrying the same error. Even if it is not possible to recognize the PCR errors, it is possible however to identify multiple PCR copies of the same fragment. In fact, copies of the same fragment will be mapped to the same position in the genome reference. If two or more copies with the same genomic position are detected, only the copy with the highest mean quality is kept and the remaining copies are discarded from further analyses.

2.2.3. Recalibration

Last step of sample preparation is the quality score recalibration. Prior to this step, Illumina pipeline assigns a quality score to each base proportional to the quality of sequencer reading. For variants detection, however, it is important to know a better estimation of the true per-base-error rate which may be obtained by recalibration. Recalibration combines together bases having the same features (same sequencing cycle, same dinucleotide bases, same Illumina quality score) and compare them to the reference genome. To avoid bias to the error rate due to real mutations, positions known as SNPs are discarded. For each group of variants, the recalibration algorithm evaluates the mismatch rate with the reference and assigns a recalibrated phred score to all variants in the group.

2.2.4. Quality and identity checks

Finally, we performed quality controls on the processed bam files, evaluating the recalibrated quality, the number of reads and the mean depth. In case of mean base quality <20 or depth lower than 2x, the samples were discarded from the further analysis and eventually resequenced. We also used verifyBamID[8] to verify the correct identity of each sample, avoiding sample swaps. The software evaluates the IBD between the sequenced sample, before variant calling, and the same sample genotyped with GWAS arrays.

3. Variant calling

The variant calling identifies the variants present in our dataset after comparing the sequences to the reference. The whole process requires many complex steps: the generation of base likelihoods at each position, the identification of variant sites, the filtration of technical artifacts and finally the refinement of genotypes by using linkage disequilibrium information of neighbor SNPs. To execute such a complex pipeline, we used and collaborated in the development of the UMAKE[9] pipeline. The pipeline allows to automating all the steps, reducing the user time requested, by using Makefiles.

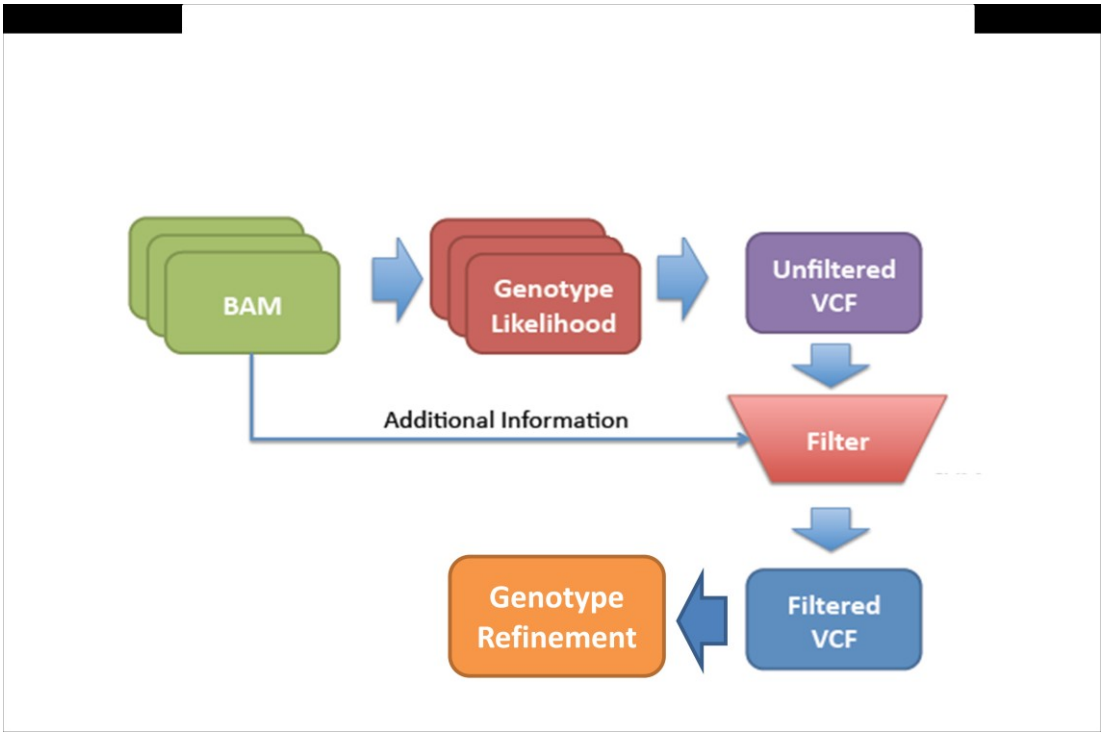

Figure 4: Workflow of the UMAKE pipeline

3.1. Likelihoods generation

This preliminary step generates 10 likelihoods (1 for each possible genotype AA, AC, AG, AT, CA….TC, TG, TT) at each position of the genome by extracting the base qualities of each nucleotide from the reads covering the position. Before generating the likelihoods, two corrections are applied to the aligned reads: Base Adjustment Quality (BAQ) and clipOverlap. The BAQ reduces artifacts due to indels by downgrading qualities at the end of the reads[10]. The clipOverlap reduces the bias generated by PCR errors generated by short reads with overlapping ends[11]. In this step each individual is analyzed separately and likelihoods are stored in GLF files for further analyses.

3.2. Variant caller

The variant caller scans the genome to identify a preliminary list of sites with at least one genotype different from the reference. In this step, we used a population-based approach that, for each position, analyzes all the individuals together. Since majority of the variants are shared among samples, the approach increases the power to detect variants, especially at positions with low coverage, and reduces false positives. The algorithm, after evaluating the likelihoods of all the individuals at each position, generates the probability of the sites being a variant. This probability is weighted with prior probabilities, derived from expected mutation rate in population genetics, to evaluate the final posterior probability of the site being a variant. If the posterior probability exceeds a specified threshold, the position is tagged as variant and genotypes are assigned to individuals who have coverage at the position. Genotypes, at this point, have low accuracy due to low coverage of the sequencing which, occasionally, may sequence only one of the two chromosomes and generate wrong genotypes.

3.3. Variant filtration

The list of variant sites generated after the variant calling contains a number of technical artifacts not detectable by an analysis based only on genotype likelihoods. For this reason, we perform further analysis on nucleotide distributions, base and read qualities to generate a better set of variant sites. As a measure of the quality of the variant sites, we evaluate the Ts/Tv ratio between the number of transition (Ts) mutations (A->G, G->A, C->T, T->C) and

transversion (Tv) mutations (the 12 remaining ones). In the DNA the expected value of the ratio is higher than 2, since transitions are more common than transversion. Instead, in case of random mutations, generated for instance by technical artifacts, the value is 0.5, thus the Ts/Tv value can be used as a good indicator of filtering quality. In the following table, it is shown the Ts/Tv for different criteria of filtering, which have, as expected, Ts/Tv lower than 2 and close to 0.5.

Filter	# SNPs	%dbSNP	Ts/tv
AB70	9302	9.7	0.93
AOI5	374	15.8	0.91
INDEL5	5010	33.6	1.16
STR20	8698	7.9	0.61
STZ5	16460	16.5	0.86
dp1160	4165	24	1.71
str-20	8846	8.4	0.63
stz-5	16975	16.5	0.89
Before filtering	422326	38.1	2.16
Filtered	41545	18.8	0.97
After filtering	380781	40.2	2.38

Table 1: Ts/Tv for different filters on chromosome 20 and variants selected before and after filtering

3.4. Genotype refinement

Genotypes generated by the variant caller, as mentioned previously, have high error rate due to low coverage sequencing of our samples. However genotype accuracy can be improved by taking into account linkage disequilibrium (LD) between SNPs. In fact, even unrelated individuals share short stretches of chromosomes which may be used to infer missing genotypes or to assign better

genotypes in positions poorly covered.

The algorithm uses a Hidden Markov Model approach, reconstructing haplotypes and LD blocks. In each iteration, the algorithm reconstructs each individual haplotypes configuration on the base of haplotypes found in the remaining individuals. Genotypes are continuously updated and used as reference until the algorithm completes the selected number of iterations. The result is a complete genotyping of the individuals at the positions detected by the variant caller and, more importantly, a complete reconstruction of the phased haplotypes. The phased haplotypes in fact, can be used as reference panel for imputation, as we will explain in chapter 4.

