Master of Science in Omics Data Analysis

Master Thesis

Bioinformatics Pipeline for Next Generation Sequencing Analysis in Association Studies of Idiopathic Pulmonary Fibrosis

by

José Miguel Lorenzo Salazar

Supervisor: Dr. Carlos Flores Infante

Unidad de Investigación, Hospital Universitario N.S. de Candelaria

Carretera del Rosario s/n, 38010 Santa Cruz de Tenerife

Department of Systems Biology

University of Vic – Central University of Catalonia

September 21th, 2015

Dedication

In Memory of María del Pilar Verdugo Rodrigo.

« *A healthy person has many wishes,*

but the sick person has only one »

(Indian proverb)

ACKNOWLEDGEMENTS

The completion of this master project has been a long journey and sometimes stressful. I could not have succeeded without the invaluable support of many people.

First of all, I would like to express my gratitude to my master thesis supervisor, Dr. Carlos Flores Infante, who brought me to this exciting area of research on complex diseases, and particularly Idiopathic Pulmonary Fibrosis. He has always offered excellent guidances, gentle encouragement, and consistent help. He has also challenged me with many interesting genetic problems during my work with his research group at the Research Unit of the University Hospital Nuestra Señora de Candelaria, Tenerife, Spain. During this research and study year, I have benefitted from the group atmosphere and research resources, as well as open discussions about genetic related subjects. Particularly, I would like to thank him for broadening my knowledge in applied Biology and Medical Science, and for providing worldwide collaborations. Special thanks are also for Dr. Carlos Flores lab's staff, Ms. Marialbert Acosta Herrera, Dr. María Pino-Yanes, Dr. Fabián Lorenzo Díaz, and Ms. Amalia Barreto-Luis.

Dr. Imre Noth and Dr. Shwu-Fan Ma (University of Chicago, Chicago IL, USA) are gratefully acknowledged for providing the NGS data of this master's work. Dr. Shwu-Fan Ma and Justin Oldham are also acknowledged for their throughly revisions of this text.

I would also like to thank the technical staff of SAII (Servicio de Apoyo Informático a la Investigación de la ULL), José Lucas Grillo Lorenzo y Luis Cerrudo Concepción, for the high quality service and fast response they have provided to manage DRAGO server facilities at La Laguna University during the development of the bioinformatics pipeline.

Special thanks are for Dr. Malu Calle Rosingana, Associate Professor at the Systems Biology Department, University of Vic – Central University of Catalonia and coordinator of the Chair of the Master's degree in Omics Data Analysis, Grup de Recerca en Bioinformàtica i Estadística Mèdica, at UVIC University, for her guidance, kind concern, and constant encouragement.

Many friends and colleagues have shown their warm caring, concern and support during this intense learning period. They have constantly encouraged me with their kind

words despite the fact they may not understand why I devote so much personal time to research. Alejandra García Marrero, the principal of the high school where I spend most of my time, understands that research requires time and effort that I always feedback to my own students. Thanks to Alejandra for allocating my schedules and allowing me to attend workshops and seminars related to my thesis. Thanks to Professors Lirta Marrero Vera and Dr. Alicia Vega Déniz, my English teachers and colleagues, for helping me write this dissertation. Thanks to Professor Carlos Baeta Bayón, for the insights he taugh me in Biology. Thanks to Dr. Pedro Salazar Carballo, for providing me inaccessible papers and his ability to always cheer me up. Thanks to Félix Martín Pérez, my friend and colleague, for his support and providing me informatics assistance when I faced hardware troubles.

I would like to express my gratefulness to full Professor Dr. José Hernández Armas (Chief of the Laboratory of Medical Physics and Environmental Radiactivity, La Laguna University, Tenerife, Spain) and Dr. Blas Fumero Fernández (Chief at the Teaching Training Service, Department of Education, Canary Islands Government, Spain). They helped me to conduct active research and participate in teacher training activities, respectively, which have allowed me to obtain finances to pursue my master studies.

I would like to thank my parents-in-law Carmita and Juan, my sister Isabel and her family, and my friends Chayo, Trini and Chano for taking care of my children while I was studying and doing research. Thanks to Juanma and Juan, at Taoro Consultores, and Clara, Guille and Dácil, at Garoé Sur, for providing facilities in the writing phase of this work. Hermana Candelaria Pérez Cejas and Colegio Pureza de María (Sant Cugat del Vallès, Barcelona) are gratefully acknowledged for their friendly logistic support.

My deep gratitude is for my beloved wife Mari Cruz who gives unfailing and encouraging support at all times. My daughters Irene and Elisa always energize me continue with my studies, despite the fact they do not yet understand why I am always studying.

Finally, there is no doubt that numerous people, probably countless, were involved in making this two-year master dissertation possible. I apologize for not mentioning them, but please know that your collaboration and help is honestly appreciated. Thank you all.

ABSTRACT

A complete bioinformatics pipeline for Next Generation Sequencing (NGS) analysis has been developed and applied to study the association of called variants with susceptibility in Idiopathic Pulmonary Fibrosis (IPF). This bioinformatics pipeline integrates the Genome Analysis Toolkit (GATK) with state-of-the-art bioinformatics tools such as quality control reporters, aligners, alternative callers (i.e. Platypus), annotators, and auxiliary tools. The pipeline executes a sequence of SBash and Bash shell scripts by queuing the programmed jobs to a SLURM queue at a cluster server provided by La Laguna University (ULL). It is also executable with a local Linux machine.

We tested the pipeline by calling single nucleotide polymorphisms (SNPs) in targeted NGS data from 192 individuals with IPF, where 16,253 variant sites were identified. The call concordance between the two utilized callers (GATK and Platypus) was estimated at 77.8% when we compared matching overlapping sites. With this data, an association study following an unmatched case-control design was performed using unrelated European individuals (n=501) from The 1000 Genomes Project as controls. Logistic regression models were applied to the phenotype trait using genotypes from the 10,245 SNPs with call rates >95%, adjusting with five principal components to account for population stratification. Despite the reduced sample size, we identified 38 variants reaching genome-wide significance ($p<5x10^{-8}$), including one previously identified in the promoter region of *MUC5B* gene (rs35705950), and several other novel susceptibility variants.

TABLE OF CONTENTS

LISTS OF TABLES

LISTS OF FIGURES

ABBREVIATIONS

Abbreviation	Description
ATP11A	ATPase, Class VI, Type 11A
BAI	Index file of a sorted by position and indexed BAM file
BAM	Binary version of a SAM file
BED	A tab-delimited text file that defines a feature track
BIM	Extended variant information file accompanying a .bed binary genotype table
BP	Nucleotide location (aliases bp, pos, position)
chr	Chromosome
CRHR1	Corticotropin Releasing Hormone Receptor 1
DPP9	Dipeptidyl-Peptidase 9
DSP	Desmoplakin
FAM	Individual information file accompanying a .bed binary genotype table
FASTA	A text file used to specify the reference sequence for an imported genome
FAM13A	Family With Sequence Similarity 13, Member A
GATK	Genome Analysis Tool Kit
gVCF	Genomic Variant Calling Format
GW	Genome-wide
GWAS	Genome-Wide Association Study
hg19	Human Genome assembly version 19
HTS	High-Throughput Sequencing
HWE	Hardy-Weinberg Equilibrium
IMP5	See SPPL2C
INDEL	Insertion / Deletion
IPF	Idiopathic Pulmonary Fibrosis

Abbreviation	Description
KANSL1	KAT8 Regulatory NSL Complex Subunit 1
MAF	Minor Allele Frequency on a certain population
MAP	Variant information file accompanying a .ped text pedigree + genotype table
MAPT	Microtubule-Associated Protein Tau
MDGA2	MAM Domain Containing Glycosylphosphatidylinositol Anchor 2
MUC5B	Mucin 5B, Oligomeric Mucus/Gel-Forming
NGS	Next Generation Sequencing
LD	Linkage Disequilibrium
OBFC1	Oligonucleotide/Oligosaccharide-Binding Fold Containing 1
p	p-value
PCA	Principal Components Analysis
PED	Standard text format for individual pedigree information and genotype calls
QC	Quality Controls
ROI	Region of Interest
SAM	A tab-delimited text file that contains sequence alignment data
STH	Saitohin
SNP	Single Nucleotide Polymorphism
SPPL2C	Signal Peptide Peptidase Like 2C
SNV	Single Nucleotide Variant
TERC	Telomerase RNA Component
TERT	Telomerase Reverse Transcriptase
TOLLIP	Toll Interacting Protein
ULL	University of La Laguna
VCF	Variant Calling Format

1. INTRODUCTION

Idiopathic Pulmonary Fibrosis (IPF) is a low incidence (4.6-16.3/100,000 person-year), devastating disease with unknown etiology and high mortality. IPF has a median survival of 3 years after diagnosis [1] and is characterized by a relentless progression in interstitial fibrosis and a progressive decline in gas exchange [2]. IPF, which typically affects adults males over the age of 65 [3], is difficult to diagnose and its clinical course unpredictable [4]. To date, lung transplantation remains the only successful treatment option for improving survival.

Single nucleotide polymorphisms (SNPs) in *TERT*, *TERC*, *RTEL1*, *SFTPA2*, and *SFTPC* genes have been firmly associated with IPF susceptibility, primarily in the familiar forms of the disease [5,6,7]. In addition, two independent genome-wide association studies (GWAS) have identified additional loci associated with IPF susceptibility, including *FAM13A* (4p22), *DSP* (6p24), *OBFC1* (10q24), *ATP11A* (13q34), *DPP9* (19p13), variants in chromosomal regions 7p22 and 15q14-15, *MUC5B-TOLLIP* (11p15), *MDGA2* (14q21.3), and *SPPL2C* (17q21) [5,8]. These findings suggest that the etiology of IPF might involve multiple genetic loci. However, although these common and rare variants have been shown to increase the risk of developing IPF, to date none of them has proven to be causal [9].

The GWAS of IPF conducted by our collaborators at University of Chicago compared 1,410 European-American IPF cases and 1,931 controls in a three stages analysis of > 10 million variants across the genome [8]. Six SNPs in three loci (*TOLLIP-MUC5B* at 11p15.5, *MDGA2* in 14q21.3, and *SPPL2C* at 17q21.31) achieved genome-wide significance with overall p-values$<5x10^{-8}$ in the second stage of the study. Common variants at 11p15.5 were also associated with IPF survival in two independent studies [6,8]. Preliminary analyses suggest that the association between *TOLLIP* genetic variants and IPF susceptibility is independent from that found at *MUC5B* [7,10], although with milder effects in disease risk compared to that observed with the latter. Interestingly, *TOLLIP* encodes the Toll interacting protein, a critical regulator of Toll-like receptor (TLR)-mediated innate immune responses and transforming growth factor-β (TGF-β) signaling pathway. While the promoter polymorphism (rs37505950) is associated with IPF susceptibility, this variant is paradoxically associated with a slower disease progression and

improved survival [6,11], suggesting that it may constitute a subset of the disease [12]. Clustering of the differentially expressed genes in IPF individuals compared with controls highlighted a plausible classification of IPF subpopulations based on molecular signatures [13,14].

The study of target enriched Next Generation Sequencing (NGS) was driven by to amend the recent evidence demonstrating the importance of rare variants in genomic regions identified by GWAS, and because these rare variants are suboptimally covered by genome-wide genotyping arrays. NGS of the three previously identified genomic regions was done in a subset of IPF patients previously identified in GWAS. This fine mapping is expected to provide additional novel IPF susceptible variants with larger susceptibility risk.

2. AIMS

The main goal of this master project is to setup and optimize a bioinformatics pipeline to analyze target enriched NGS data at loci identified in previous GWAS [8] and to perform a case-control association study to identify IPF susceptible variants.

The constructed bioinformatics pipeline will focus particularly on the following subaims:

- To integrate data quality assessment and control steps following standard protocols for association analysis.
- To compare and contrast the consensus of variants identified by the Genome Analysis ToolKit (GATK) and Platypus callers after functional annotation provided by snpEff and ANNOVAR.
- To conduct an association study using individuals of European descent from The 1000 Genomes Project to discover novel variants with larger effects on IPF susceptibility by means of GATK and PLINK (a toolset for whole-genome association and population-based linkage analysis) software packages.

With these goals in mind, we have developed a pipeline on a desktop machine running Linux Ubuntu 14.04 LTS and tested on DRAGO, a server that is suitable for shared memory computing at La Laguna University, Tenerife, Spain.

3. METHODOLOGY

3.1. Study design

To accomplish the described goals, we have performed the following steps:

1. Select cases for targeted sequencing
2. Conduct next generation sequencing of the three regions of interest (ROIs)
3. Perform initial NGS data management
4. Develop a pipeline for genotype data extraction
5. Perform association studies with IPF susceptibility

Steps 4 and 5 are the main research context of this master's work.

Data was derived from 192 subjects with an average age at diagnosis of 68 years (Table 3-1) with respiratory symptoms including dyspnea on exertion and/or cough for at least three months. A high-resolution computed tomography scan with a definite or probable usual interstitial pneumonia (UIP) pattern was required from each patient in accordance with published guidelines [15]. A surgical lung biopsy confirming UIP was obtained in 37.3% of affected subjects utilized. None of the subjects had a record of a clinically significant exposure to known fibrogenic agents or suffered from other known causes of interstitial lung disease.

The range of cases is diverse in severity and source. Patients were selected from the University of Chicago (n=149), "Correlating Outcomes with biochemical Markers to Estimate Time-progression" (COMET) study in idiopathic pulmonary fibrosis (n=22), and the "Anticoagulant Effectiveness in Idiopathic Pulmonary Fibrosis" (ACE) study (n=21). We have intentionally selected a diverse array of cases with the aim of evaluating variants within the three regions (chromosomes 11, 14, and 17) previously associated with disease susceptibility. Establishing the deep coverage sequence of these regions will allow us to conduct the requisite analyses for determining significant variants and their relationship to IPF susceptibility.

Table 3-1. Clinical characteristics of the 192 patients with idiopathic pulmonary fibrosis.

Age at diagnosis, mean years (IQR)	68 (63-75)		
Sex (%)	Female	48 (25.0)	
	Male	144 (75.0)	
Smoking status (%)	Known	Never smoker	46 (24.0)
		Ever smoker	133 (69.3)
	Unknown	13 (6.7)	
FVC (% predicted)	65.8		
D_LCO (% predicted)	46.0		

IQR: Inter-quartile range. FVC=forced vital capacity. D_LCO=diffusion capacity of lung for carbon monoxide.

DNA was extracted from peripheral blood using QIAamp® DNA Blood Maxi kit from Qiagen (Valencia, CA) following manufacturer's protocol. The quality of genomic DNA (gDNA) samples was assessed by either TapeStation (Agilent Technologies, Santa Clara, CA) or 1% eGel (Life Technologies, Carlsbad, CA). Sample concentrations were determined using the Qubit dsDNA BR Assay (Life Technologies). One microgram of each high quality gDNA sample was sheared to an average peak size of 200 bp using the Covaris S-220 acoustic shearing device (Covaris Inc., Woburn, MA) according to the manufacturer's instructions. The proper size distribution of sheared gDNA fragments was confirmed by TapeStation analysis (Agilent Technologies). Sequencing (>100x depth coverage, 1.3 Gb/sample) was performed using the SureSelect™ Target Enrichment System XT2 kit (Agilent Technologies) to derive target-enriched DNA samples with a custom-design capture of approximately 1.7 Mb of human genome sequence. Briefly, the sheared gDNA samples were end-repaired, A-tailed and ligated with pre-capture indexing adaptors. The adaptor-ligated libraries were then amplified in five PCR cycles with the Herculase II Master Mix (Agilent Technologies). Library sizes were checked by TapeStation (Agilent Technologies). The concentration of each library was determined by Qubit dsDNA HS Assay (Life Technologies). Libraries were combined to form pre-hybrization pools with a total of 1500 ng from equal molar contribution of eight individual indexed libraries. The library pools were concentrated to 7 μL in a vacuum centrifuge to prepare

for target enrichment. Hybridization with the designed capture library was carried out for 24 hours at 65°C. Post-hybridization bead enrichment and stringent washes were executed as described in the standard protocol. The enriched library pools were amplified with 10 PCR cycles with the Herculase II Master Mix (Agilent Technologies). Correct library size and concentration were again determined by TapeStation and Qubit dsDNA HS, respectively. Libraries were further pooled to contain equal amounts of enriched DNA from each of 64 samples. The final library pools were quantified by quantitative PCR with the Kapa Biosystems Library Quantification Kit - Illumina (Kapa Biosystems Inc., Wilmington, MA). Paired-end reads of 100 bases were then generated on the HiSeq2500 platform from Illumina (San Diego, CA). All services were provided by DNA Services Facility of Research Resources Center at the University of Illinois at Chicago (http://www.rrc.uic.edu/dnas).