4. Sardinia Sequencing

4.1. Study design: motivations

The goal of our project is the investigation of Sardinian specific variation and, more in general, of the Sardinian variation not accessible by direct GWAS genotyping or after imputation on 1000G or HapMap.

The best method to assess the complete genomic variability present in our

dataset is the deep sequencing of all the individuals available to us. However the project would be extremely costly. Therefore we decided to choose a cheaper approach, by low pass sequencing a subset of 2120 individuals and to genotype *in silico* the remaining part of the cohort. In this chapter we will discuss the motivations behind our choices.

4.1.1. Why a sequencing project?

First of all, we decided to choose a sequencing project in opposition to a more extensive classical GWAS genotyping plus imputation, previously used in the past in our studies. The GWAS approach worked extremely well to detect common variation but also a fraction of rare variation shared between Sardinian and the public reference panel populations (typically Europeans). However, despite the large number of loci found associated with many different phenotypes, the variation detected by such approaches cannot explain the complete heritability of phenotypes. Some recent findings in our SardiNIA cohort, based on fine-mapping based arrays, like Metabochip, showed that a fraction of the missing heritability is due to rare variants not directly genotyped or tagged by proxies using previous GWAS platforms[12]. Furthermore some of the variability can be due to complex variants, like indels, CNV which would require additional genotyping to be analyzed. With a sequencing approach, we will be able to investigate the complete spectrum of variation, especially the Sardinian specific variation, and the complex variants, without additional genotyping. Furthermore, since we will generate a reference panel, we will be able to impute these variants in future studies involving Sardinian individuals.

4.1.2. Why a sequencing project in Sardinia?

As mentioned previously, GWAS approaches have limitations in detecting the population specific variability. The variation not accessible by such approaches is proportional to the genetic distance between the population under analysis (i.e. Sardinians) and the population used as reference (i.e. Europeans). The Sardinian population has a large genetic distance (Fst) to other geographically neighbor populations. Published results[13] from the Human Genome Diversity Panel(HGDP) reports an Fst=$5.3x10^{-3}$ with Italians and Fst=$5.7x10^{-3}$ with North Italians, showing a pattern similar to other isolated populations, like Basque. Furthermore using PCA analysis Sardinian individuals have been found to be outliers in a genetic map of Europe[14].

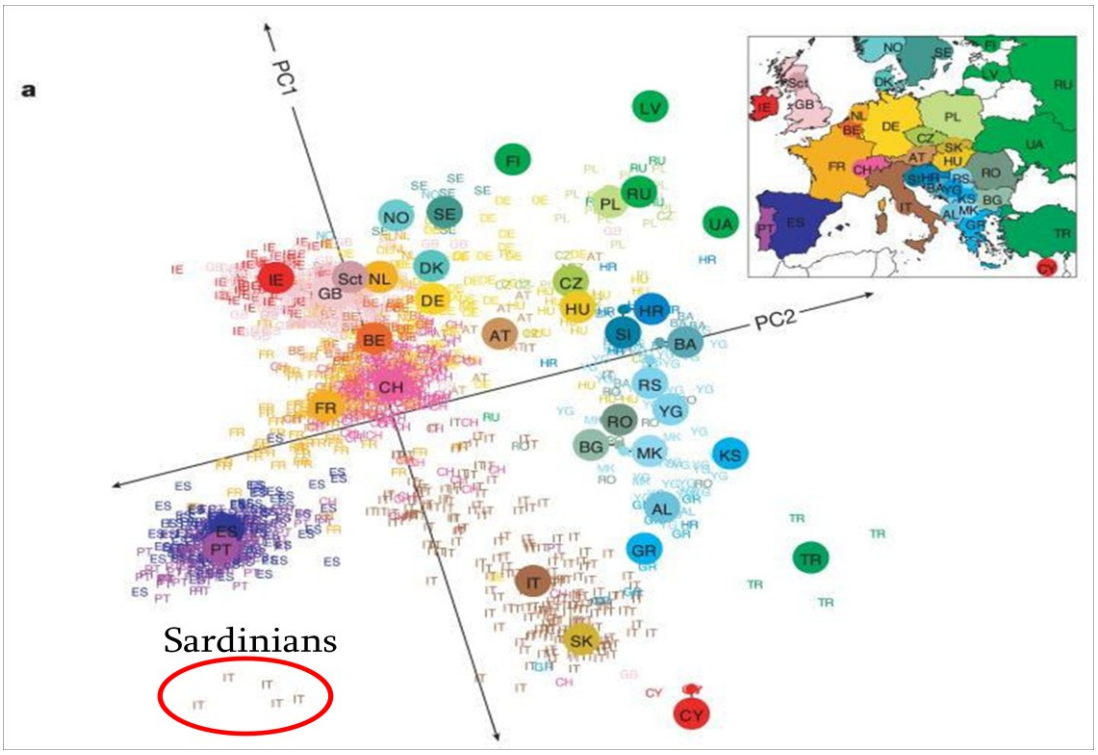

Figure 5: The PCA genetic map matches the geographic map of Europe. Afterwards Italian outliers were assigned to Sardinian population, (figure by Novembre et al, 2008)

For these reasons we expect that, compared to other non isolated populations, we will have a larger fraction of variation not accessible using GWAS arrays or reference panel based on European populations.

4.1.3. Why whole genome low pass sequencing?

Next generation sequencing approaches offer many different study designs and possible alternatives on the number of individuals to sequence, on the genomic regions and on the coverage(the number of reads for each position). In this paragraph we will discuss the possible alternatives and the reasons behind our choices.

The best method to investigate the complete variability of our sample would be whole genome deep sequencing (30x or more) of all individuals present in our dataset. This approach would give us the possibility to investigate the whole genome with very good genotyping accuracy and a good assessment of rare variants. However its costs are extremely large and, nowadays unfeasible, thus we will need to reduce the amount of sequences generated by reducing the number of individuals sequenced or by reducing the sequencing coverage.

Published results[15], based on simulations on 45000 chromosomes, show a comparison between two different approaches using the same amount of sequencing (12000x), sequencing of 3000 individuals at 4x or sequencing of 400 individuals at 30x.

Population Maf (%)		0.1–0.2	0.2–0.5	0.5–1	1–2	2–5	>5
Statistic	**Design**						
% Discovery	**400@30x**	65.4%	87.1%	100.0%	100.0%	100.0%	100.0%
	3000@4x	58.1%	94.3%	100.0%	100.0%	100.0%	100.0%
Overall genotypic concordance	**400@30x**	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	3000@4x	99.8%	99.7%	99.6%	99.7%	99.6%	99.8%
Heterozygote concordance	**400@30x**	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	3000@4x	82.4%	81.9%	90.3%	97.2%	98.8%	99.8%
Dosage r2	**400@30x**	99.4%	99.6%	99.7%	99.8%	99.8%	99.9%
	3000@4x	63.9%	68.9%	80.2%	91.9%	95.7%	99.2%
Information content (n * r2)	**400@30x**	398	398	399	399	400	400
	3000@4x	1917	2069	2406	2758	2873	2978

Table 2: Comparison of high-coverage (400 @ 30X) and low-coverage (3000 @ 4x) sequencing designs, given the same total sequencing effort. Simulated data.

Looking at discovery of variants present in the population, we can see that for variants with frequency higher than 0.5 % both approaches behave in the same way discovering 100% of variability. At lower frequencies, as expected, both approaches miss a fraction of the variation in the 45000 chromosomes, due to the low number of individuals sequenced in the case of 400 individuals at 30x coverage and to the low coverage in case of the design with 3000 individuals at 4x.

The design with 400 individuals at 30x coverage shows much better concordance with genotypes and better accuracy after imputation (dosage r2). Instead, looking at the effective sample size (information content), the design with 3000 individuals, even with the 66% reduction at lowest frequencies, outperforms the design with 400 individuals, whose sample size cannot exceed the fixed limit of 400.

However, since our goal is the detection of association signals in the

population, a better parameter of interest is the statistical power to detect population signals using different depth of sequencing plus imputation in the remaining individuals not sequenced in the cohort.

	Population Minor Allele Frequency			
Design	**0.1%**	**0.5%**	**1%**	**3%**
400@30x	4.00%	6.40%	7.40%	11.60%
400@30x + imputation in remaining 2600	10.40%	14.20%	15.00%	34.60%
1000@12x	12.20%	13.40%	14.60%	17.80%
1000@12x + imputation in remaining 2000	15.60%	20.80%	25.60%	41.80%
2000@6x	54.40%	57.80%	61.60%	82.20%
2000@6x + imputation in remaining 1000	56.20%	59.40%	61.80%	83.60%
3000@4x	61.60%	75.60%	82.80%	90.40%

Table 3 : Comparison of power to detect disease-SNP association

The power of detecting signals is much higher in the design with 3000 individuals sequenced at 4x because of the large number of individuals sequenced, which contain a larger number of variants (especially rare). This allows to including them in the reference panel and more importantly to have multiple copies of rare haplotypes which may be imputed with better accuracy.

4.2. Imputation using reference panels from sequencing

GWAS studies require large sample size to detect low effect association signals, however sequencing costs are currently too large to make possible to reach typical GWAS sample size of sequenced individuals. For this reason we

decided to sequence only a subset of the samples in our cohorts and to genotype *in silico*, trough imputation, the remaining part of the cohort, using the reference panel generated from sequence.

4.2.1. A population based reference panel

The variant calling pipeline described in chapter 3 generates a dataset containing a phased haplotypes for each sequenced individual at the positions filtered after variant calling. We used these phased haplotypes to build a reference panel containing Sardinian individuals and haplotypes.

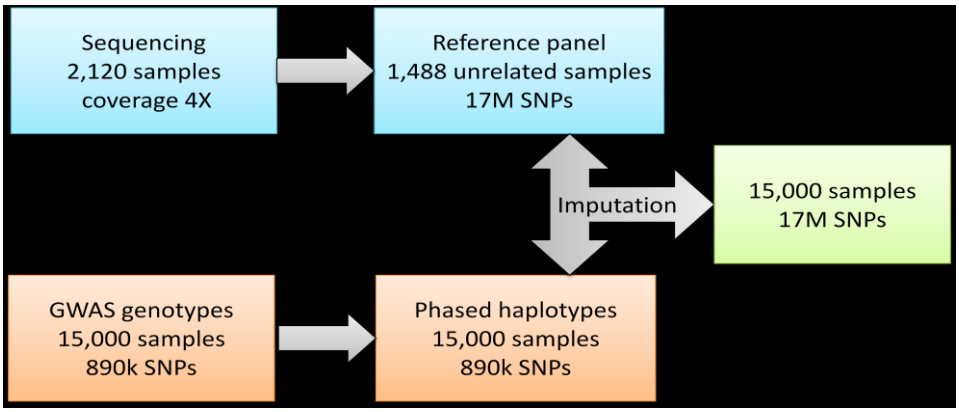

Figure 6: Workflow of imputation of sequence into GWAS genotypes

Starting with the 2120 individuals sequenced at 4x coverage, we removed the related samples (mostly child of a trios) and we generated a reference panel containing 2976 (1488 individuals x 2) phased haplotypes.

In parallel with the sequencing project we genotyped, using GWAS commercial arrays, a total of 15,000 samples, from the 2 projects described in chapter 1, and we phased them using MACH[16] .

We then combined the two datasets using imputation, genotyping *in silico* the 15,000 samples from the two projects at the 17 million variants generated by sequencing.

4.2.2. Comparison with 1000G reference panel

To better understand the benefits of a population based reference panel, we performed a comparison between our panel and the most complete and large reference panel publicly available: the 1000G reference panel. The panel contains 1092 individuals (sequenced at 4x), from 4 different macro areas in the world: Europe, Africa, Asia and America. Of these, 381 individuals belong to the European population, see chapter 5.3.1 for more details on other populations.

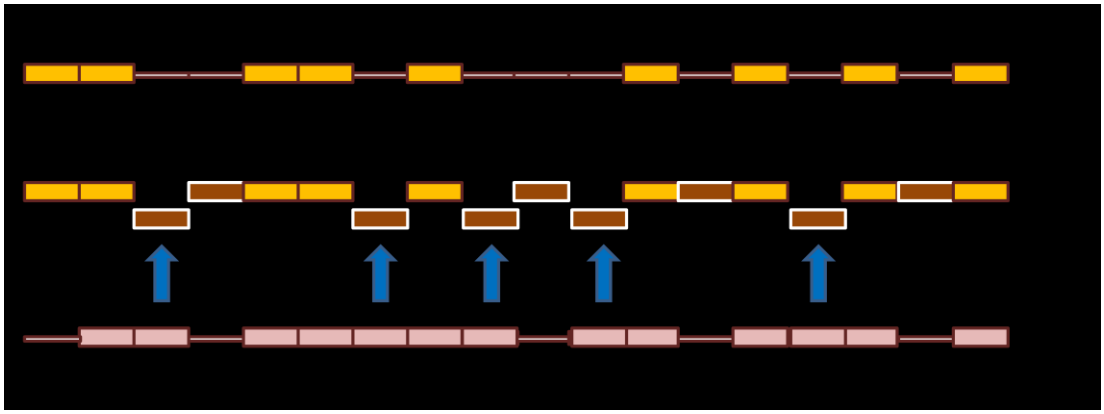

Figure 7: Schematic of reference panels comparison, genotypes indicated by the blue arrows are used to evaluate imputation accuracy

We compared the performances of the two reference panels by using two GWAS arrays in the Sardinia cohort. The genotypes from the first array (scaffold array) have been phased and used for imputation; the genotypes from the second array (test array) have been used as golden standard to compare the accuracy of

imputation. We removed the SNPs directly genotyped in the scaffold array to test only the SNPs genotyped *in silico* (indicated by blue arrows in the figure 7) and have a better estimation of the true imputation accuracy. We also excluded from the comparison the Sardinian individuals in common between the reference panel and the scaffold array.

Reference panel (sample size)	Imputation accuracy (r^2)			
	MAF 0.5-1%	MAF 1-3%	MAF 3-5%	MAF >5%
1000G Europeans (381)	56.2%	78.0%	86.7%	92.0%
Sardinians (347)	70.5%	87.6%	92.1%	95.6%
1000G ALL (1092)	59.2%	79.6%	87.1%	92.2%
Sardinians (831)	79.0%	92.0%	95.5%	97.1%
Sardinians (1488)	87.3%	94.8%	97.1%	98.2%

Table 4: Comparison of the Sardinian and 1000G reference panels

Comparing the first two reference panels with the same number of individuals (347 Sardinians versus 381 Europeans), it is possible to see that for common variation (MAF > 5%) both panels have good performances, with a correlation higher than 90%. When moving to lower frequencies, however, the Sardinian reference panel outperforms the 1000G reference panel, because the Sardinian reference panel contains the rare and population specific variation (and haplotypes) not present in the 1000G European individuals. Even including the whole set of 1092 individuals from 1000G it is possible to see the same trend. Notably, the inclusion of the individuals from Africa, Asia and America does not improve the imputation of common variation, meaning that 381 Europeans are enough to impute with good accuracy the Sardinian individuals at common SNPs.

The small improvement in the lower frequencies is probably due to rare haplotypes shared between Africans and Sardinian not present in the Europeans.

By increasing the sample size of the Sardinian reference panels, we can see a large increase in the different interval of frequency. Strikingly, the imputation accuracy of rare variants reaches 87.3% when using the latest panel with 1488 panel.

5. Sequencing results

In this chapter we will describe the results of sequencing the 2120 individuals, the number of variants detected, their accuracy and their quality. We will show some comparisons with the 1000G project variants, to understand how much of the Sardinia variability can be identified by other sequencing projects containing European individuals. Finally, we will show an example of application to the SardiNIA project to better understand how the next generation sequencing approaches are essential to detect some Sardinian specific loci.

5.1. Data freezes, effects of increasing number of individuals sequenced

In opposition to GWAS studies, generation of sequence data takes a very long time, typically one week per 10 samples. In our case, it took 2 years to complete sequencing of the 2120 samples. For this reason, we decided to periodically generate data freezes with the available samples to evaluate the quality of the data and to start testing the tools and methods to apply to future analyses.