Table 3-2. Location of regions-of-interests (ROIs).

Chromosome	Size (bp)	Start position	End position
11	218,123	1,212,769	1,430,892
14	835,629	47,308,828	48,144,457
17	655,030	43,672,710	44,327,740

3.2. NGS pipeline overview

The collected NGS data were first analyzed using an Illumina's Exome pipeline modified by the Bioinformatics Core within the Center for Research Informatics (CRI, http://cri.uchicago.edu) at the University of Chicago. This analysis consisted of performing raw data quality controls (QC), filtering, and mapping (including SeqPrep, FastQC, NovoAlign, Picard, and SAMtools) to generate the BAM/BAI files that constitute the starting point for this project.

To complement and extend the downstream analysis of this data, we have programmed a complementary pipeline (Figure 3-1) comprising the use of the following software:

— **Qualimap** [16], a platform-independent application written in Java and R that provides both a Graphical User Interface (GUI) and a command-line interface to facilitate the QC of alignment sequencing data and its derivatives like feature counts.

— **Genome Analysis Tool Kit** (**GATK**) [17-19], a package developed at the Broad Institute (MIT and Harvard University) to analyze high-throughput sequencing data.

— **Platypus** [20], a variant caller developed by The Wellcome Trust Centre for Human Genetics, was used for comparative purposes.

— **SnpEff** [21], together with **ANNOVAR** [22], for functional annotation of genetic variants from high-throughput sequencing data.

— **LASER** [23], a program to estimate individual ancestry by directly analyzing off-target reads using CEPH-HGDP data from >900 worldwide samples as reference genotyped for >630.000 SNPs.

— **SAMtools** [24], a suite of programs for interacting with high-throughput sequencing data (including SAMtools, BCFtools, and HTSlib).

— **VCFtools** [25], a software suite that implements various utilities for processing VCF files, including validation, merging, comparing, etc.

— **PLINK** [26], a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale case-control analyses in a computationally efficient manner.

BEDtools [27], **fcGENE** [28], and other software used in the design and testing of this pipeline were listed in **Appendix A1**. Aforementioned, given the small overlap of regions between CEPH-HGDP reference panel utilized by LASER and the three regions of interests (ROIs), the principal component analysis (PCA) provided by **Eigensoft** [29,30] was used instead.

3.3. NGS pipeline description: phases and steps

The advances in massive sequencing technologies and related software make genome variations easy to identify and quantify. However, genotyping of these variations are still challenging because each variant caller relies on its own algorithm to assign quality scores to individual base calls, resulting in different calls even when they are applied to the same sequencing data. Using the most stringent QC metrics the reproducibility of single nucleotide variant (SNV) call is around 80%, suggesting that erroneous variant calling can be as high as 20-40% in a single experiment [31]. With this in mind, a bioinformatics pipeline aimed at accurately discovering variants in high-throughput sequencing (HTS) data has been programmed. We have followed the DNAseq best practices workflows suggested by the GATK development team at the Broad Institute [17-19].

Among the wide spectrum of SNPs and structural variants (including copy number variants [CNVs], insertions and deletions; inversions and translocations), we optimized the bioinformatics pipeline using SNVs identified in defined genome regions harboring several genes.

This pipeline was designed with three distinct phases [32] and depending on the features and hardware resources of the computing system hosting the bioinformatics pipeline (Figure 3-1), allows for sequential runs while executing phase-specific steps in parallel. These three phases are the followings:

- Data preprocessing (performed by the hosting group)
- Variant discovery, genotyping, and filtering
- Variant evaluation and refinement of results

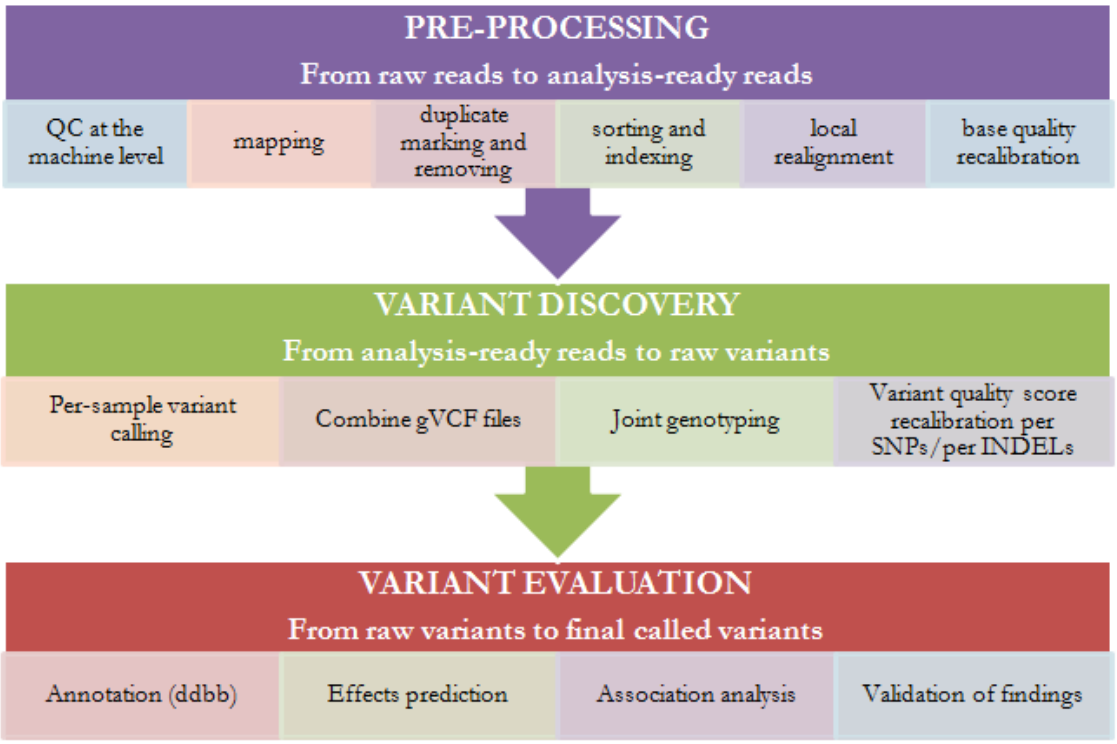

Figure 3-1. Pipeline for calling variants in DNAseq data from cohorts of IPF samples. Shown here is the pipeline with three distinct phases: pre-processing of raw data, variant discovery and variant evaluation.

The bioinformatics pipeline is a collection of shell Bash scripts to be executed in a sequential order on a Linux shared memory computing system equipped with a queuing system, though it is possible to run the complete pipeline on a desktop machine. It can be easily adapted to any other DNAseq experiments (e.g. exome sequencing) by modifying the programmed scripts.

To make data suitable for the variant calling analysis several data quality assessment and phase-specific steps must be performed. We will briefly summarize these steps in the following subsections.

3.3.1. Pre-Processing

As a first step, SeqPrep was used to remove sequence adaptors and merge overlapping reads from the same DNA fragment. In parallel, the quality of the raw reads was checked using FastQC, which produces a summary of the read quality, including %GC, per base quality, duplicate level, etc.

Data preprocessing is a required phase to make raw sequencing data suitable for downstream analysis. It comprises non-GATK and GATK steps (Figure 3-2).

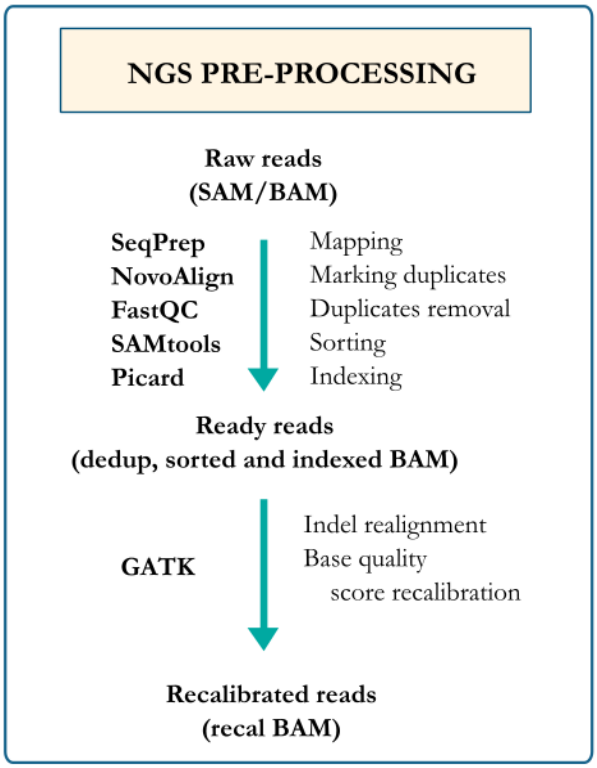

Figure 3-2. DNAseq pre-processing steps. In this stage of the pipeline, we move from raw reads (SAM/BAM files) provided by the sequencer software into recalibrated reads (recal BAM files).

3.3.1.1. Mapping and Marking Duplicates

As the goal of the experiment was to achieve high coverage in the targeted genome regions, there is a relatively high degree of duplication in the experimental data set. Therefore, removing of duplicates after alignment was also performed.

After QC, raw reads of each sample were mapped against the hg19 human reference genome using the NovoAlign mapper. Of note, the Burrows-Wheeler Aligner (BWA) [33] was used previously by the hosting group to test this data, but resulted in suboptimal results, especially within repetitive regions in the ROI in chromosome 17 (not shown). Alignment files were generated in compressed BAM format, which contains information on the numbers of aligned reads from each original read file, and the number of duplicates (based on alignment to identical regions of the genome, not on sequence identity) identified. It also lists the numbers of reads that fall within the targeted regions, and the degree of enrichment within the targets relative to the rest of the genome. For each sample, reads were enriched (>200x times) within the ROIs (on-target) vs. outside regions (off-target).

After aligning the individual groups by reads for each individual sequence file, alignment files from the same sample were merged. Reads aligning outside the targeted regions were discarded, and duplicates again removed. These alignment files were then merged to create a single alignment file representing all reads in all samples that fall within the ROIs. Sorting and creation of index BAI files were also performed.

Mapping and marking of duplicates were performed by the University of Chicago Bioinformatics Core following their pipeline named '*Illumina Exome pipeline, with modifications*' as described in the preceding paragraphs. As a result, aligned BAM files against hg19 were provided as the starting point for this work.

3.3.1.2. Realignment around Indels

BAM/BAI files generated by the previous step constitute the input of the GATK pipeline described in this masterwork. A preliminary QC with FastQC and Qualimap software was initially performed to check that duplicates were in fact removed and to make sure that aligned sequences were sorted and indexed.

The detection of genetic variants from NGS data is prone to errors due to multiple factors such as base-calling, alignment errors, read coverage, etc. Therefore, identification of genetic variants is an area of active research and many statistical methods are being developed to improve and quantify the large uncertainty associated with genotype calling [17,34,35].

Algorithms used in the initial mapping step tend to produce various types of artifacts, such as reads aligned on the edges of insertions/deletions (indels), which represent the most common structural variants implicated in the pathogenesis of various disease states [36,37]. Indels often get mapped with mismatching resulting in a false SNP in that locus. Our pipeline was therefore designed to perform a new realignment around indels within the BAM. GATK parses the BAM files, generates a list of target intervals, and performs a local realignment around indels.

3.3.1.3. Base Quality Score Recalibration

Variant calling algorithms depend on the quality scores assigned to the individual base calls in each sequence read. As these quality scores may be affected by different sources of systematic error, they must be recalibrated. GATK applies a machine-learning algorithm to model these errors empirically and adjust the quality scores accordingly [17]. This recalibration is run twice and R script is used to generate several graphical plots to visualize the effects of the recalibration process (Figure 3-3).

a)

b)

c)

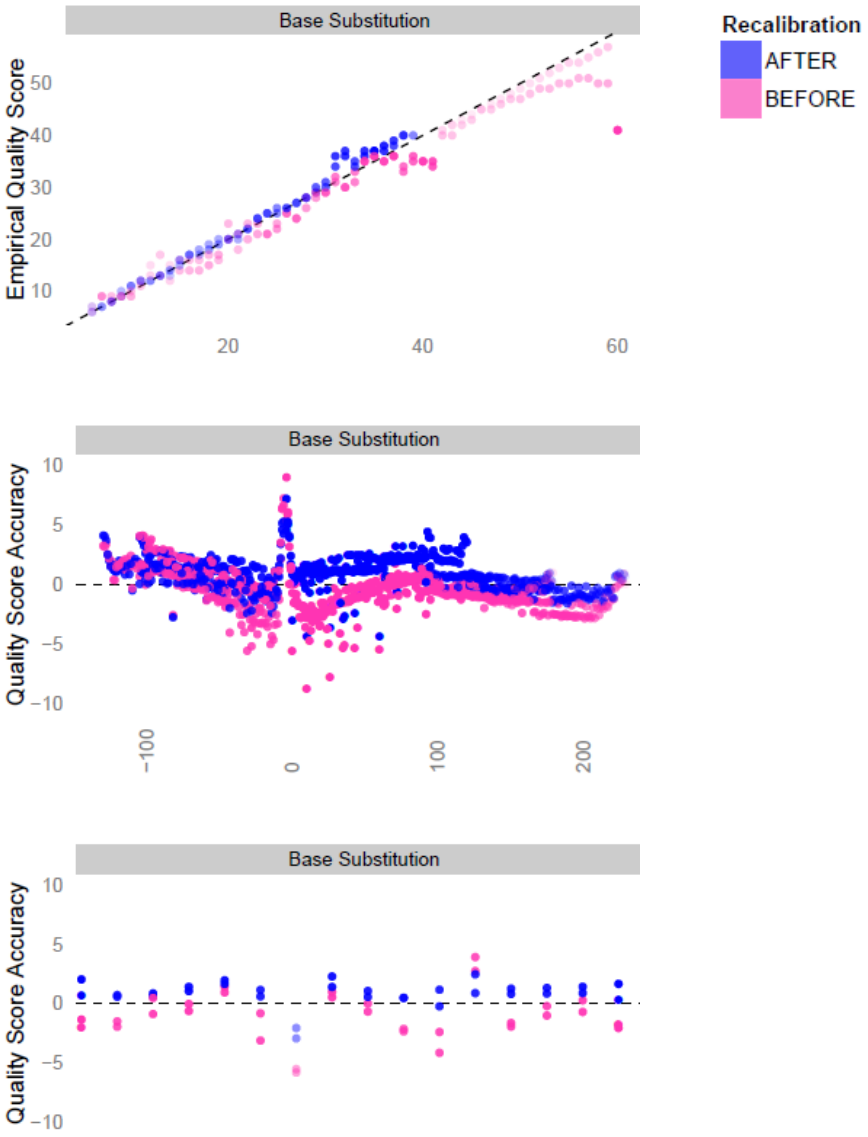

Figure 3-3. 'Before' and 'After' recalibration plots for a recalibrated BAM file corresponding to one IPF individual for the event 'Base Substitution' (insertion and deletion events are also analyzed). a) Empirical quality score; b) Quality Score Accuracy for pair-end reads; c) Quality Score Accuracy residuals.

The recalibration process is divided in two phases. The 'before-and-after' recalibration plot displays first pass recalibration values in pink, which are obtained from applying the GATK BaseRecalibration walker on the original alignment. Second pass recalibration values are shown in blue, and correspond to results obtained from the application of the GATK BaseRecalibration walker on the alignment recalibrated using the first pass tables.

3.3.2. Variant Discovery, genotyping and filtering

This phase is the bottleneck of the variant calling. The success in this phase is a trade-off between minimizing false negatives or Type-II errors (sensitivity gain) and minimizing false positives or Type-I errors (specificity gain). To do so, GATK uses separate steps (Figure 3-4): variant calling (performed on a per-sample basis), joint genotyping (performed per-cohort of the 192 sample files) and variant filtering (performed per-cohort). The first two steps are designed to maximize sensitivity, while the filtering step aims to deliver a level of specificity that can be customized for each project.