5.1.1. Variants discovered increasing number of individuals

The project started in March 2010 and the first dataset, in June 2010, includes 66 individuals. Since then, we performed a total of 6 data freezes with increasing sample sizes that allowed us to continuously improving our methods of analysis and to testing how increasing sample size affects our population based variant calling and genotyping refinement.

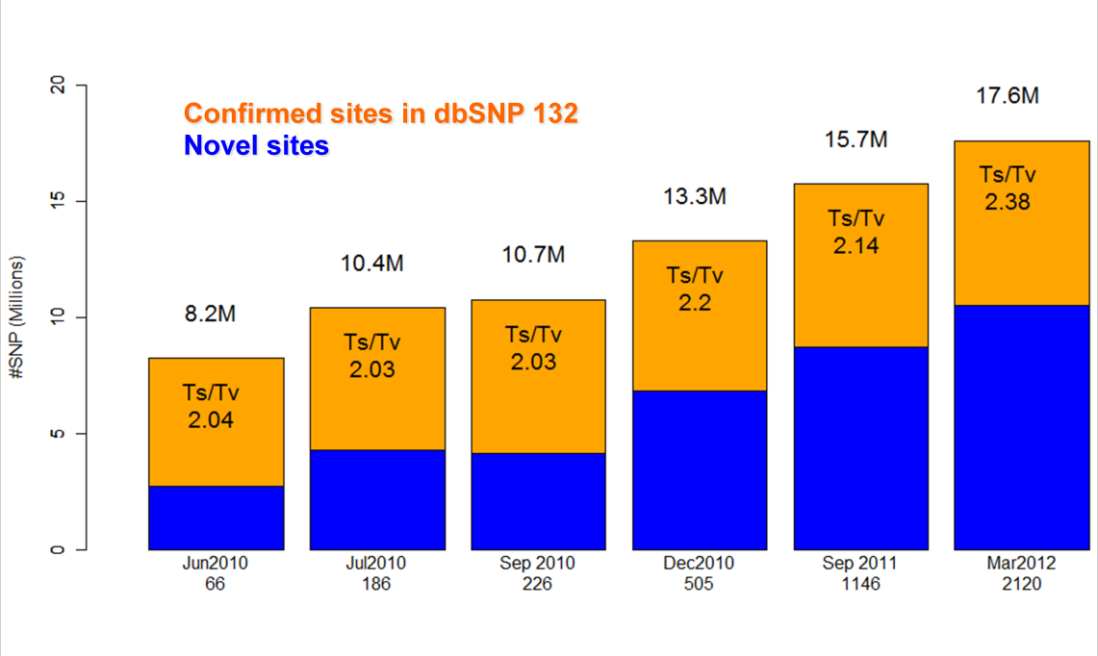

Figure 8: Data freeze timeline, with date and number of individuals for each data freeze

An increasing in the number of individuals corresponds, as expected, to a large number of variants detected and more importantly to a large number of novel variants, not present in common database and then not accessible using GWAS approaches. These are the variants we are most interested in, since some of them may represent Sardinian specific variability that we would probably never access without using a sequencing approach on Sardinian individuals.

As mentioned before, we are using the Ts/Tv ratio as a measure of quality of our dataset. Its value is quite stable, confirming the quality of the variant sites in every data freeze. It is possible to see a small improvement in the latest data freezes, due to the improvements in analysis methods that we developed during the two years of our work that leaded to more sophisticated and accurate variant calling, filtering and genotyping refinement algorithms.

5.1.2.Accuracy increasing number of individuals

Since we are using a population approach, we expect that having more individuals will result in better variant calling and genotyping. To understand the size of this effect, we evaluated, for each data freeze, the heterozygous genotype concordance between the genotype generated by the sequencing approach after the genotype refinement and the Metabochip (a fine mapping based arrays that contains variants in the whole frequency spectrum).

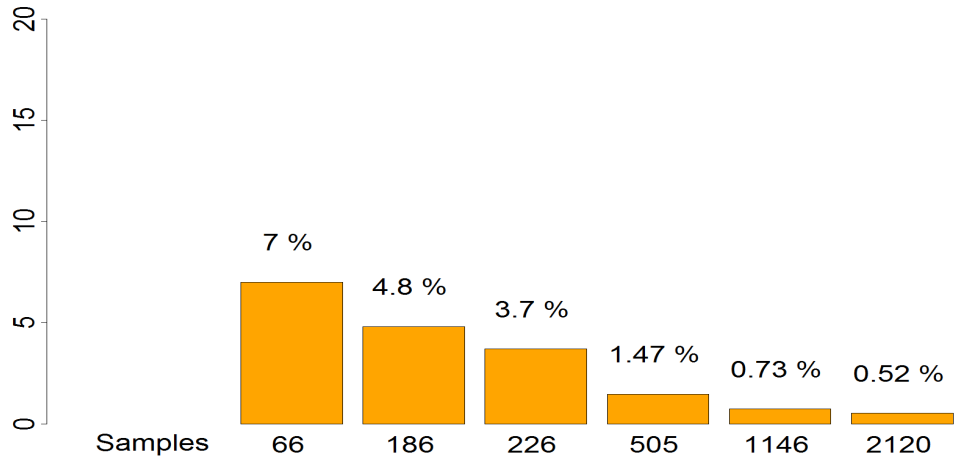

Figure 9: Mismatch rate between Metabochip GWAS genotypes and sequence genotypes after genotype refinement as a function of the number of individuals

Results show a clear improvement in the genotype accuracy of sequence data, due to the increase of the number of individuals sequenced. By calling together more individuals we have two major benefits. The first one is that it is easier to call a variant if it is carried by more than one individual. When calling a variant with frequency 1%, it appears as a singleton in 100 individuals, instead it appears 10 times when calling 1000 individuals together. This effect gives larger

advantage to rare variants, which shows a better improvement in the accuracy. The second benefit affects the genotype refinement algorithm, by increasing the number of individuals we have a larger pool of haplotypes to assign to each individual, allowing a better reconstruction of the haplotipic configuration and then a better genotype and phasing accuracy.

5.1.3. Saturation

In the paragraph 5.1.1 we described how, by increasing the number of individuals, we obtain a larger number of variants. In this paragraph we will discuss the limits of number of variants found in the population and whether we can have an indication on the number of variants we may find by sequencing more and more individuals.

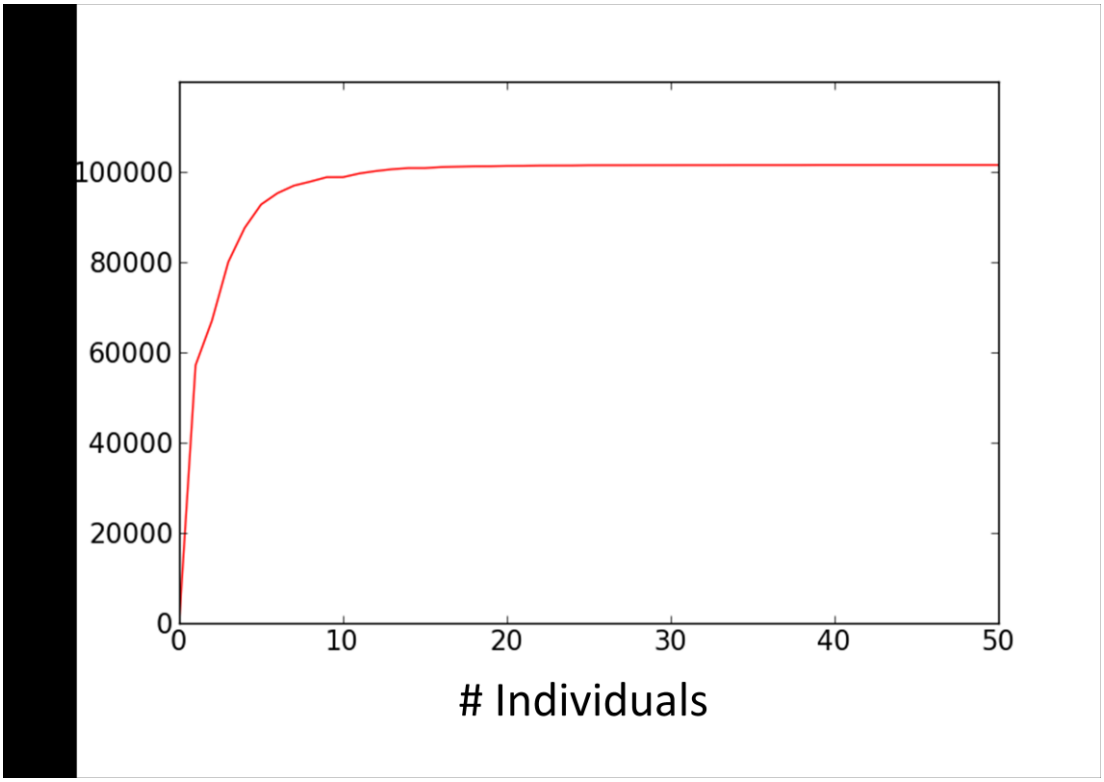

Figure 10: Number of common variants (MAF > 10%) detected as a function of the number of individuals

In the figure above, we are plotting the number of common variants (MAF> 10%) detected as a function of the number of individuals. In this case, by sequencing one individual we find almost 60,000 variants (~60% of the total). By adding more individuals, up to 15, it is still possible to see an increment on the number of variants. After 20 individuals, however, the number of variants detected stops increasing since these 20 individuals already contains all the common variants in our dataset.

Looking at different frequency bins, the pattern is similar but there are some interesting considerations about.

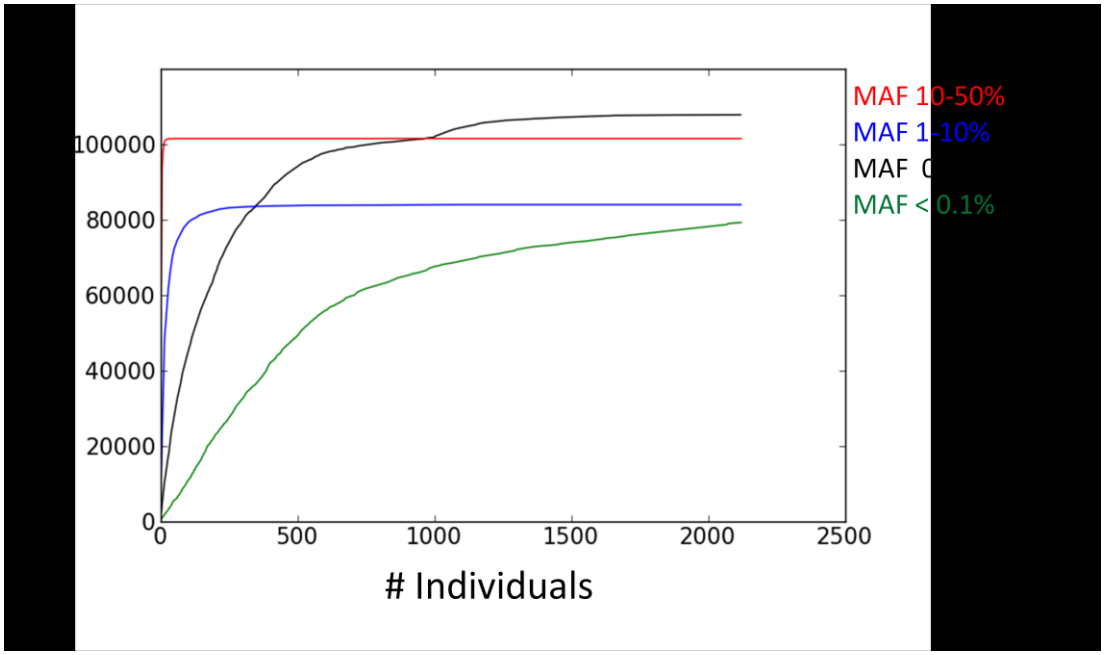

Figure 11: Number of variants detected as a function of the number of individuals and of the minor allele frequency

For MAF 1-10% (blue line) we find the same trend of the common variants, the number variants increase up to ~250 individuals then we reach the saturation since we detect all the variants in the dataset

Before analyzing the frequency bin with MAF 0.1-1%, it is important to highlight individual's distribution in the plot. The first 997 individuals are from the autoimmunity case control cohort from Cagliari and Sassari; the remaining 1123 individuals are from the SardiNIA cohort in Lanusei, in the most isolated part of Sardinia. Looking at the black line (MAF 0.1-1%) there is a step at 1000 individuals due to the introduction of the individuals from the SardiNIA cohort which belong to a different sub population with a different set of rare variants. Since the small step appears only in frequencies lower than 1%, we can conclude that there are not substantial difference in the variants with MAF >1% found in the two cohorts (even though there are some differences in the allelic frequencies detectable by PCA analysis). Furthermore for frequencies below 1% we cannot

infer anything about the total number of variants present in Sardinia, since adding other Sardinian cohorts to our dataset may increase the number of variants in the frequency bin 0.1%-1%.

Finally, by analyzing the frequency bin with MAF < 0.1%, we cannot find saturation on the number of variants, as expected, since this frequency bin include also the singletons and potentially every new individual added to the dataset carries a moderate amount of singletons.

To conclude, even if we reach saturation of the number of variants for some frequencies bins, sequencing more individuals helps in the detection of rare variants and more importantly helps in genotyping with better accuracy the rare variation currently detected, which results in higher power to detect association's signals generated by rare variants.

5.2. Statistics on variants

In this chapter we will focus our attention on the results of sequencing 2120 individuals and on the number of variants detected.

Before filtering, we detect a total of 19.4M variants with a Ts/Tv of 2.02, after applying the filters the number of variants reduced to 17.6 with a higher Ts/Tv of 2.19, confirming the validity of the filters.

FILTER	#SNPs	#dbSNP 135	%Novel	Known Ts/Tv	Novel Ts/Tv	Overall Ts/Tv
PASS	17,617,122	12,220,588	30.6	2.22	2.12	2.19
FAIL	1,790,250	598,918	66.5	1.49	0.8	0.98
TOTAL	19,407,372	12,819,506	33.9	2.18	1.76	2.02

Table 5: Number of variants, percentage of known SNPs and Ts/Tv ratios for the dataset with 2120 individuals

To better evaluate the quality of our data we are evaluating the Ts/Tv ratio separately for known and novel variants. Known variants are detected by direct genotyping and confirmed also by other groups, so we expect that they have lower rate of false positives. Since, at genomic level, there should not be any difference between known and novel variant, we expect that the two categories have the same Ts/Tv ratio. Looking at the variants before filtering, we notice that there is a small imbalance (2.18 vs. 1.76), however, after filtering the imbalance disappears and the Ts/Tv ratio become more similar (2.22 vs. 2.12).

The percentage of novel SNPs is 30.6%, however it is important to stratify the result by different minor allele frequencies.

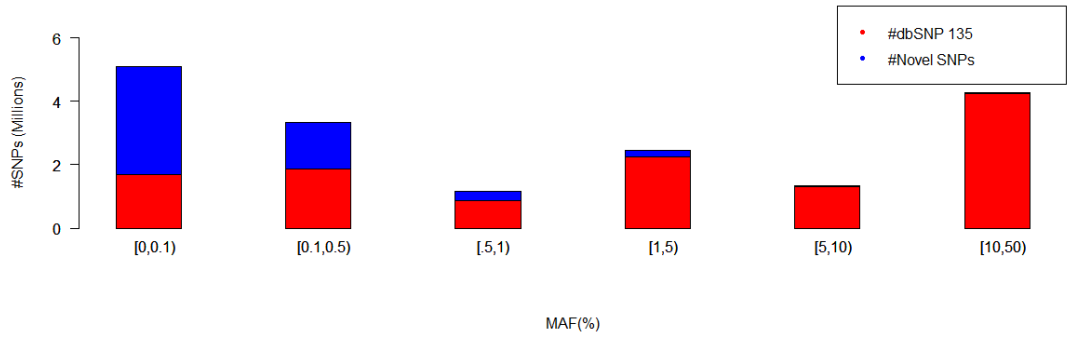

Figure 12: Distribution of known and novel SNPs among different minor allele frequencies

Looking at the Figure 12, the common variants show a very small fraction of novel variants (0.4%). This is quite expected since they are shared worldwide, can be detected by other studies and then are already included in dbSNP 135.