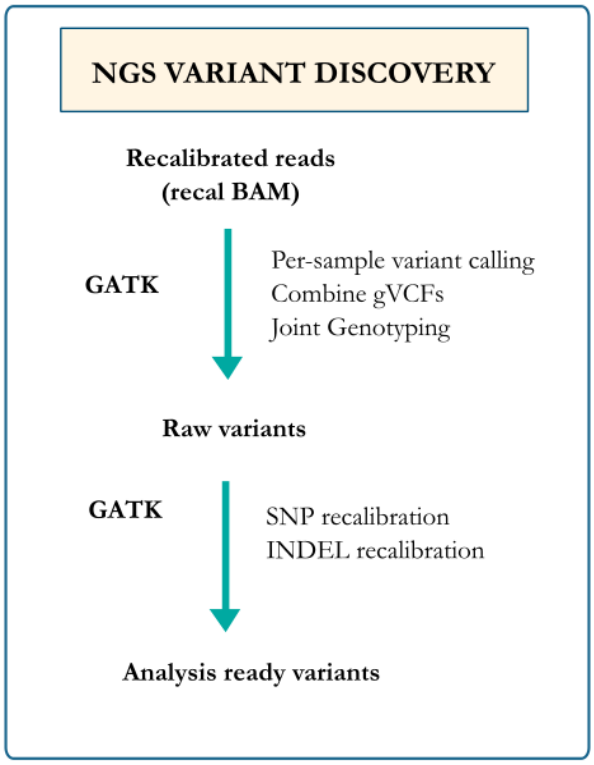

Figure 3-4. Flow diagram depicting the variant discovery phase. In this stage of the pipeline, we move from recalibrated reads into called variants for downstream analysis.

In the variant discovery, genotyping, and filtering phase, we moved on from reads (i.e. de-duplicated, sorted and indexed BAM files) to variants (i.e. analysis-ready VCF files). As a rule, the variant discovery depends on the type of sample (whole genomes, exomes, etc.) and other parameters related to the sequencing, such as: coverage, depth, quality of reading, etc. (Figure 3-4).

3.3.2.1. Per-Sample Variant Calling

In this step, GATK generates a multi-sample SNP and indels calling with the HaplotypeCaller walker. The aim is at simultaneously calling SNPs and indels via local de-novo re-assembly of haplotypes in active sequences of the targeted regions. As a result, 192 genomic VCF (gVCFs) files are prepared. As a second step, the bioinformatics pipeline combines all the gVCFs files into a single cohort gVCF file prior to performing joint genotyping.

3.3.2.2. Joint Genotyping

The GenotypeGVCFs walker creates a set of raw SNP and indel calls from the gVCF cohort file that will undergo a variant quality score recalibration. According to the GATK DNAseq best practices, this cohort-wide analysis empowers sensitive detection of variants at complex loci. This is a multi-sample joint aggregation step and merges the records together in a sophisticated manner: at each position of the input gVCF, this tool will combine all spanning records, produce correct genotype likelihoods, re-genotype the newly merged record, and then re-annotate it.

3.3.2.3. Variant Quality Score Recalibration

Variant recalibration is based on a machine learning method that assign a calibrated probability to each variant call in the raw call set. In this step, GATK functions to reduce the chance of missing real variants and to discard false positives. GATK then uses this variant quality score to filter the raw call set in a second step, thus producing a subset of calls with a desired level of quality, fine-tuned to balance sensitivity and specificity. This calibration is processed in two steps separately: one for SNPs and one for indels. GATK also produces graphical plots to visualize the results of the mapping quality parameters and the models used in the Variant Quality Score Recalibration (VQSR) (Figure 3-5).

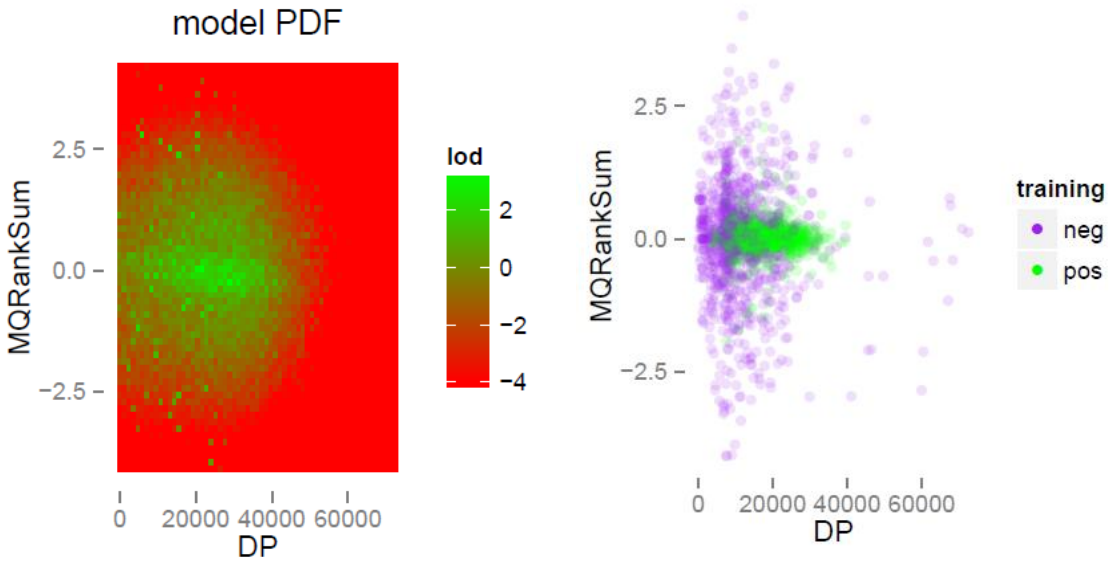

Figure 3-5. 2D projection of mapping quality parameters used by GATK in the Variant Quality Score Recalibration for an IPF individual.

In Figure 3-5, the MQRankSum is plotted against DP (Depth of Coverage). This variant-level annotation compares the mapping qualities of the reads supporting the reference allele to those supporting the alternate allele. The ideal result is a value close to zero, which indicates there is little to no difference. A negative value indicates that the reads supporting the alternate allele have lower mapping quality scores than those supporting the reference allele. Conversely, a positive value indicates that the reads supporting the alternate allele have higher mapping quality scores than those supporting the reference allele [17]. VSQR develops a continuous, covarying estimate of the relationship between SNP call annotations (e.g. MQRankSum, HaplotypeScore, DP, etc.) and the probability that a SNP is a true genetic variant versus a sequencing or data processing artifact, resulting in plots similar to the one shown above.

In addition, the VQSR provides a continuous estimate of the probability that each variant is true, allowing one to partition the call sets into quality tranches defined by the user, typically at 90, 99, 99.9, and 100% thresholds that correspond to levels of sensitivity relative to the truth sets used in the training. The basic idea behind this procedure is that, with well-calibrated variant quality scores, the user can generate call sets in which each variant does not require a binary answer as to whether it falls within the set. If a very high

accuracy call set is required, one can use the highest tranche. Conversely, if a larger or a more complete call set is of higher priority, one can use progressively lower tranches.

3.3.3. Annotation

Once variants have been annotated, detailed evaluation and refinement of results are required to ensure calls are not artifacts (Figure 3-6). This is a necessary step before accepting a putative variant as a risk, neutral, or protector element in the association with the disease. In this phase, genotype refinement, functional annotation and additional QC are performed.

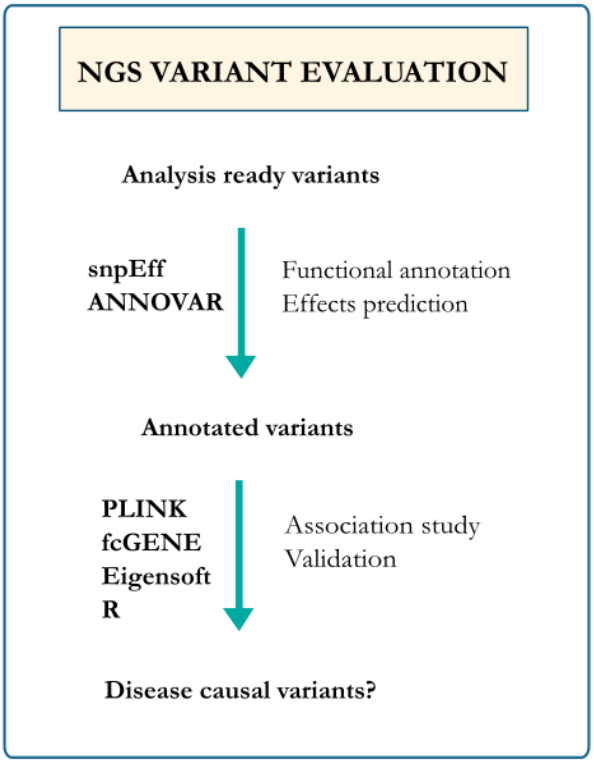

Figure 3-6. DNAseq variant evaluation and refinement of results. In this stage of the pipeline, the analysis-ready variants are annotated and further analyses are performed (i.e. association studies and validation of variants).

As a result, GATK provides a VCF file that inputs the variants evaluation phase. The designed bioinformatics pipeline combines two annotator alternatives, snpEff and ANNOVAR.

SnpEff software analyzes the input, annotates the variants and calculates the effects they produce on known genes (e.g. amino acid changes).

ANNOVAR allows for very flexible annotations, and can include any number of a wide variety of annotation types including, in principle, any track from the UCSC genome browser. For this project, the VCF file has been annotated with the allele frequencies from a variety of public data sets (e.g. NHLBI Grand Opportunity Exome Sequencing Project, The 1000 Genomes Project, Complete Genomics, COSMIC, and dbSNP).

In order to gain a better understanding on the variant calling and association study results, the GATK and Platypus VCFs were annotated with snpEff and ANNOVAR with different functional categories, including gene context, evolutionary conservation, and various functional predictions.

3.4. GATK walkers commands

GATK may be considered an ecosystem of specialized tools, called 'walkers', which one can use out of the box, individually or chained into scripted workflows. To do so, we have programmed a chain of shell Bash scripts to run a complete variant discovery workflow, from "BAM-to-VCF" (Table 3-3).

Table 3-3. GATK and non-GATK steps involved in the bioinformatics pipeline designed for the analysis of IPF NGS data.

PRE-PROCESSING			
Variant calling steps	**Description**	**Software**	**GATK 'walker'**
Mapping and Marking Duplicates (de-dupping)	1. Identify read group information	SeqPrep FastQC	---
	2. Generate a SAM file containing aligned reads	NovoAlign aligner	---
	3. Convert to BAM file, sort and mark duplicates	Picard	---
Local Realignment Around Indels	1. Create a target list of intervals to be realigned 2. Perform realignment of the target intervals	GATK	RealignerTargetCreator IndelRealigner
Base Quality Score Recalibration	1. Analyze patterns of covariation in the sequence dataset	GATK	BaseRecalibrator
	2. Do a second pass to analyze covariation remaining after recalibration	GATK	BaseRecalibrator
	3. Generate before/after plots	GATK	AnalyzeCovariates
	4. Apply the recalibration to your sequence data	GATK	PrintReads

VARIANT DISCOVERY			
Variant calling steps	**Description**	**Software**	**GATK 'walker'**
Variant Discovery	Calling Variants with HaplotypeCaller 1. Determine the basic parameters of the analysis	GATK	gVCF mode (-ERC GVCF cohort analysis workflow)
	2. Per-sample Variant calling	GATK	HaplotypeCaller (-emitRefConfidence GVCF)
	3. Optional data aggregation step	GATK	CombineGVCFs on batches of ~200 gVCFs
	4. Joint genotyping	GATK	GenotypeGVCF
	5. Variant Quality Score Recalibration	GATK	VariantRecalibrator ApplyRecalibrator
VARIANT EVALUATION			
Variant calling steps	**Description**	**Software**	**GATK 'walker'**
Selection	Selection of variants of interest, merging of VCF files (cases and controls), comparison of VCF contents, exporting toe PLINK format, analyze genotypes concordance, etc.	GATK	SelectVariants, VariantsToBinaryPed, VariantsToTable, VariantsToVCF, CombineVariants, GenotypeConcordance
Annotation	1. Annotate GATK VCF file for cases 2. Prepare VCF file for controls and annotate	SnpEff ANNOVAR	--- ---
Association analysis	1. Study of susceptibility in IPF	PLINK	---

3.5. Overview of GATK scripting for DRAGO cluster

The designed bioinformatics pipeline is suitable for the analysis of NGS data as well as whole-genomes and exomes data. It integrates a sequence of shell Bash scripts and can be executed on remote cluster servers or local machines. The scripting has been debugged in an i7-8 cores and 16 GB RAM machine running Ubuntu 14.04.02 LTS, and tested on a cluster server, DRAGO, provided by the University of La Laguna (Tenerife, Canary Islands, Spain).

DRAGO is a Red Hat-based workstation suitable for shared memory processes. It has four nodes with four processors per node, with 10 processor cores each, making a total of 160 cores available to researchers. It also has 1 TB of RAM and 4 TB in hard drives. This configuration currently lets DRAGO work as a single shared memory machine with 160 cores.

We access to DRAGO remotely by using a SSH tunnel from the LINUX shell or by using applications such as Putty (jobs management) or Filezilla (file managements). It uses a queue system based on SLURM (an open-source resource manager designed for Linux clusters of all sizes). It has five queues with different maximum execution time and number of available cores (sequential: 168 h/1 core; test: 5 min./9 cores; fast: 30 h/30 cores; medium: 168 h/40 cores; batch: 12 h/80 cores).

The basic scheme of Bash scripts to be queued at DRAGO is summarized in Box 3.5-1. An excerpt of the whole script is provided in **Appendix A2**. Additional excerpts of GATK and related commands are shown in **Appendix A3**.

Box 3.5-1. DRAGO SBATCH and BASH scripting scheme.

```
#!/bin/bash

###Queue name
#SBATCH --partition=<queue-name>

###Job name
#SBATCH -J <job-name>
```

```
###Number of nodes (i.e. 1)
#SBATCH --nodes=1

###Total number of processes (1 process = 1 core)
#SBATCH --ntasks=<cores>

###Execution Time
#SBATCH -t <HH:MM:SS>

###Email options: BEGIN, END, FAIL, ALL
#SBATCH --mail-type=ALL
#SBATCH --mail-user=<username@domain>

###Log outputs
#SBATCH --error=<output.log.err>
#SBATCH --output=<output.log.out>

###Load profile
source /etc/profile

### Loads specific modules in memory
source /etc/profile

### Loads Java 1.7 to run GATK
module add java/sun1.7

### Loads R 3.1.2 (make sure that ggplot2, gplots, reshape, grid, tools, and gsalib
libraries are already installed)
module add R/3.1.2

### Full path to software, auxiliary databases and sample data:
$INPUT(i)=/full-path-to-inputs/...

### Full path for outputs
$OUTPUT(j)=/full-path-to-outputs/...

## non-GATK or GATK 'walker' commands here
GATK <command><options> input.file output.file
```

3.6. Platypus variant caller

Besides GATK, we tested Platypus [20] as an alternative variant caller which also utilizes a haplotype-based variant calling algorithm similar to GATK.

An alignment-base approach is the most common and simple to call variants. This is done by aligning reads to a reference genome and finding locations where bases differ from the reference nucleotide. While this approach has a high sensitivity and does not require large computing resources, it has limitations. One is that alignment-based approaches focus on a single variant type, like SNP or indel, resulting in errors and high false positive rates around indels and larger variants.

By separating the processes of identifying and genotyping variants, even a weakly supported variant in one sample can be confidently called if it is strongly supported by another sample or samples. This approach reduces the rate of false negative calls due to downward fluctuations in read coverage, a feature that is important in comparisons of tumors and metastases, population-based studies and pedigrees including parent-offspring trios in de novo discovery designs [20]. The use of multi-sample variant calling helps in borrowing information between samples to call variants determined to be unreliable in a single sample. These methods are integrated into the GATK HaplotypeCaller and Platypus variant callers.

Platypus uses an alternative variant calling approach that plots reference-free sequence assembly builds on de Bruijn graphs to find evidence of polymorphisms. Such an approach works on the local haplotype level rather than on the level of individual variants and does well on highly divergent regions. This approach has large computational requirements, though Platypus is able to deal with targeted NGS data in notably reduced computing times as compared to GATK.

According to its authors, Platypus achieves high sensitivity and specificity for SNPs, indels and complex polymorphisms by using local de novo assembly to generate candidate variants, followed by local realignment and probabilistic haplotype estimation. It is an order of magnitude faster than existing tools and generates calls from raw aligned read data without preprocessing.