We detect a very large number of rare variants, ~30% of the variants has frequency lower than 1/1000 and ~50% of the variants has frequency lower than 5/1000, under the expectation of population genetics theory. These two frequency bins, have a larger number of novel variants (67% and 44% respectively) for a

total of almost 5 millions of novel variants. Since they are so rare in Sardinia, they can be even rarer outside Sardinia, and therefore they probably will never be present in the individuals involved in public datasets. By using our sequencing approach we are able to genotype them trough sequencing and, more importantly, we are able to genotype them *in silico* in any Sardinian individuals genotyped with a GWAS array.

Other than evaluating the raw number of individuals we evaluated the number of variants according to their functional annotation.

	Genome	Synonymous	Nonsynonymous	Nonsense
Total SNPs	17.6M	57,716	74,037	1,387
Novelty rate dbSNP135	30.6%	23.0%	35.0%	55.7%

Table 6: #Variants according functional annotation

As already mentioned, we found a total of 17.6 millions variants and 31.6% of these are already in dbSNP 135.

We also found 57,716 synonymous variants (23% novel) 74,037 non synonymous variants (35% novel) and interestingly 1387 nonsense variants (56% novel). The higher rate of novel non-synonymous and nonsense variants can be explained by their frequency. In fact since they are altering or damaging, they have a lower frequency compared to synonymous and intronic variants whose mutation does not modify any phenotypes.

Per individual statistics				
	Genome	Synonymous	Nonsynonymous	Nonsense
5th percentile	3,376,149	10,359	9,006	57
Mean	3,403,628	10,512	9,168	66
95th percentile	3,427,757	10,664	9,322	76

Table 7: Number of variants per individual

Looking at the results at individual level, we find a mean of 3.4 millions variants per individual, with very small deviation in the whole dataset. In average we also find 10,512 synonymous and 9,168 non synonymous variants, in addition to 66 nonsense variants. At a first glance, the number of nonsense variants per individuals looks too high, but many of the non sense variants may belong to genes with no essential functionality. Further studies are needed to evaluate the impact of these stop-codon variants.

5.3. Sardinia vs. 1000 Genomes

In this chapter we will compare the dataset we generated with the largest publicly available dataset generated with next generation sequencing.

5.3.1. The 1000 Genomes project

The 1000G project is an international cooperation, involving the most important universities and the research centers, with the aim of describing the

worldwide common and rare variation by sequencing more than 1000 individuals from different parts of the world. The project also compares different strategies of sequencing by doing exome sequencing and deep sequencing of limited number of individuals. For our purposes however we will focus only on the whole genome low coverage sequencing.

The currently available data contain 1092 individuals from 4 continents (379 Europeans, 286 Asians, 246 Africans and 181 Central Americans), each of them sequenced at a coverage of ~4x. Since we used the same sequencing technology and we performed similar steps of analysis, we can compare the variants present in the two dataset without significant adjustment to the results to compensate different methods of variant calling.

5.3.2.Population comparison

Since 1000G project includes different populations, we decided to compare their variants with the variants detected in our dataset. For the comparison, in Figure 13 we are plotting on the Y axis the fraction of Sardinian variants rediscovered by the 1000G population of interest and on the X axis the number of 1000G individuals, using an approach similar to the one we used in the saturation.

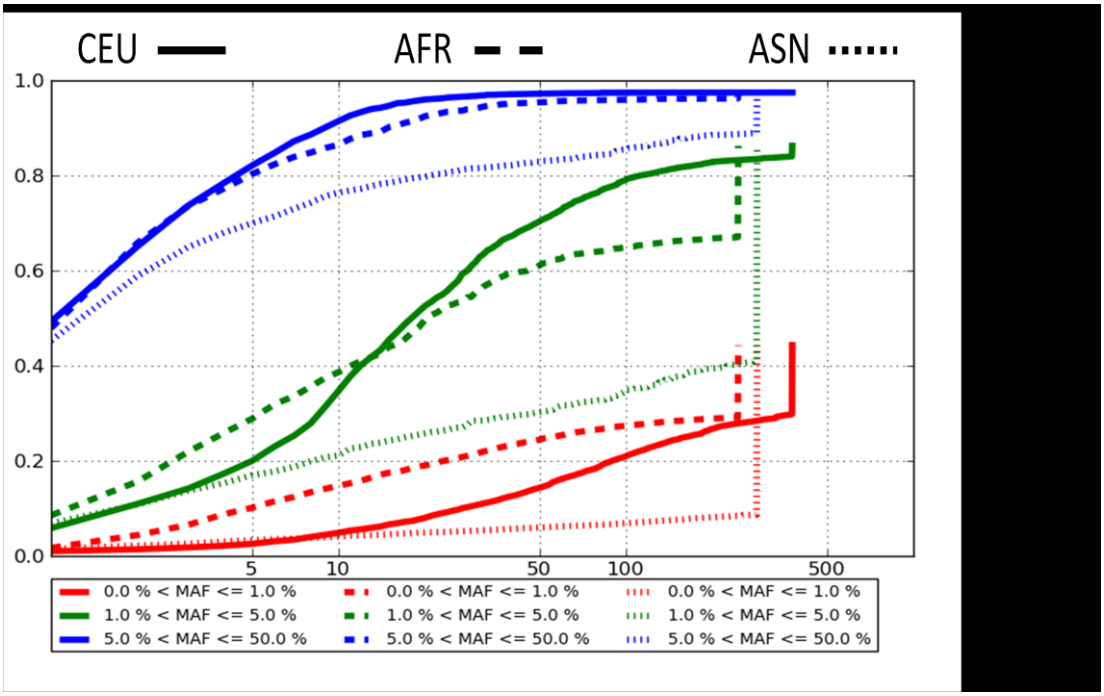

Figure 13: Fraction of Sardinian variants rediscovered by the 3 major 1000G populations (Europeans, Africans, and Asians). Different lines represent different populations; different colors represent different frequency intervals. Height of the last point represents the variation discovered by the remaining individuals not included in the population of interest

From geographic considerations, we expect a larger sharing between Sardinians and Europeans, followed by Africans and finally Asians which should show lower sharing for all the frequency intervals. As expected, genetic results match geographic expectations; however some more considerations have to be done.

For common frequencies (MAF > 5%) Europeans and Africans share with Sardinians approximately the same amount of variation, with a slightly higher value for Europeans. In both cases however, about 50 individuals are enough to discover 95% of the common variation present in Sardinia. Part of the missing variation may be due to high frequency Sardinian specific variants, to variants filtered out by quality controls in the 1000G or to some unfiltered artifacts in our

dataset. Further investigation of these variants has to be done.

Frequencies between 1 and 5% show an interesting difference between Africans and Europeans, in fact 10 Europeans individuals share a smaller number of variants when comparing to 10 Africans individuals since they belong to an older population and carry a large number of rare variants. Increasing the number of individuals however, the sharing between Sardinian and Europeans increases arriving up to 84% compared to a 68 % of the Africans individuals. The 16% difference may be due to rare variants recently mutated and spread in Europe and Sardinia but not present in Africa.

Finally, looking at very rare variants with frequency <1%, the sharing between Sardinians and Europeans/Africans is extremely low, around 27%, because the majority of variants in this bin are very rare variants originated in Sardinia and not spread in the Europe/Africa. It is also interesting to see how the variant detected by Europeans and Africans represent almost a disjoint set, and even if they separately reach 27% , the total number of variants detected by the union of the two populations reaches 54% of the very rare variation present in the Sardinian population.

To remark the consideration on paragraph 4.2.2, about imputation quality, the 1000G reference panel is a good dataset to investigate common variation present in Sardinia. However, since our goal is the investigation of the rare and Sardinian specific variation, the sequencing and the construction of an internal reference panel is an essential requirement for our goal.