Platypus performs a three-stage algorithm pipelined fully transparent to the user as follows:

1. Read alignment and local assembly to produce 2^n candidate haplotypes.
2. Align reads to candidate haplotypes by fitting population frequencies and individual haplotypes.
3. Marginalize over individuals, variant-level filtering, and individual-level filtering to produce final called variants

Contrary to GATK, which generates many temporary files, Platypus runs the whole pipeline without using intermediate files or separate processes. With mapped and sorted BAM files are used as input, merging, sample de-multiplexing and read de-duplication are performed by Platypus under the same command execution.

3.7. Association study

3.7.1. Design

The association study design can be considered as an unmatched case-control one where a binary categorical disease outcome (be a case/be a control) is logistically regressed over the genotypes and other covariates on single SNP basis after careful QC.

The cases are composed of 192 individuals with IPF. A total of 501 unrelated controls were selected based on a reported European descent in The 1000 Genomes Project (1KGP) [38], available at http://www.1000genomes.org/data. Controls included individuals from the following regions: Utah Residents (CEPH) with Northern and Western European Ancestry (CEU); Finnish in Finland (FIN); British in England and Scotland (GBR); Iberian Population in Spain (IBS); and Toscani in Italy (TSI).

VCF files corresponding to chromosomes 11, 14, and 17 of the selected controls from the 1KGP Phase III (release date 20130502) were downloaded from its online server at the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/.

After using the aforementioned bioinformatics pipeline to produce VCF files for the targeted genome regions in cases and controls, the GATK CombineVariants walker

was used to combine variants. Next, VCFtools was used to remove indels and to filter out non-biallelic variants. Finally, VCFtools provided PED/MAP files for downstream QCs and association analysis with PLINK (Figure 3-7).

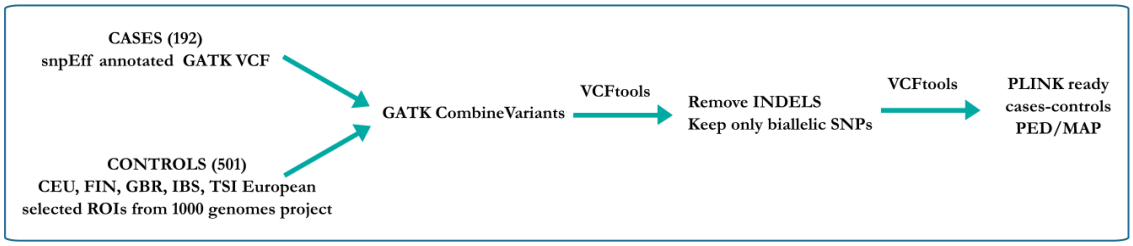

Figure 3-7. Worklow diagram showing the steps to be followed after GATK variant calling. Cases and controls are merged and indels are removed.

The association analysis was performed with PLINK 1.07 using the workflow depicted in Figure 3-8.

The preliminary part of the association study requires the addition of the phenotype information (step 1) and a basic study of the allelic frequencies and missing data (step 2) to account for missingness information. Increased false positives are particularly likely for cases-control studies, where case and control samples are likely to be collected and genotyped separately [39]. Due to the special design of this association study, where controls are gathered from an external database, QC for missingness issues on an individual and SNP-basis are required steps [40].

The removal of a sub optimal SNP is pivotal for the success of the association study, but care must be taken because the removal of a poorly genotyped marker might miss a potential disease-related variant [19]. Frequently, SNPs with a call rate < 95% are removed from the calling set [41,42]. For studies with small sample sizes, a call rate threshold of >97-99% and a marker of minor allele frequency (MAF) > 5% are used to keep only high quality genotyped individuals [43]. This is because the expected small size of the heterozygote and rare homozygote (homozygote for the alternative allele) clusters makes these variants difficult to call using current genotype calling algorithms and often results in false positives. As the power to detect rare variants is quite low [44], their removal will not affect the overall study. Given that we do not want to lose information at this stage, variants from the full spectrum of frequency will be considered for association,

albeit further inspections of those associated will be performed to discern whether associated variants are true variants or simply artifacts.

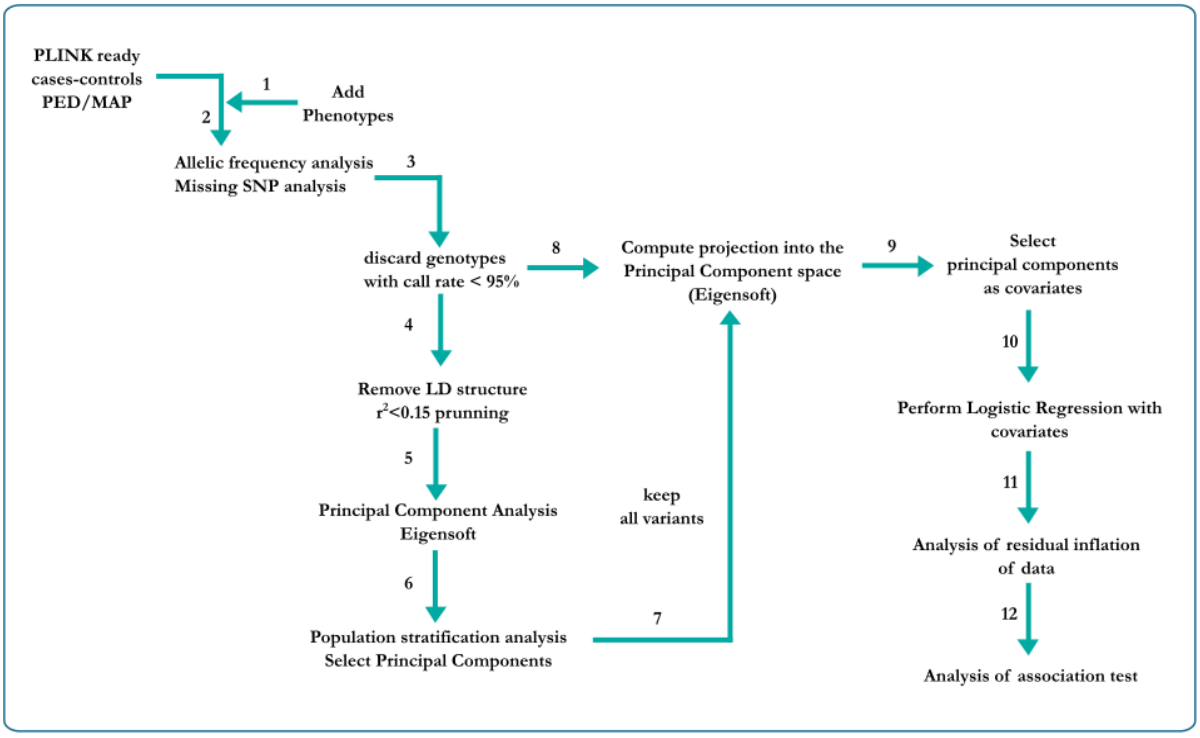

Figure 3-8. Association study workflow. LD=Linkage disequilibrium.

To do so, we discarded those genotypes with a call rate < 95% (step 3). SNPs showing a small-scale linkage disequilibrium (LD) are filtered out using a pruning with r^2<0.15 (step 4) in order to remove the linkage correlation structure with PLINK. These filtered data sets underwent a PCA (step 5) to determine the set of eigenvectors that optimally reflect the population structure (step 6).

Spurious association results may be the consequence of unaccounted genotyping issues. Therefore, it is necessary to perform additional post-association QCs. In genetic association studies, the deviation from the Hardy-Weinberg Equilibrium (HWE) of each SNP is frequently assessed as an initially QC to identify variants with questionable genotypes. HWE testing assumes that genotypes are sampled from the general population and are therefore trested only in controls [45]. Deviations of HWE can be due to inbreeding or random mating, population stratification (absence of mutation or migration) or lack of selection according to genotype [46,47], but it can be also represent a disease

association [48]. Given that in the current design, our controls were derived from the highly curated 1KGP database, this QC step was not considered. However, HWE p-values were assessed during the evaluation of the top IPF associated variants along with other QC metrics.

3.7.2. Population Stratification and Confounding

According to Astle and Balding [49] the main causes of confounding in GWAS are:

- *Population structure* or the existence of major subgroups in the population, where differences in allele frequencies are confounded with subpopulation.
- *Cryptic relatedness*, i.e. the existence of small groups (often pairs) of highly related individuals.
- *Environmental differences* between subpopulations or geographic locations.
- *Differences in allele call rates* between subpopulations.

The presence of population stratification and its effect upon the association test cannot be discarded in principle. Population stratification or population substructure represents systematic differences in allele frequencies between subpopulations within a sample, possibly due to different ancestry. As allele frequencies and disease stratification or admixture can confound the association between the disease trait and the genetic marker, it increases type I error (false positives) of association studies. This is particularly true when both the genotypic and phenotypic data differ between populations [50].

The number of principal components or main eigenvectors (step 7) selected to assess population stratification is frequently based on formal metric, such as the Tracy-Widom statistic (statistical significance of each principal component) [29,30,51]. Once the local LD features are removed, the top principal components should account for population structure, which on average affect all SNPs in the same manner. We keep for downstream analysis the original PED/MAP files with genotype calling rate > 95% (step 8). We then projected the whole variants into the principal components space and kept them as covariates (step 9) for the logistic regression (step 10) between the phenotype trait and the sample genotypes.

In a case-control study, the parameter of interest is the odds of disease. Because sampling is not at random, the odds of disease is not directly measurable. However, the odds of disease is mathematically equivalent to the odds of exposure, which can be directly calculated from exposure frequencies [52]. The quotient of exposure odds between cases and controls produces an odds ratio (OR), which quantifies the probability of disease based on exposure.

To demostrate this, let us consider a variant consisting of a single biallelic locus with alleles A and a. Unordered possible genotypes are A/A (homozygous for the reference allele), A/a (heterozygous) and a/a (homozygous for the alternative allele). The allelic OR describes the association between disease and allele by comparing the odds of disease in an individual carrying allele A to the odds of disease in an individual carrying allele a [51]. The genotype OR describes the association between the disease and genotype by comparing the odds of disease in an individual carrying a defined genotype (i.e. A/a or a/a) to the odds of disease in an individual carrying the reference genotype (i.e. A/A).

A logistic regression between the disease trait and sample genotypes is carried out adjusting for the PCA-derived scores from non-correlated SNPs to account for population structure in the analysis [53-56].

The logistic regression model establishes a relationship between a binary outcome variable (the phenotype trait: case versus control) and a group of predictor variables (genotypes and other covariates). It models the logit-transformed probability as a linear relationship with the predictor variables. More formally, let y be the binary outcome variable indicating cases/controls with 0/1 and p be the probability of y to be 1, p = prob(y=1). Let $x_1, \dots, x_k$ be a set of k predictor variables. The logistic regression of y on $x_1, \dots, x_k$ estimates parameter values for $\beta_0, \beta_1, \dots, \beta_k$ of the following equation:

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \qquad (1)$$

The left-hand side of Eq. (1) is the log of the ratio of the probability of disease compared to the probability of being a control (not having the disease), or the log of the odds of disease. Thus, the logistic regression model is a linear model for the log odds:

$$\left(\frac{p}{1-p}\right) = Odds \qquad (2)$$

The model parameters might be interpreted as the difference in log odds of the outcome associated with a one-unit change in the predictor variable while the other variables are kept constant [56].

The regression coefficients, β_k, let us compute how much the odds increase multiplicatively with a one-unit change in each independent variable.

In addition, the so-called genomic control technique can be applied to detect and compensate for the presence of a fine-scale or within-population stratification during association testing. If genomic control operates, population stratification is treated as a random effect that causes the distribution of χ^2 association test statistics to have an inflated variance and a higher median value than would otherwise be observed [54]. The test statistics are assumed to be affected uniformly by an inflation factor (λ). This factor is estimated from a set of selected SNPs by comparing the median of the observed test statistics with the median of their expected values under the assumption of no population stratification. The residual inflation is visualized by means of quantile-quantile or QQplots from Eigensoft outputs and/or using PLINK function "--adjust" to compute the level of residual inflation variation that persists after this preliminary adjustment (step 11). Under genomic control, if population stratification exists, $\lambda>1$ and the correction is simply applied by dividing the actual association χ^2 statistic values by λ to get a deflated-by-λ distribution.

The last step of this workflow performs the association study (step 12) by means of a logistic regression using the main eigenvectors to account for the population structure.

As already stated, the selection of case individuals was focused on European-American IPF patients following precise guidelines. Thus, we prepared two complementary workflows to detect the existence of such a population structure and estimate qualitatively the ancestry of IPF cases to avoid spurious results in the association.

Our initial approach was based on the use of LASER [23,58], a program to estimate individual ancestry by directly analyzing shotgun sequence reads without calling genotypes.

LASER relies on the availability of a set of reference individuals whose genome-wide SNP genotypes and ancestral information are known. Then it constructs a reference coordinate system by applying principal components analysis to the genotype data of the

reference individuals. Next, for each sequencing sample, it uses the genome-wide sequencing reads to place the sample into the reference PCA space. With an appropriate reference panel, the estimated coordinates of the sequencing samples identify their ancestral background and can be directly used to correct for population structure in association studies or to ensure adequate matching of cases and controls. These coordinates are computed by means of a Procrustes analysis to project the new ancestry map into the reference PCA space [58]. Very few targeted NGS sequence bases overlapped the 632,958 reference markers of CEPH-HGDP panel [59] so this approach using LASER did not provide results as expected (data not shown).

The alternative approach utilized (Figure 3-9) was based on PCA applied to selected genotypes within the same ROIs of 2,504 controls from 1KGP corresponding to five worldwide biogeographical populations. As a first step, VCF files from cases and controls are merged (1) by means of GATK CombineVariants walker. Then genotypes with a call rate < 95% are discarded (2). The small-scale LD structure is removed imposing a pruning filter of r^2<0.15 (3). PCA is performed using Eigensoft (4) and principal components are selected (5) after considering how population stratification is modified by increasing the number of eigenvectors. Ancestry of IPF samples is then estimated by plotting all samples, cases and controls, in the eigenvectors space.

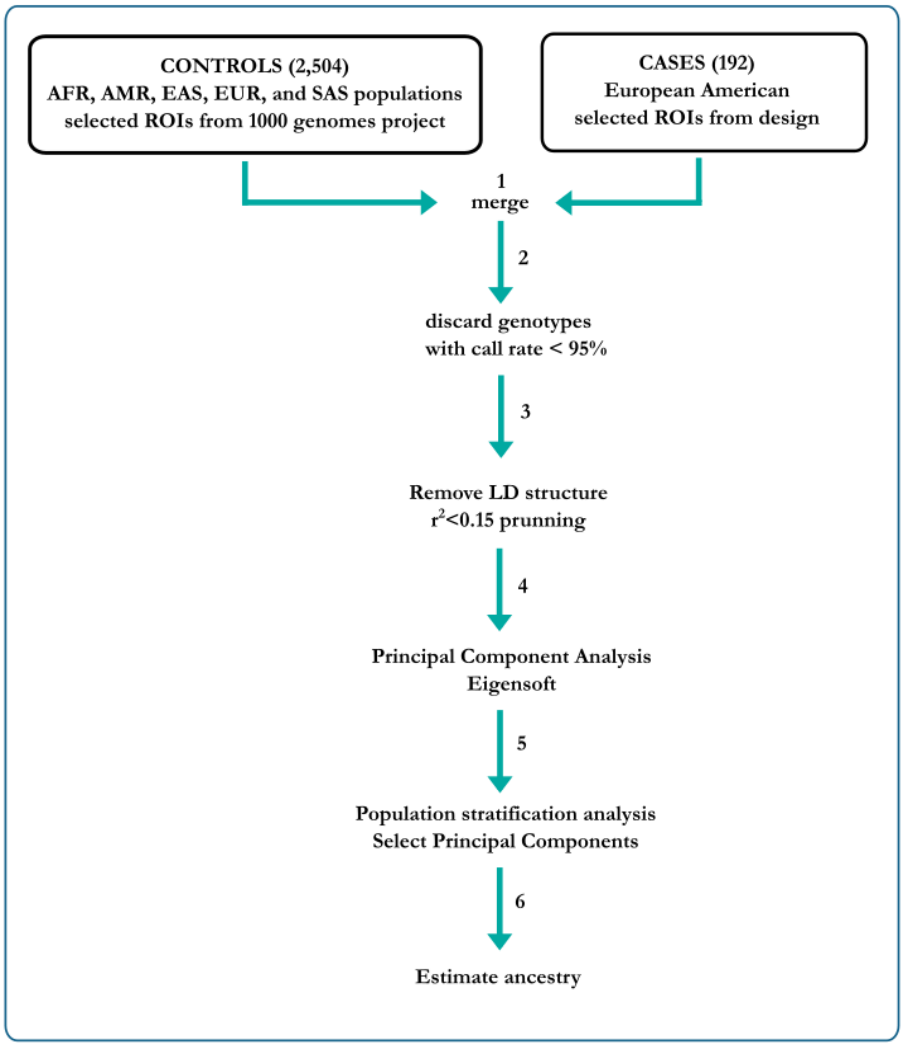

Figure 3-9. Workflow to estimate ancestry of IPF cases by means of PCA. LD=Linkage disequilibrium.

4. RESULTS AND DISCUSSION

4.1. Quality Controls

Three different ROIs were studied in IPF cases within chromosomes 11, 14, and 17 (Table 3-2). A total of 1.71 Mbp were fine mapped by NGS. A visual inspection of the processed BAM files (Figure 4-1) —in bigWig format— showed that the mean coverage depth at each of the ROI was >100x (Table 4-1).

a)

b)

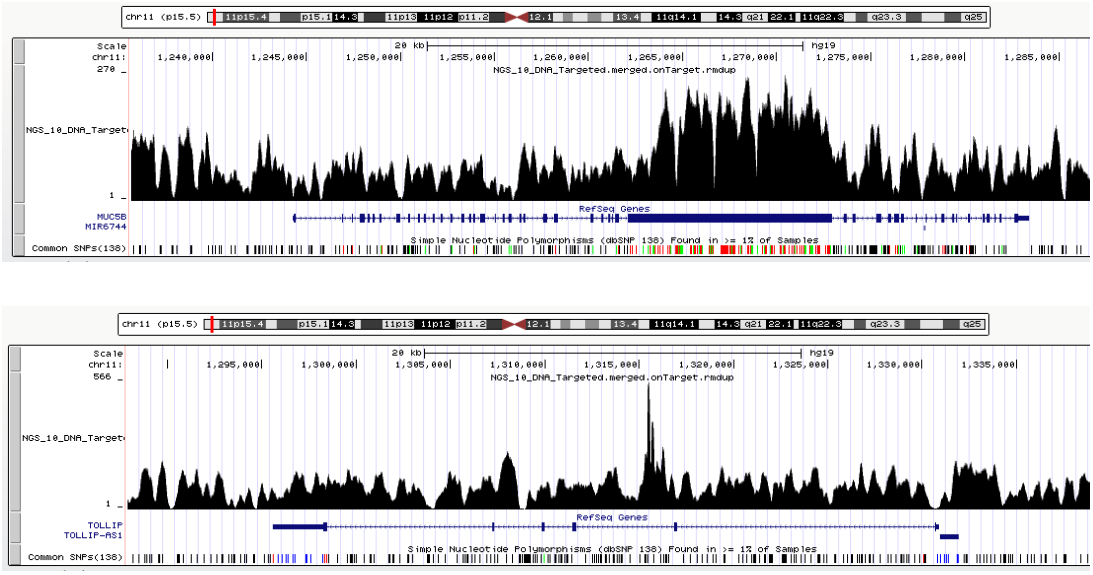

Figure 4-1. ROI in chromosome 11 visualized as an added track to UCSC Genome Browser for an IPF individual in bigWig format: a) *MUC5B* region; b) *TOLLIP* region.

Table 4-1. Quality control results as observed with Qualimap for all IPF samples.

Chromosome	Length	Mapped bases	Mean coverage
11	217,440	24,175,349	111.2
14	835,421	95,654,226	114.5
17	655,685	74,276,346	113.3
			113.6 ± 9.13

In addition, more than 98% of genomic locations within ROIs were covered at >89x (Figure 4-2).

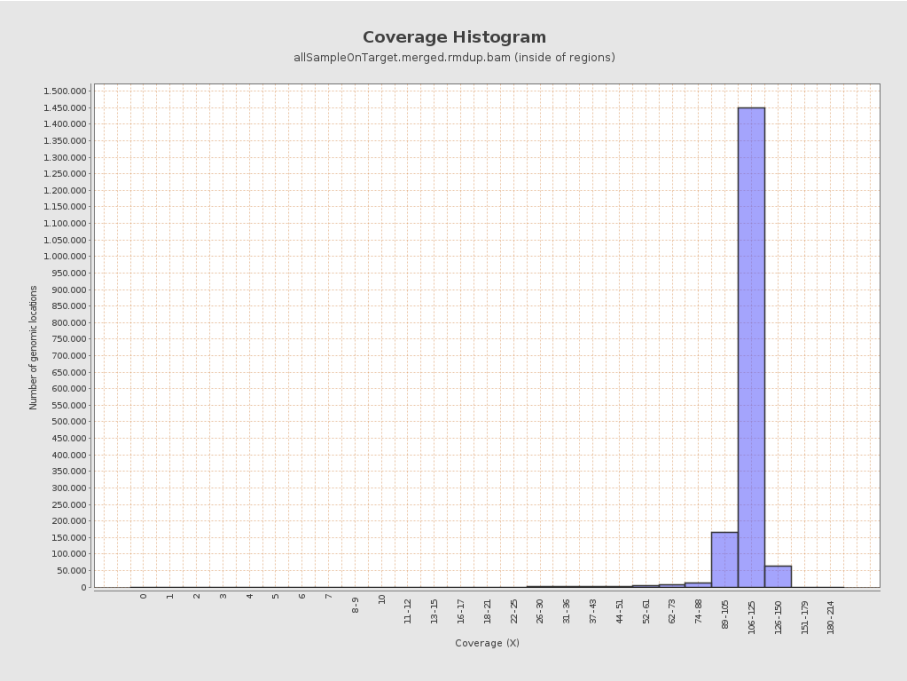

Figure 4-2. Coverage histogram within the sequenced ROIs for all IPF samples.

As an example of the exhaustive QC required by a NGS experiment, some quality parameters corresponding to technical replicates of paired-end analysis for an IPF patient are shown in Table 4-2.

Table 4-2. Quality control results as observed with SeqPrep in replicates of paired-end analysis for an IPF patient.

Input sequence files	L02_R1_01	L02_R2_01	L04_R1_01	L04_R2_01
Number of reads	2,931,172	2,931,172	3,597,809	3,597,809
Percent GC	43	43	43	44
Percent duplicates (sequence identity in the first 50 bp of the reads)	13.65	12.66	16.54	16.24

Input sequence files	L02_R1_01	L02_R2_01	L04_R1_01	L04_R2_01
Sequence quality at the 36th position	38.0	36.8	38.9	37.7
Number of reads left unmerged by SeqPrep	1,136,788	1,136,788	2,024,372	2,024,372
Number of reads merged	1,790,721	1,790,721	1,558,492	1,558,492
Fraction merged	0.611	0.611	0.433	0.433

The quality of the genotyping process is of crucial relevance in NGS experiments. Quality scores, expressed in Phred scale (a logarithm function of base-calling error probabilities) versus position in read are shown in Figure 4-3. As indicated, Phred scores were larger than 30 (less than 1 error in 1,000 bases) for all ROIs.

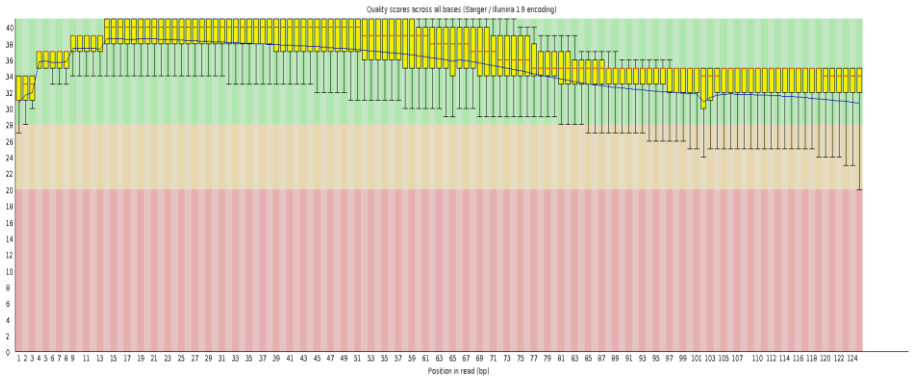

Figure 4-3. Quality scores across all bases as observed with FastQC in a replicate paired-end analysis for an IPF patient.

Similarly, the average Phred score was roughly 38 (Figure 4-4), indicating that the probability of an incorrect base call was less than 1 in 10,000 on average (>99.99% base call accuracy).

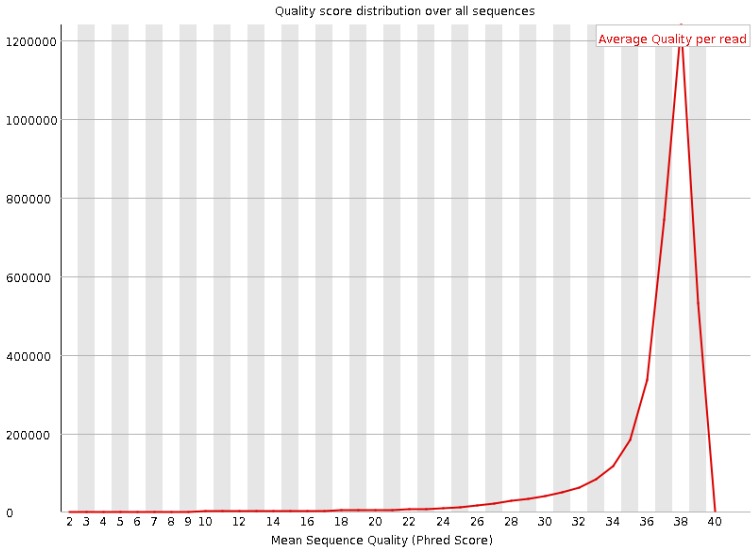

Figure 4-4. Quality scores distribution over all sequences as observed with FastQC in a replicate paired-end analysis of an IPF patient.

In relation to the quality of mapping, despite the left tail of the distribution of the number of genomic locations, most of reads presented mapping scores around 70 (i.e. less than one wrong alignment expected in 10 million alignments) (Figure 4-5).

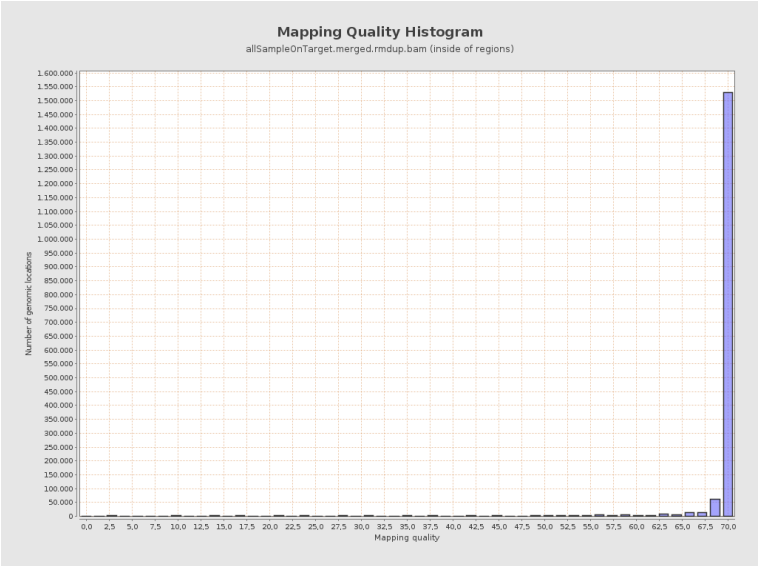

Figure 4-5. Mapping quality histogram for all IPF samples.

4.1.1. MAF and genotype consensus between sequencing platforms: NGS versus genome-wide genotyping

Recent scientific literature highlights that experimental errors in NGS are more frequent than expected. The reported base call accuracy for leading NGS technologies varies largely, from one error in one thousand nucleotides (99.9%) [60] to one error in ten million nucleotides (99.9999%) [61]. Therefore, artifacts (false variants) may be erroneously considered as de novo, rare or somatic putative variants [62].

Genome-wide (GW) SNP data (421,814 SNPs after QC procedures) for a total of 115 IPF individuals out of the 192 that were sequenced for this work were available. To measure the concordance of variants measured by using these two distinct technologies —the Affymetrix 6.0(Affymetrix, Santa Clara, CA) chip and the Illumina HiSeq2500 (Illumina) Sequencer— we computed the intersection of SNPs measured with both methodologies. We found a total of 231 SNPs common to both data sets. Minor allele frequencies in the two technologies showed a very high linear correlation (Pearson correlation, R^2=0.998) (Figure 4-6).

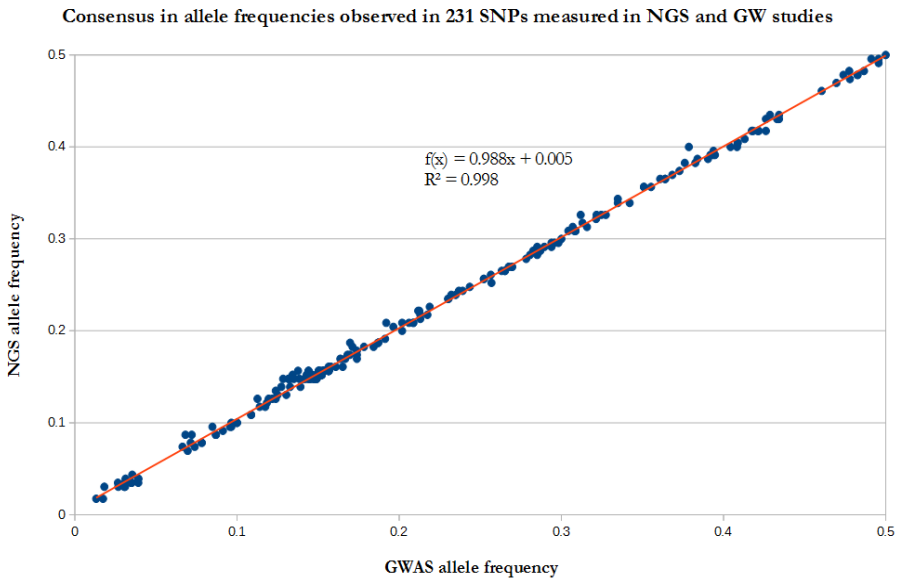

Figure 4-6. Minor allele frequency in NGS and GW SNP data for the overlapping set of 231 SNPs in 115 IPF samples with data from the two technologies.

In addition, a very high concordance was also observed for genotypes measured by both technologies in the same individuals. PLINK 1.07 provides a fast merge mode to

check the level of concordance between these sets. From a total of 26,565 overlapping data, 26,244 had a present call in both datasets, and 25,229 were concordant. Thus, the observed concordance rate was estimated in 96.1% (95% CI: 95.9% - 96.4%).

4.1.2. Consensus between two haplotype-based variant callers: GATK versus Platypus

We used PLINK and VCFtools to obtain concordance metrics for called variants from GATK and Platypus. GATK was able to identify a total of 16,253 variant sites, whereas Platypus identified 15,937. By overlapping sites (13,652 common sites), that is considering variants identified within the same genomic position or locus by the two callers or common sites, we found a concordance of 84.0% (95% CI: 83.4%-84.6%) and 85.7% (95% CI: 85.1%-86.2%) for GATK and Platypus, respectively (Figure 4-7). If we only consider the matching overlapping sites (12,522), this concordance was estimated in 77.0% (95% CI: 76.4%-77.7%) and 78.6% for GATK and Platypus, respectively (Figure 4-8).

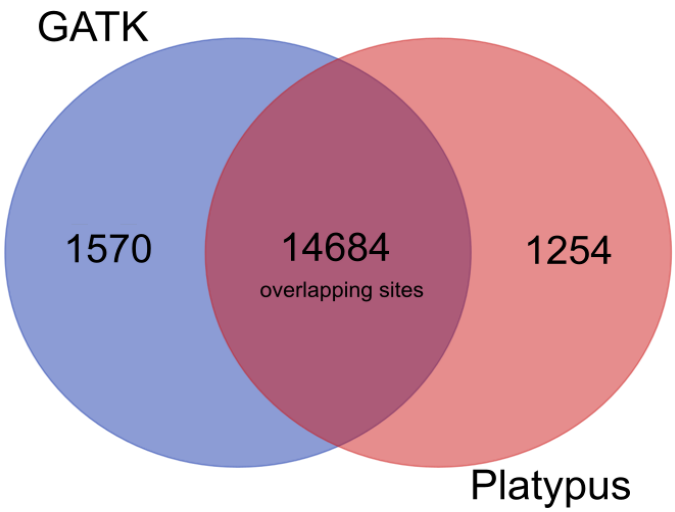

Figure 4-7. Concordance between GATK and Platypus callers per individual overlapping sites. The sum of overlapping sites and sites found only by GATK or Platypus yields the total number of identified variants with each caller (16,253 and 15,937, respectively).