5.4. A proof of concept: the LDL cholesterol in the Sardinia cohort

So far we described the improvement and the benefits of the sequencing project to the variants discovery and to the imputation accuracy and discussed how these benefits may lead to higher power to detect Sardinian specific association's signals. In this chapter we will show an example of a real case in which the Sardinia sequencing allows us to detect a variant not accessible using GWAS arrays or 1000G imputation.

5.4.1. GWAS genotyping versus imputation after sequencing

Here, we will focus on the association signals detected using as outcome the LDL cholesterol in the Sardinia cohort.

The study of this trait involves 5949 phenotyped individuals from the Sardinia cohort. We tested the association using linear regression model adjusted using standard covariates, age ,squared age, sex. Since the individuals belong to families, we tested them using the Merlin[18] software, able to test association taking into account pedigree kinship to weight associations found in related individuals.

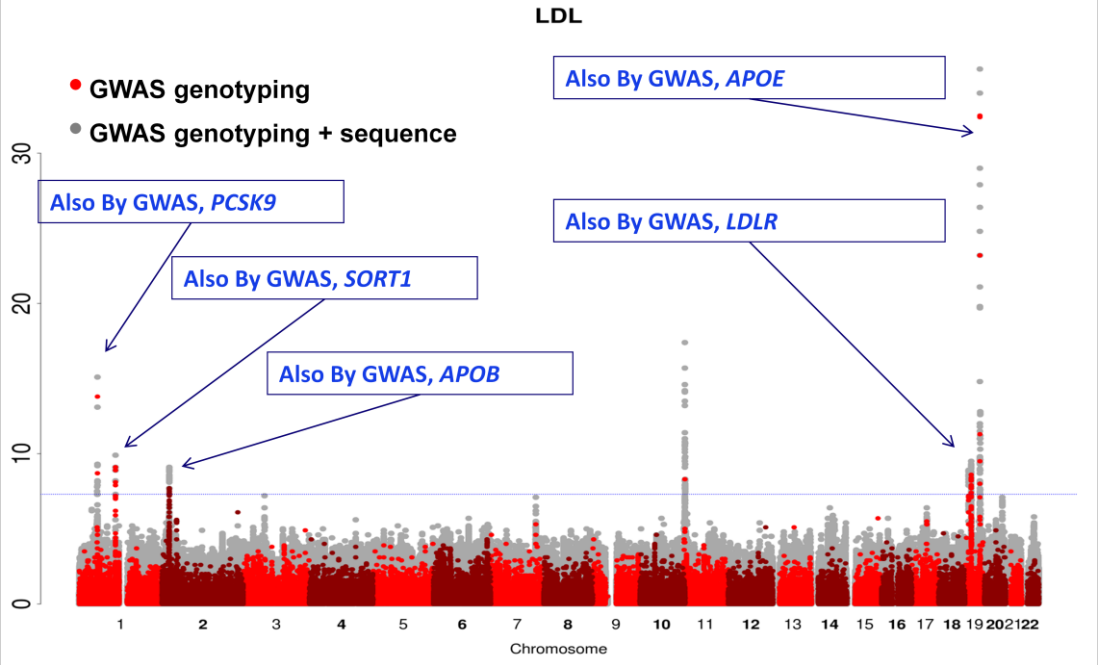

Figure 14: Manhattan plot of LDL cholesterol in the Sardinia cohort. Red points represent associations using GWAS arrays only, gray point the background represent association after imputation with the Sardinian reference panel

In the Figure 14, we are showing the Manhattan plot of LDL associations, the scale on Y axis is the negative $\log_{10}$ of the association's p-values and the horizontal blue line represents the significance threshold for GWAS ($5x10^{-8}$). On the X axis there is the position in the genome, separated by chromosome. The red points are generated testing the association using GWAS arrays on a total on 831,515 SNPs coming from the merging of Omni Express, Metabochip, Immunochip and Exomechip. The gray points in the background are generated testing the genotypes after the imputation on the Sardinia reference panel with 1488 individuals, a total of 17.6 millions SNPs.

Looking at the loci from previous GWAS studies (genes *PCSK9, SORT1* on chromosome 1, *APOB* on chr2, *APOE* and *LDLR* on chr19), we can notice how both approaches are able to detect the signals, since these signals come from

common variants present in the genotyping arrays. These variants have been found in the European populations and then they are present in the GWAS arrays. Their association p-value is similar to the p-value after imputation meaning that they are good proxies of the signal found in the gene.

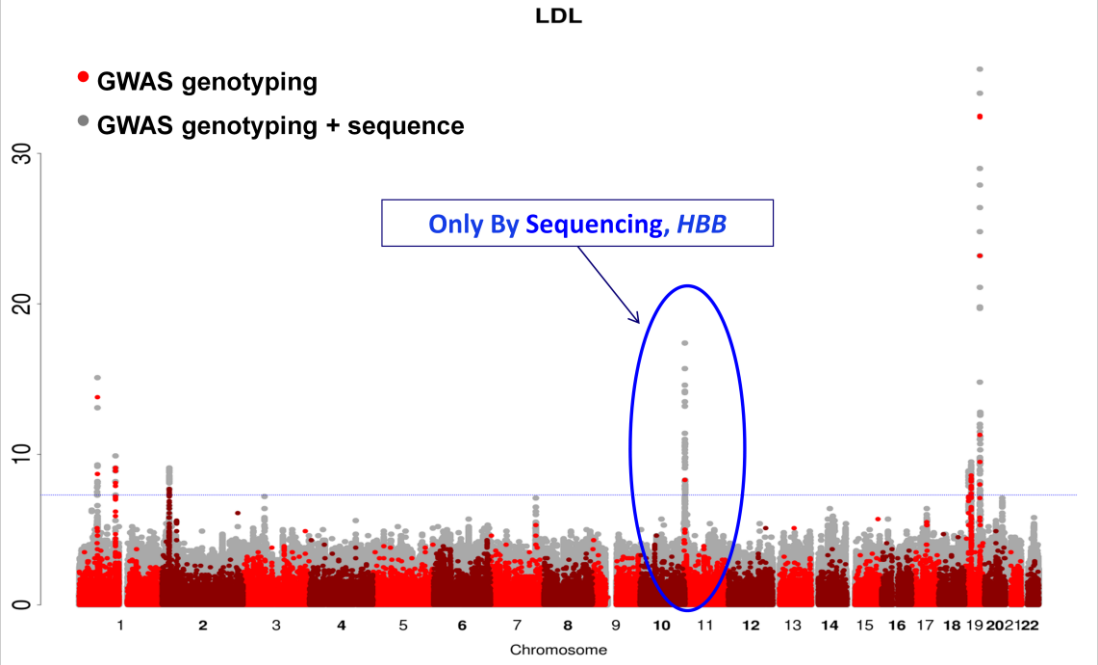

Figure 15: Manhattan plot of LDL cholesterol in the Sardinia cohort. The association on chr11 shows higher association when using imputation

However, focusing the attention on chromosome 11, the situation is quite different. Here, there is one GWAS SNP associated with a p-value $3.7x10^{-10}$, while the top SNP associated using sequencing is associated with a p-value $4.3x10^{-21}$ (confirmed with direct genotyping). The SNP in the GWAS arrays, in this case, is a bad proxy of the top SNP found with sequencing and its association value does not reflect the true association of the locus. Furthermore the association detected with GWAS is driven by a single SNP, missing the usual pattern of association driven by multiple SNPs in LD with similar p-values.

5.4.2. The HBB locus

Investigating more carefully the locus detected after imputation on Sardinia reference panel, we found that the top SNP rs76728603 (chr11:5248004) is a stop-codon mutation on the *HBB* gene. This mutation is extremely well known in Sardinia. In fact, the mutation is responsible for 95% of cases of β0–thalassemia in Sardinia[19] but is also known to be protective against malaria. Due to very high positive selection effect against the malaria the mutation has an increased frequency in Sardinia (5%) compared to its extremely low frequency in Europe (0.03%) thus it is very hard to be detected when sequencing European individuals and it is very unlikely that this variant is included in GWAS arrays.