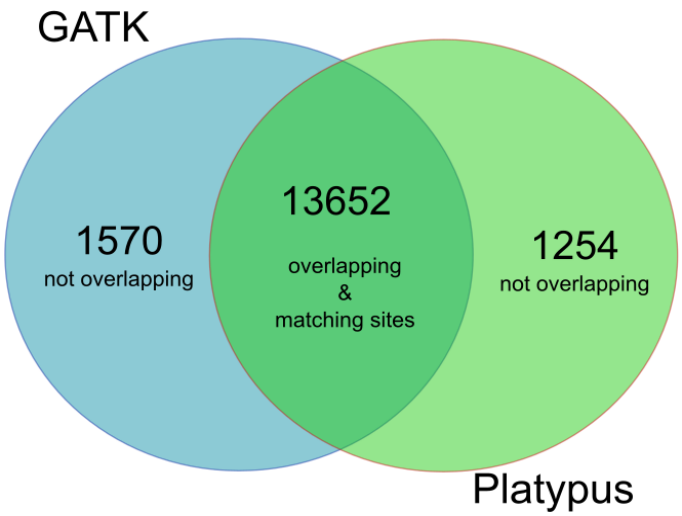

Figure 4-8. Concordance between GATK and Platypus callers per individual overlapping-matching sites. Taking into account the number of overlapping (14,684) and matching sites (13,652), and sites found only by GATK (1,570) or Platypus (1,254), we can derive the total number of identified variants with each caller (16,253 and 15,937, respectively).

Among exonic variants, up to 70% were missense (70.74% and 69.52% for GATK and Platypus, respectively). In addition, silent variants were estimated in 27% (26.26% and 28.25% for GATK and Platypus, respectively), while a very low number of nonsense variants was called (3.00% and 2.23% for GATK and Platypus, respectively). The consensus median according to the number of effects by function was high (101.8%) between both callers. However, we must take into account that between 7.8% and 9.7% of called variants are exclusive to GATK or Platypus (Figure 4-8).

The number of called variants per type is shown in Table 4-3. The number of SNPs identified by GATK is larger than those identified by Platypus. On the contrary, Platypus showed a higher number of insertions and deletions compared to GATK. Upon simplification of the multiple nucleotide polymorphisms (MNPs) provided by Platypus into more basic/primitive alleles by means of the GATK VariantsToAllelicPrimitives walker, a higher consensus is reached (omitted here because it is beyond the scope of this master work). The median consensus between GATK and Platypus per type of called variants was estimated in 90.6%.

Table 4-3. Number of called variants per type.

Type	Platypus	GATK	Consensus % (GATK as reference)
SNP	12,617	13,932	90.6
MNP	658	0	---
INS	1,847	2,200	84.0
DEL	2,310	2,102	109.9
Total	17,432	18,234	90.6

SNP=Single Nucleotide Polymorphism; MNP=Multiple Nucleotide Polymorphism; INS=Insertion; DEL=Deletion.

SnpEff provides different types of functional annotation. The number of called variants affecting certain genome elements is presented in Table 4-4. The most frequent variants identified by both callers were INTRONIC, UPSTREAM/DOWSTREAM, INTERGENIC, UTR and EXONIC types (in that order).

Table 4-4. Percentage of called variants affecting a certain genome element.

Affected genome element	Platypus	GATK	Consensus % (GATK as reference)
CODON related variations	0.135	0.2	67.5
SPLICE related variations	0.032	0.019	168.4
EXONIC	1.707	1.67	102.2
INTRONIC	44.88	43.97	102.1
INTERGENIC	10.58	10.52	100.6
NON_SYNONYMOUS_CODING	0.87	0.816	106.6
SYNONYMOUS_CODING	0.615	0.637	96.5
START/STOP related variations	0.065	0.078	83.3
FRAME_SHIFT related variations	0.566	0.523	108.2
UTR related variations	1.95	1.951	99.9
UPSTREAM/DOWSTREAM	38.61	39.62	97.4
		median	102.3

Again, the concordance between the variants called by GATK and Platypus are within ±10% if we analyze the number of effects by region or affecting a certain genome element, with the exception of variants annotated as CODON, SPLICE and START/STOP. Given the functional relevance of variants in these gene elements, these results are concerning for the clinical application of NGS technologies. In fact, O'Rawe et al. [63] carried out a very wide comparative of concordances between genotyping platforms, different aligners and variant-calling pipelines (SOAP, BWA-GATK, BWA-SNVer, GNUMAP, and BWA-SAMtools) using exomes and single whole genomes. Due to the large variations in the studied concordances, they suggested that more caution should be exercised in genomic medicine settings when analyzing individual genomes, including interpreting positive and negative findings with scrutiny, especially for indels.

The ratio of transitions (pyrimidine-pyrimidine or purine-purine changes) to transversions (pyrimidine-purine or purine-pyrimidine changes), the so-called Ti/Tv ratio, can be used as another quality metric of human NGS data. The Ti/Tv for human whole-genome sequence data was estimated in the range of 2.0 to 2.2 [18,64], while a higher ratio (~3.0) is expected in exomes due to the presence of methylated cytosine in CpG dinucleotides in exonic regions [65]. The bias in favor of mutations between bases of similar chemical properties (transitions) over those with dissimilar properties (transversions) is independent on both CpG and the GC content of the genome. Therefore, the Ti/Tv may be used as a useful diagnostic tool to measure the quality of the NGS data generated [19,66]. GATK and Platypus provided Ti/Tv ratios in the range of those expected (2.192 and 2.234, respectively) for whole genomes. This result is not unexpected as a large fraction of the ROIs are non-exonic sequences.

4.2. Association study results

Liu et al. [67] performed a comparative analysis with distinct callers (SAMtools, GATK, glftools, and Atlas2) using single-sample and multiple-sample variant-calling strategies on whole exomes. They concluded that GATK had the highest rediscovery rate (0.9969) and specificity (0.99996), and its Ti/Tv ratio was closest to the expected value of 3.02. In addition, variant genotypes called by exome sequencing versus exome arrays were more accurate, although the average variant sensitivity and overall genotype consistency rate were as high as 95.87% and 99.82%. Such an extensive comparative analysis has yet to be conducted for Platypus. Based on this evidence, downstream association studies were performed considering only the variant calls generated by GATK.

Once programmed and tested the bioinformatics pipeline using GATK provided a VCF file with 16,253 variant loci for the whole set of cases. After combining the data from cases (192) and unrelated controls (501), a total of 61,374 variants were observed. After filtering out of indels, a total of 57,696 biallelic sites were observed. These variant loci were further processed following the workflow depicted in Figure 3-8. A summary of remaining SNVs available for downstream analysis is presented in Table 4-5.

Table 4-5. Summary of SNVs observed in cases and controls.

SNV set	Sites	Pruning LD r^2<0.15
After merging of data from cases and controls	61,374	---
Biallelic SNVs	57,696	---
Biallelic SNVs with call rate >95%	10,245	2,342

4.2.1. Allelic frequencies and missing data

We used PLINK 1.07 to obtain allelic frequencies and missing data summaries jointly for cases and controls. The average number of individuals missing a certain genotype was 0.175%, with a maximum of 34 individuals. After identifying SNVs with an excessive missing rate (>5%), we kept 10,245 out of 57,696 variants and all individuals (cases and controls) for downstream analysis.

4.2.2. Ancestry estimation in the IPF patients

As explained in section 3.7.2, we prepared a double strategy to study the ancestry of samples. Due to the small size of the ROIs, there was insufficient sequence overlapping between the CEPH-HGDP reference panel (which contains 938 individuals and 632,958 markers) and the IPF NGS data. Therefore, LASER could not be applied successfully in this study. An alternative approach was prepared by gathering 2,504 individuals as a set of population references from 1KGP corresponding to 26 different populations from many different locations around the globe. These populations have been genetically divided into 5 biogeographical population groups: AFR, African; AMR, admixed American; EAS, East Asian; EUR, European; and SAS, South Asian.

We extracted the selected ROIs from those individuals and merged them with cases. As a result, a set of 2,696 individuals (2,504 reference individuals / 192 cases) and a total of 62,528 variants were obtained. We discarded variants with a calling rate lower than 95%, and from these filtered SNVs, we kept only those with a LD r^2<0.15 in order to remove the LD structure. Finally, a total of 1,876 SNPs were used for PCA (2,696 individuals x 1,876 SNPs) using R and the snpRelate package.

A plot of the first two eigenvectors or main principal components derived from the filtered set of 1,876 SNPs (Figure 4-9, left panel) showed clusters of individuals matching biogeographical population groups (AFR, AMR, EAS, EUR, and SAS). The IPF individuals (depicted as orange circles) are clustered with European (EUR) and American (AMR) individuals, supporting that the IPF cases are, in fact, of European-American ancestry (Figure 4-9, right panel). This result is in agreement with the careful selection of IPF individuals followed by Noth et al. [8].

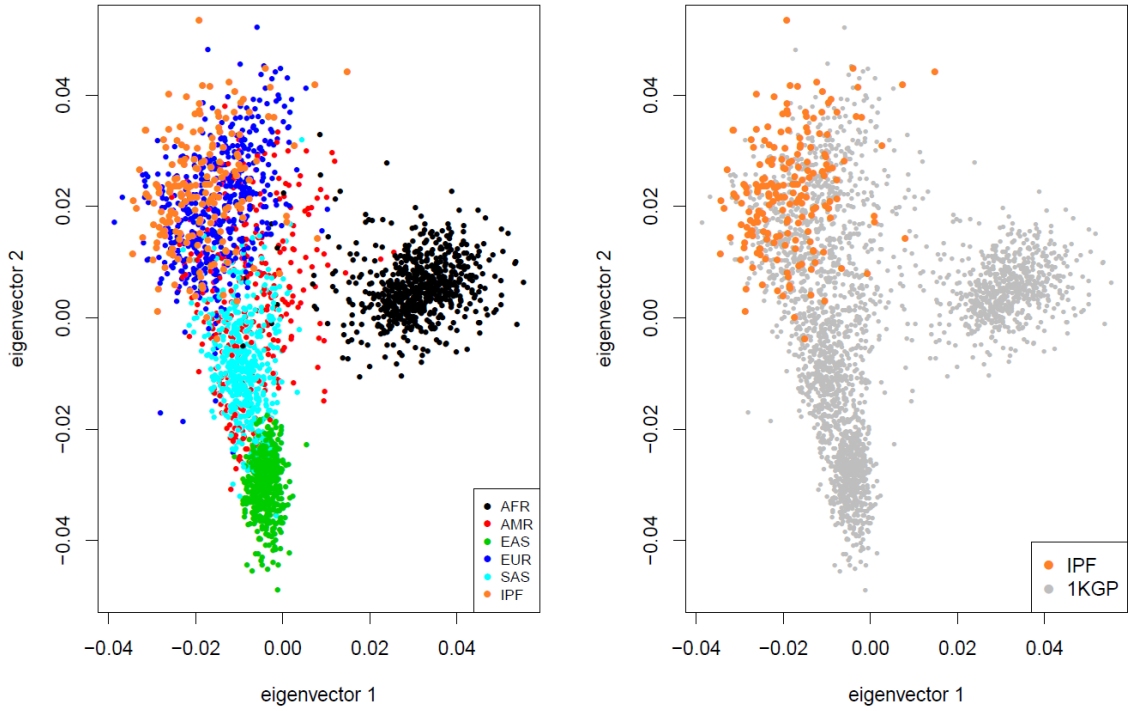

Figure 4-9. Plot of the first two principal components for IPF cases and 1KGP individuals from the five biogeographical population groups. IPF individuals are indicated with orange circles and cluster with European and admixed American population groups. AFR=African; AMR=Ad Mixed American; EAS=East Asian; EUR=European; SAS=South Asian.

4.2.3. Population stratification

Using the filtered subset of 10,245 SNVs, we deduced a principal component space formed by a pruned set of variants (with r^2<0.15) by means of Eigensoft. Eigensoft was also used to analyze the dependence of inflation of the statistic test with the number of principal components. A QQ-plot was prepared to compare the observed distribution of the association test (χ^2) versus the expected quantiles after adjusting by the selected number of principal components (Figure 4-10). We observed that five principal components were enough to account for the population stratification present in the association study, as they provided a negligible residual inflation of association results (λ=1.00).

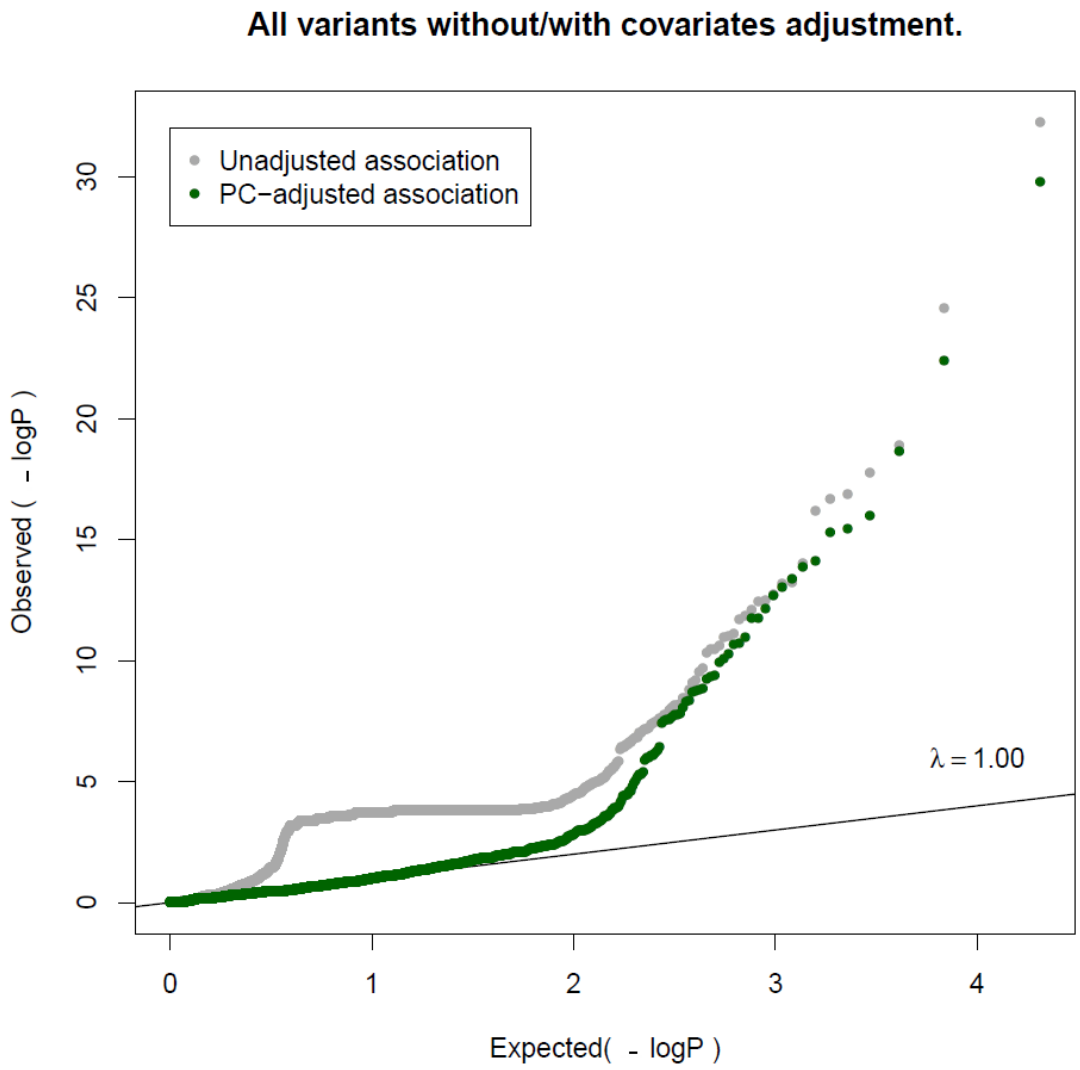

Figure 4-10. QQ-plot representing the IPF association results for the 10,245 SNPs. Grey circles represent the inflated distribution of the statistic due to population stratification. Green circles represent the deflated distribution of the statistic after adjusting for five principal components.