Even if a variant associated with thalassemia and malaria may sound unrelated to cholesterol, the association is known since 1989[20]. The effect of the variant is the truncation of beta-hemoglobin chain, resulting in a reduced life span of red blood cells. To keep the right amount of red blood cell, individuals carrying the variant need an accelerated erythropoiesis which uses a larger amount of plasma lipids (compared to individuals without the variant) and then the carriers have a reduced amount of LDL cholesterol.

5.4.3. Detecting the Q40X using public datasets

We evaluated whether the variant is accessible by using different approaches by checking its frequency in public datasets commonly used for imputation.

In the HapMap project, the variant is not present. In this case the limitation comes from the genotyping technology, in fact the HapMap project is

based on Affymetrix and Illumina GWAS arrays. These are two commercial platforms designed on general worldwide populations and designed to contain tag SNPs to access the largest possible variation among individuals worldwide.

Another publicly available dataset, containing exome sequencing on 6500 individuals (2200 African Americans and 4300 European Americans) has only a copy of the alternative allele on a total of 13,000 chromosomes. Unfortunately this is an exome sequencing based study and it is not possible to use the data for imputation.

Finally we analyzed the variant in the 1092 individuals sequenced by the 1000G project. Again, the variant is just a singleton in 2184 chromosomes and unexpectedly the only carrier of the heterozygote genotype is a Mexican individual, confirming once again its low frequency in Europe. Since 1000G data generated a reference panel for imputation we decided to run the imputation to understand whether the association signal is detectable by using the 1000G external reference panel.

We evaluated the association signals around the HBB variant, after imputation on the same GWAS array with two different reference panels, the 1000G (on the left) and the Sardinian reference panel (on the right).

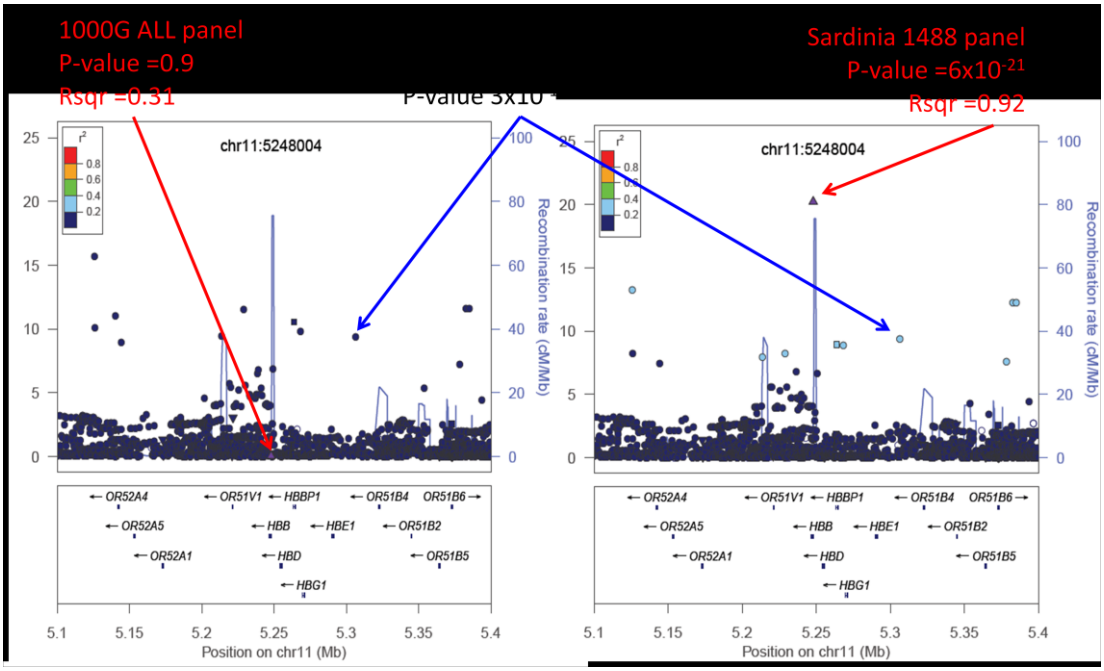

Figure 16: LocusZoom on the HBB mutation region. Left panel plots the results after 1000G imputation, right panel plots the results after Sardinian reference panel imputation. Dot colors represent the degree of linkage disequilibrium (r^2) with the HBB variant

Using the 1000G panel, the HBB Q40X variant shows an association p-value of 0.9, thus the association signal is completely missed due to the very low imputation accuracy (rsqr = 0.31). Since the 1000G reference panel does not contain the Sardinian haplotypes with the Q40X mutation and the Q40X mutation is a singleton, the imputation quality is extremely low, generating wrong genotypes and producing a p-value completely different from the real value.

When using the Sardinian reference panel, the imputation quality is much higher (rsqr=0.92) and the p-value on imputed genotypes matches the p-value obtained after direct genotyping of the variant, confirming once again the benefits of using the reference panel generated by sequencing individuals from the Sardinia population.

6. Conclusions

6.1. Analyses performed

With the goal of investigating the Sardinian specific variability we sequenced 2120 individuals using whole genome low pass sequencing approach. We performed alignment and quality checks of the sequence samples and we called the variant sites using a population based approach. We filtered variants using stringent criteria and applied genotype refinement algorithms to increase the genotype accuracy and to produce phased haplotypes. After removing the related individuals we generated a reference panel for imputation and we genotyped *in silico* a total of 15,000 individuals involved in a study of quantitative traits in the Ogliastra region and in a study of autoimmune diseases in Sardinia. We evaluated the benefits of using a Sardinian reference panel to genotype *in silico* Sardinian individuals and we showed the improvement in term of genotyping accuracy and power to detect association signals especially for rare and Sardinian specific variation.

We detected a total of 17.6 millions variants, 30.6% of these variants are novel when comparing to the latest release of dbSNP and more than 5 millions variants (50% of the rare variants) are novel in the allelic frequencies below 1%.

We compared the variant sites detected in our project with the variant found by the 1000G consortium in the European, African and Asian population, evaluating the allele sharing between these populations and the Sardinian population.

Finally, we provided a real example on the detection of a stop-codon variant in the *HBB* gene and we showed how the detection of this signal is problematic when using large scale approaches based on GWAS arrays and imputation using reference panels not containing Sardinian individuals.

6.2. Future work

Our future plans include the analyses and application of the existing pipeline to an extension of the existing dataset to arrive to a total sample size of 3400 individuals sequenced.

We will also perform the variant calling of complex variants, like CNV, indels and structural variation to better understand the Sardinian variability and to have a more complete overview of the genetic underlying the phenotypic variability in the quantitative traits of the SardiNIA project and in the autoimmune diseases studied in our case-control study.

We will also use the sequencing data to start some analysis on the population genetics of Sardinia, trying to evaluate coalescence times for our sequences, differences in allele frequencies with Europeans and selection signals that may affect genetic diversity in Sardinia.

7. References

[1] NHGRI GWA Catalog www.genome.gov/GWAStudies

[2] International HapMap consortium, *The International HapMap Project*, 2003

[3] Pilia et al, *Heritability of cardiovascular and personality traits in 6,148 Sardinians*, 2006

[4] Pugliatti et al, *The epidemiology of multiple sclerosis in Europe*, 2006

[5] Sanna et al, *Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis* , 2010

[6] Mardis, *Next Generation DNA sequencing methods,* 2008

[7] http://bio-bwa.sourceforge.net/

[8] http://genome.sph.umich.edu/wiki/VerifyBamID

[9] http://genome.sph.umich.edu/wiki/UMAKE

[10] http://samtools.sourceforge.net/mpileup.shtml

[11] http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap

[12] Sanna et al, *Fine mapping of five loci associated with low-density*

lipoprotein cholesterol detects variants that double the explained heritability, 2011

[13] Veeramah et al, *Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity*, 2011

[14] November et al, *Genes mirror geography within Europe*, 2008

[15] Li et al, *Low-coverage sequencing: Implications for design of complex trait association studies*, 2011

[16] http://www.sph.umich.edu/csg/abecasis/MACH/

[17] 1000G Project Consortium, 2010 and 2012

[18] Abecasis et al, *Merlin-rapid analysis of dense genetic maps using sparse gene flow trees*, 2002

[19] Cao et al, *Beta Thalassemia*, 2010

[20] Maioli et al, *Plasma lipids in beta-thalassemia minor*, 1989