Once association results were obtained, we first compared the findings with those reached in the GWAS by Noth et al. [8]. In stage two of that study, six SNPs located in three loci were significantly associated at genome-wide significance: three *TOLLIP* SNPs (rs111521887, rs5743894, rs5743890) and one *MUC5B* SNP (rs35705950) at 11p15.5; one *MDGA2* SNP (rs7144383) at 14q21.3; and one *SPPL2C* SNP (rs17690703) at 17q21.31. For that, regional plots of association results, generated by the on-line tool *LocusZoom* [68], were centered on rs35705950 in chromosome 11, rs7144383 in chromosome 14, and rs17690703 in chromosome 17 (Figure 4-11). *LocusZoom* plots the minus log-10 of the p-

values resulting from the association tests (y-axis) versus the chromosome position, together with the recombination rate (in cM/Mb), pairwise LD values (r^2) between SNPs in European populations from 1KGP, and proximal genes (x-axis).

In addition, we annotated each of the 10,245 tested SNVs according to their MAF in controls for plotting. We classified the variant as 'rare' (represented by squares) when the control MAF was < 5% and 'frequent' when MAF ≥ 5% (depicted in circles).

This study replicated previous findings at 11p15.5, where a variant in the promoter region of the gel-forming *MUC5B* gene (rs35705950) constitutes the most significant hit. However, while several other SNPs reached genome-wide significance in the region (details below), none of the three *TOLLIP* SNPs reported as risks by Noth et al. [8] (rs111521887, rs5743894, rs5743890) were significant in this study.

As for the 14q21.3 and 17q21.31 loci, none of the two top hits reported by Noth et al. [8] were nominally significant in this study (rs7144383, p=0.181; rs17690703, p=0.639). SNP rs4898572, another intronic variant in strong LD with rs7144383 in *MDGA2 gene* as reported by Noth et al. [8], was also not significant in this study (p=0.191). These discrepancies may be explained by the small sample size and the weak reported effects for these SNPs, as many other nearby variants in the two loci were detected at genome-wide significance (details below).

Further validation studies are being conducted to be able to discern if newly discovered risk variants are real polymorphisms or artifacts. In particular, the region in chromosome 17 shows a complex structure [69], where we found frequent and rare variants that we may consider as putative risk variants at this stage. This region spans 440 kb partially or entirely involving five genes (*CRHR1*, *IMP5*/*SPPL2C*, *MAPT*, *STH*, and *KANSL1*) and harbors a high number of copy number variants (CNVs). To make it more complex, the evidence supports the existence of a large inversion that has been positively selected in Europeans [38]. This structural complexity may have affected the genotype calling by the utilized algorithms. Therefore, further work will be necessary to empirically validate the robustness of key variants associated in this region.

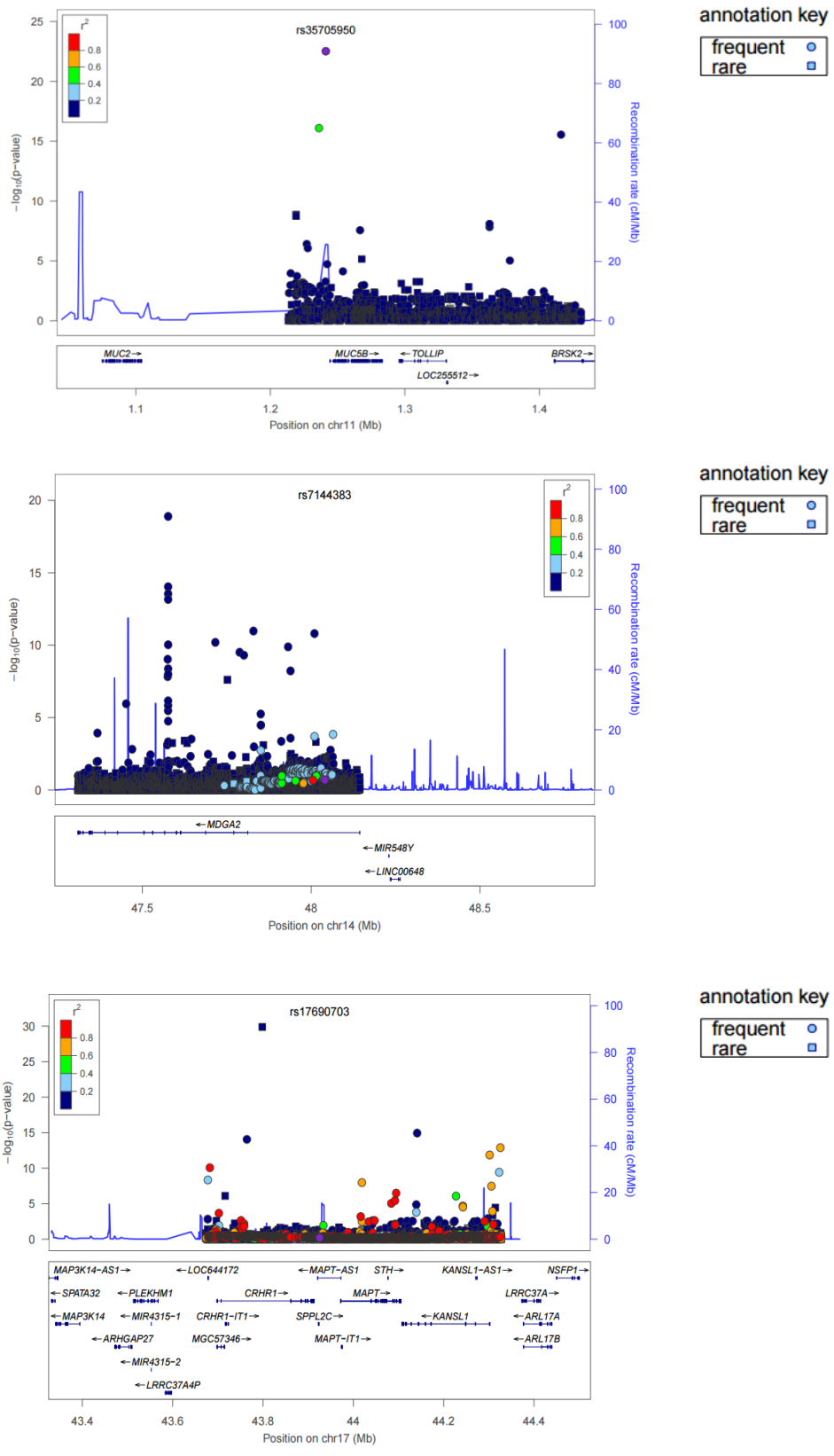

Figure 4-11. Regional association plots in chromosomes 11 (top), 14 (middle) and 17 (bottom), centered in previously reported top significant SNPs by Noth et al. [8] (depicted as purple circles).

4.2.4. Top associated SNPs

The logistic regression of the binary trait (phenotype case/control) with five principal components as covariates identified 38 top significant SNPs: nine in chromosome 11, seventeen in chromosome 14, and twelve in chromosome 17 (Table 4-6). Significance is defined here as the probability of a test statistic being equal to or greater than the observed test statistic if the null hypothesis of no association is true. Thus lower p-values show that in case of no association, the chance of seeing this result is extremely low. A common approach to define a threshold to declare statistical significance is based on the concept of genome-wide significance. For European-descent populations, this threshold has been estimated at $7.2x10^{-8}$ [70]. A more recent accepted standard is $5x10^{-8}$ [71], which is the threshold utilized in this association study. A summary of associated variants reaching genome-wide significance broken by whether they are described for the first time or not is shown in Table 4-6. As previously indicated, the variant with the highest significance is the SNP in the promoter region of *MUC5B* (rs35705950; $p=3.97x10^{-23}$), related with a slightly larger effect compared to that reported (OR= 6.12, 95% CI: 4.28-8.76). Although we did not test whether this effect is statistically distinct from that reported by previous GWAS, studies with small sample sizes such as that of this master's, are expected to translate into biased OR estimates.

Table 4-6. Summary results for the 38 genome-wide significant SNPs associated with IPF.

chr	Annotated	No annotation	Total
11	9	0	9
14	16	1	17
17	8	4	12
			38

Setting a threshold of 5% in controls to declare a variant as frequent/common (MAF>5%) or infrequent/rare (MAF<5%), we found that three of the top SNPs on chromosome 11 are rare (but are quite frequent in IPF cases). Similarly, for chromosomes 14 and 17, only one SNP on each region was classified as rare among the significant results.

4.2.5. Further evaluations of top associated SNPs

As explained in the Methodology Section, the GATK VQSR walker provides a continuous estimate of the probability that each variant is true, allowing one to partition the call sets into quality tranches when GATK operates in SNP-mode. We have used four thresholds of sensitivity relative to the truth sets: 90, 99, 99.9, and 100%. These tranches are applied to the GATK output VCF file using the FILTER field. In this way, we can choose to keep all or only some filtered records (Figure 4-12).

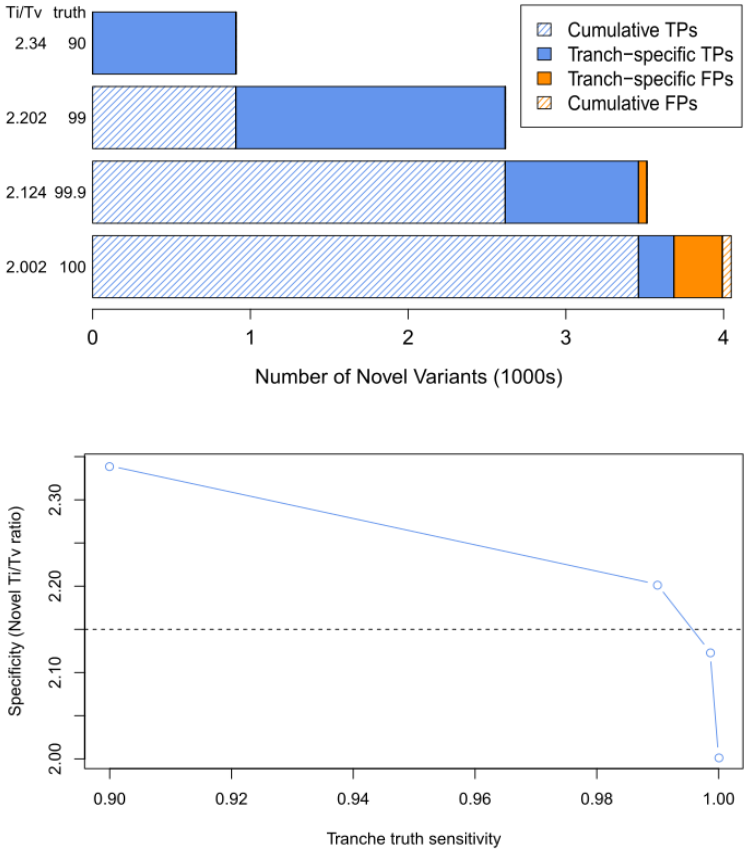

Figure 4-12. Variant Quality Score Recalibration for NGS IPF called variants using GATK: tranches plot (top) and specificity versus tranche truth sensitivity (bottom).

In the tranches plot, the x-axis represents the number of novel variants called while the y-axis shows two quality metrics: novel transition to transversion ratio (Ti/Tv) and the overall truth sensitivity.

The GATK VariantRecalibrator walker was used setting to 4 the maximum number of different Gaussian-clusters of variants in the identification process (*--maxGaussians 4*).

From the lowest tranche (90) to the highest (100), we move from low sensitivity /high specificity to high sensitivity/low specificity. In other words, each subsequent tranche in turn introduces additional true positive calls along with a growing number of false positive calls.

We have computed the significance of HWE deviations in cases by means of '*HardyWeinberg*' R library for each of the top identified SNPs (Figure 4-13). While deviations from HWE expectations may be found among patients because of the disease status, large deviations in the context of complex diseases may also suggest issues related to the quality of genotyping data. Out of the 38 top significant SNPs, 11 showed a large deviation from HWE expectations in cases after a Bonferroni adjustment (computed as 0.05 / 10,245 SNPs = 4.88×10^{-6}).

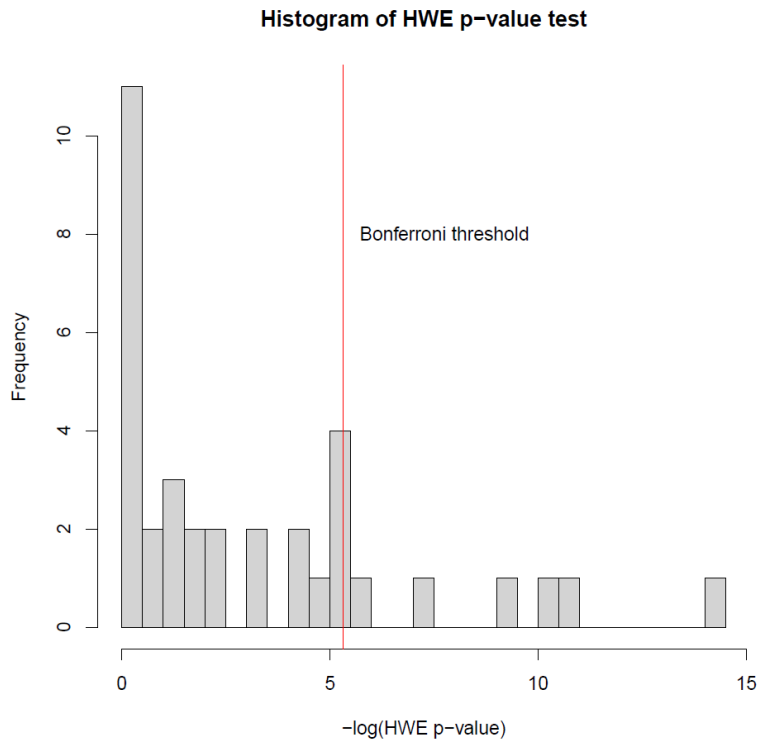

Figure 4-13. Histogram of HWE tests in cases for the 38 top significant SNPs. The vertical red line depicts the Bonferroni threshold (4.88×10^{-6}) to declare a HWE p-value as a sign of departure from the HWE.

In addition, we have revised the FILTER field in the VCF file provided by GATK to study the correspondence between the statistical significance of the association and the tranche of each finding (results not shown since they are under current evaluation).

A total of 12 out of 38 SNPs passed the GATK VQSR procedure with high confidence of being true variants in the NGS data set. A total of 19 out of 38 SNPs were found in the tranche between 99.90 and 100.00 thresholds, where a relatively high number of true negatives SNPs are expected. The rest of SNPs (n=7) were found in the tranche between 99.00 and 99.99 thresholds. The SNP in the promoter region of *MUC5B* (rs35705950) was identified by both the GATK and Platypus callers (Figure 4-14).

In this exploration, we focused on one particular SNP, rs371630624, to illustrate the inherent difficulties of current genotype calling algorithms for NGS data, particularly for rare variants. This SNP is a rare variant that was found strongly associated with the disease in this study ($p=7.37x10^{-13}$; OR=1,822, 95% CI: 234.2-14,170). However, the lack of genotype counts corresponding to the expected rare homozygous (homozygous for the alternative allele or "1/1" genotype; Figure 4-14) while the population of heterozygous is large, has a consequent violation of the HWE expectations (HWE $p=9.88x10^{-15}$), highlighting plausible issues that artifact the results for this SNP [63,72]. Moreover, while all the reference allele homozygous patients were supported by >80% of reads for that allele, none of the heterozygous patients were supported by >20% of reads for both alleles (average proportion of alternative allele reads was 12.4%). In comparison, genotypes for variant rs35705950 passed the GATK Variant Quality Score Recalibrator with a '*PASS*' value in its filter, while rs371630624 showed a '*VQSRTrancheSNP99.90to100.00*' warning in its corresponding filter. This means that the variant was in the range of VQSLODs (the probability that each call is real) corresponding to the remaining 0.1% of the training set, which is considered as a false positive. Thus, a further inspection of this putatively associated variant was required. Furthermore, rs371630624 was not identified by Platypus.

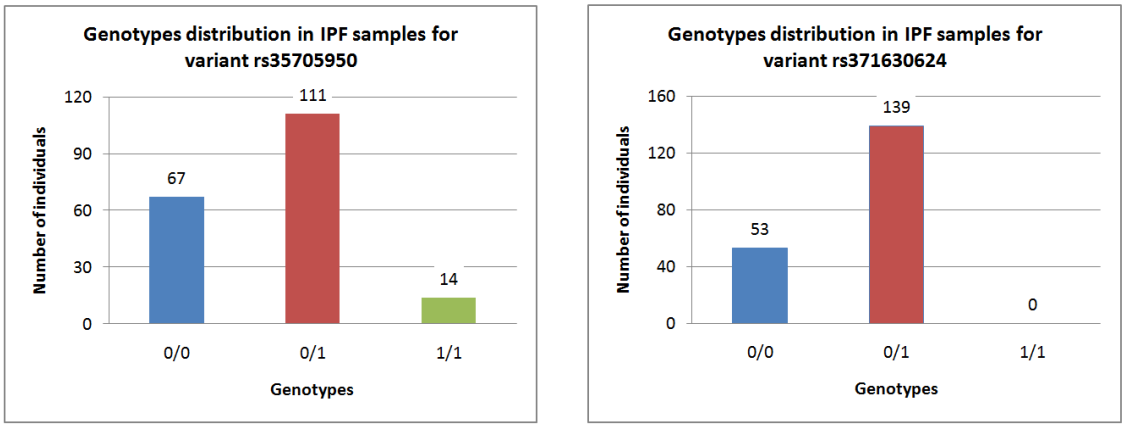

Figure 4-14. Genotypes distribution of variant rs35705950 (left) and rs371630624 (right) in IPF individuals.

When focusing of the most significant SNP from each of the three regions, each showed a relatively high number of reads and a high Phred score base quality within ±50 bp flanking the variant position, but a variable mean base quality mapping (Table 4-7).

Table 4-7. Quality features of top associated SNPs per chromosome.

Top SNP at chr.	Mean number of reads	Mean Mapping Quality (MAPQ)	Mean Phred score (QUAL)
11 (rs35705950)	104	70.0	38.8
14	41.7	49.9	40.6
17	226	70.0	39.7

Interestingly, both GATK and Platypus called these three SNPs, although with variable concordances for the genotypes called (Table 4-8).

Table 4-8. Genotype concordance of top SNPs at each region between GATK and Platypus calls.

Top SNP in chr. 11 (rs35705950)	Genotypes (number of individuals)			
Caller	**0/0**	**0/1**	**1/1**	**./.**
GATK	67	111	14	0
Platypus	67	111	14	0

Top SNP in chr. 14	Genotypes (number of individuals)			
Caller	**0/0**	**0/1**	**1/1**	**./.**
GATK	141	27	10	14
Platypus	121	49	4	18

Top SNP in chr. 17	Genotypes (number of individuals)			
Caller	**0/0**	**0/1**	**1/1**	**./.**
GATK	137	51	4	0
Platypus	137	51	4	0

There was a perfect match in the called genotypes for variants rs35705950 and the top SNP of chromosome 17 (100% concordance in both cases), whereas the concordance between GATK and Platypus was as low as 75% for the top variant in chromosome 14. GATK and Platypus called this variant but warned about the calling confidence setting their filters to "VQSRTrancheSNP99.90to100.00" and "alleleBias", respectively.

4.2.6. Functional analysis of top associated SNPs

Linking associations between phenotypes and genotypes and their impact on health is challenging, both in terms of improved or new therapies and reliable prognosis. A key step is to search for the mechanisms linking the presence of a variant to an altered *in vivo* gene product function [73].

A SNV may disturb the function of a gene product through a wide spectrum of mechanisms, including transcription factor binding, miRNA interactions, messenger RNA splicing, structure and half-life, translation efficiency, and non-synonymous substitution

effects [73]. The challenge of the interpretation is even greater if we consider non-coding variants, given the diversity of non-coding functions, the incomplete annotation of regulatory elements and the potential existence of still unknown mechanisms of gene regulation [74].

A number of strategies are currently used to perform a functional impact analysis of gene variants. Among them include reference functional genomics, chromatin state maps, nucleotide-resolution regulatory annotations, predictive models of variants effects, comparative genomics between related species, and evolutionary conserved biochemical activity [74].

To help in the interpretation and prioritization of variants, several online resources are available to researchers, including HaploReg [75], RegulomeDB [76], ENSEMBL's SNP Effect Predictor [77]. Variant annotators such as snpEff [21] and ANNOVAR [22] allowed us to use this knowledge from different curated and non-curated databases to automatically provide the annotation of variants. This information was further supplemented with empirical data generated by the ENCODE project [78] as reported by HaploReg v3 and RegulomeDB. In addition, conserved regions that exhibit evidence of selective constraint were identified by GERP [79] and SiPhy [80] scores.

Among top significant SNPs, seven were identified as coding non-synonymous (five of them as part of *MUC5B* gene, and two as part of *MUC2*), three variants were identified as promoter variants (one in chromosome 11 and two in chromosome 17) and 12 were annotated as enhancer variants (nine in chromosome 14 and three in chromosome 17) according to current ENCODE data. Most of these regulatory variants (at promoters or enhancers) were also associated with DNAse I hypersensitivity sites and/or with the alteration of motifs. One of the promoter variants identified was, in fact, rs35705950 in *MUC5B,* which associates with histone marks in 9 organs, DNAse I hypersensitivity sites in 5 organs, binding of 11 proteins at this locus, and 4 altered motifs because of the allelic changes. This variant was one of two that showed consistent evidence of phylogenetic conservation both in GERP and SiPhy (a second one was an intronic variant in chromosome 17). It is present in Europeans and Americans with a MAF of 8% and 10%, respectively. We have observed that this variant is present in 36.2% of IPF cases, whereas controls showed a MAF of 10.7%, similar to the MAF declared for European Americans. None of the top significant SNPs was related with eQTLs or protein changes. Interestingly,

no functional information was found for the rare variant rs371630624, again supporting that this putative novel variant is, in fact, an artifact of the variant calling.

4.3. The bioinformatics pipeline computing times

We have implemented a bioinformatics pipeline available in a cluster server based on the GATK and auxiliary tools (see **Appendix A1** for a comprehensive list of used software). Since the use of different queues with a different number of processors and cores is possible at the cluster server, it is not easy to provide an exact computing time for each of the tasks comprising the GATK pipeline. However, we provide an estimation of the computing time in each of the basic steps, with an approximate time of 220 hours (**Appendix A4**). During the execution time, about 2,170 temporal and definitive files are produced, occupying 130 GB of extra data. Taking into account the original sample file sizes of about 200 GB, a total of 330 GB hard drive storage was required.

5. CONCLUSIONS

1) We have developed a GATK-based bioinformatics pipeline to process and manage targeted Next Generation Sequencing data to extract germline-variants. This pipeline has been extended to study the association of called variants with susceptibility to Idiopathic Pulmonary Fibrosis.
2) We have used this bioinformatics pipeline to perform an association study of genetic variants with susceptibility to Idiopathic Pulmonary Fibrosis, by comparing the data obtained from regions of interest in chromosomes 11, 14 and 17 between 192 patients and 501 unrelated European subjects from The 1000 Genomes Project.
3) A total of 38 SNPs reached genome-wide significance, five of them described for the first time in this study. The previously described risk SNP at the promoter of *MUC5B* showed the top significance in the region of chromosome 11. On the contrary, in regions from chromosomes 14 and 17, the study revealed novel variants with stronger significance than those SNPs identified by the previous GWAS.

6. APPENDICES

6.1. Appendix A1. List of software used to design and test the bioinformatics pipeline

Name	Version	Core	Description
ANNOVAR		Perl	Annotate called variants, a variant annotator. URL: http://annovar.openbioinformatics.org/
BEDtools	2.17.0	C++	A Swiss-army knife of tools for a wide-range of genomics analysis tasks: intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. URL: http://bedtools.readthedocs.org/
Eigensoft	6.0.1	C/Perl	The EIGENSOFT package combines functionality from our population genetics methods (Patterson et al. 2006) and our EIGENSTRAT stratification correction method (Price et al. 2006). URL: http://data.broadinstitute.org/alkesgroup/EIGENSOFT/
FastQC	0.11.2	Java	FastQC is a QC application for high throughput sequence data. It reads in sequence data in a variety of formats and can either provide an interactive application to review the results of several different QC checks, or create an HTML based report, which can be integrated into a pipeline. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
fcGENE	1.0.7	C++	It converts genotype SNP data into formats of different imputation tools like PLINK MACH, IMPUTE, BEAGLE and BIMBBAM, second to transform imputed data into different file formats like PLINK, HAPLOVIEW, EIGENSOFT and SNPTEST. URL: http://sourceforge.net/projects/fcgene/
Filezilla	3.13	C	A FTP/SFTP client for remote file management. URL: https://filezilla-project.org/
GATK	3.3	Java	Genome Analysis Tool Kit, a variant caller. URL: https://www.broadinstitute.org/gatk/
HaploReg	3	Online	HaploReg is a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. URL: http://www.broadinstitute.org/mammals/haploreg/

Name	Version	Core	Description
IGV	2.3.34	Java	It is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations. URL: https://www.broadinstitute.org/igv/
NovoAlign	3.02.00	C/Perl/R	A mapper for NGS reads to a reference database. It is an aligner for single-ended and paired-end reads from the Illumina Genome Analyser. URL: http://www.novocraft.com/products/novoalign/
LocusZoom		Online	LocusZoom is a tool to plot regional association results from genome-wide association scans or candidate gene studies. URL: http://locuszoom.sph.umich.edu/locuszoom/
Platypus	0.8.1	Python/C	A variant caller designed for efficient and accurate variant-detection in high-throughput sequencing data. By using local realignment of reads and local assembly, it achieves both high sensitivity and high specificity. URL: http://www.well.ox.ac.uk/platypus/
Picard	1.119	Java	A set of Java command line tools for manipulating high-throughput sequencing data (HTS) data and formats. URL: http://broadinstitute.github.io/picard/
PLINK	1.07	C/C++	It is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner. URL: http://pngu.mgh.harvard.edu/~purcell/plink/
PLINK2	1.9	C/C++	It is a complete rewrite of the original code and represents a very significant improvement in overall speed and functionality. URL: https://www.cog-genomics.org/plink2
Putty	0.65	C	A client program for the SSH, Telnet and Rlogin network protocols. URL: http://www.chiark.greenend.org.uk/~sgtatham/putty/
Qualimap	2.1.	Java/R	Qualimap 2 is a platform-independent application that provides both a Graphical User Interface (GUI) and a command-line interface to facilitate the QC of alignment sequencing data and its derivatives like feature counts. URL: http://qualimap.bioinfo.cipf.es/

Name	Version	Core	Description
R	3.2.1.	C IDE (RStudio)	R is a free software environment for statistical computing and graphics. URL: https://www.r-project.org/ Many libraries and dependent-packages are required by GATK, Qualimap, etc. Some of them are: • Available from CRAN: ggplot2, gsalib, optparse, HardyWeinberg. • Available from Bioconductor: NOISeq, Repitools, Rsamtools, GenomicFeatures, rtracklayer, snpRelate.
RegulomeDB		Online	RegulomeDB is a database that annotates SNPs with known and predicted regulatory elements in the intergenic regions of the H. sapiens genome. URL: http://www.regulomedb.org/
SAMtools	1.2	C	SAMtools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format. URL: http://samtools.sourceforge.net/
SepPrep	0.4	C	A program to merge paired end Illumina reads that are overlapping into a single longer read. URL: https://github.com/jstjohn/SeqPrep
snpEff	4_1h	Java/Perl/ Python	A genetic variant annotation and effect prediction toolbox. It annotates and predicts the effects of variants on genes (such as amino acid changes). Features: URL: http://snpeff.sourceforge.net/
VCFtools	0.1.12b	Perl/C	VCFtools is a program package designed for working with VCF files, such as those generated by The 1000 Genomes Project. The aim of VCFtools is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files. URL: https://vcftools.github.io/
WinSCP	5.7.5	C	WinSCP is an open source free SFTP client, FTP client, WebDAV client and SCP client for Windows. URL: https://winscp.net

6.2. Appendix A2. Overview of SBATCH/BASH scripts running on DRAGO cluster server

Note: use "#" for running SBATCH commands on DRAGO (code shown in blue color) and "###" for comments (code shown in green color). If DRAGO is not used, BASH commands works as expected (no "#" is required; code shown in black color).

```
#!/bin/sh

### Script name: script-name.sh

### Target: pipeline for IPF NGS DNA-seq using DRAGO cluster at drago.saii.ull.es

### SBATCH and SLURM management
### Send job to a queue: > sbatch task1.sh
### Query the status of a queue: > squeue
### Cancel the job: scancel <job_id>
### Query the status of a job: scontrol show job <job_id>
### See available queues: sinfo

### Capture a log file with console results
### bash ---.sh &> ---.log


### Asigns name to a job (default name: PBS)
#SBATCH -J ...

### Select cluster nodes and cores
#SBATCH-N 1 # 1 node for DRAGO
#SBATCH-n 20 # 1 cores for DRAGO

###Available SLURM queues are: "sequential, test, fast, medium y batch"
### Select SLURM queue
### sequential: sequential applications (168 h/1 core); test: small tests (5 min./9 cores); fast: rapid executions (30 h/30 cores); medium: long-term runs (168 h/40 cores); batch: high demanding jobs (12 h/80 cores; default queue).

### Use 'test' queue
#SBATCH -p test

### Or use 'fast' queue
#SBATCH -p fast

### Or use 'batch' queue
#BATCH -p batch
```

```
### Or use 'medium' queue
##SBATCH -p medium

### Stores the standard log output
#SBATCH -o script-name.out

### Stores the standard err output
#SBATCH -e script-name.err

### Do not repeat job in case of failure
#script-name.sh -r n

### Exports environment variables
#script-name.sh -V

### Options to send email: BEGIN, END, FAIL, ALL
#SBATCH --mail-type=BEGIN
#SBATCH --mail-user=username@domain

#SBATCH --mail-type=END
#SBATCH --mail-user= username@domain

#SBATCH --mail-type=FAIL
#SBATCH --mail-user= username@domain

### Loads profile
source /etc/profile

### Load specific modules in DRAGO memory
### Loads Java 1.7 to run GATK
module add java/sun1.7
### Loads R 3.1.2 (make sure that ggplot2, gplots, reshape, grid, tools, and gsalib libraries are already installed)
module add R/3.1.2

### Paths to files. The script check a dummy config file where a variable 'machine' identifies the type of machine to be used for running this pipeline (1=cluster server; 2=desktop machine; 3=laptop machine)
machine=$(awk 'NR==1 {print $1}' machine/machine)

### Check in which machine the pipeline will be run
if [ "$machine" = "1" ]; then
echo "We work on DRAGO cluster..."
ROOT=/username/a-folder
ROOT2=/username/another-folder
GATK=$ROOT/apps/GATK-3.3-0/GenomeAnalysisTK.jar
fi

if [ "$machine" = "2" ]; then
```

```
echo "We work on a Desktop machine..."
ROOT=/home/username/a-folder
ROOT2=/username/another-folder
GATK=$ROOT/apps/GATK-3.3-0/GenomeAnalysisTK.jar
fi

if [ "$machine" = "3" ]; then
echo "We work on a laptop machine..."
ROOT=/home/username/a-folder
ROOT2=/username/another-folder
GATK=$ROOT/apps/GATK-3.3-0/GenomeAnalysisTK.jar
fi

### Set full paths to reference genome and auxiliary databases
REFERENCE=$ROOT2/hg19/ucsc.hg19.fasta
GATK_BUNDLE=$ROOT2/gatk_bundle


### Auxiliary files for GATK routines:
INDELS=$GATK_BUNDLE/Mills_and_1000G_gold_standard.indels.hg19.vcf
DBSNP=$GATK_BUNDLE/dbsnp_138.hg19.vcf
HAPMAP=$GATK_BUNDLE/hapmap_3.3.hg19.vcf
OMNI=$GATK_BUNDLE/1000G_omni2.5.hg19.vcf

### Input path for sample files and text file with a list of sample filenames
SAMPLES=$ROOT/scripts/samples/samples_task---
INPUT=$ROOT2/samples

### Output path for results
OUTPUT=$ROOT2/outputs/task---
CONTROL=$OUTPUT/task---.log

echo "============================================="
echo "= PIPELINE TO CALL VARIANTS FROM DNA-seq data using GATK ="
echo "=cflores-lab / HUNSC 2015="
echo "============================================="

### DO NOT DELETE THIS LINE: FOR-LOOP is built based on the sample folder
### when iteration over the files contained in a defined folder is required
cd $SAMPLESFOLDER

### Example of a GATK task
### TASK01: Compute Indel realignment for each file ($file stores current file name) ---
> prepares the list of targets

### For each IPF DNA-seq sample
while read file
do
echo "----------------------------"
```

```
echo "TASK01: Compute Indel realignment for each file ($file stores current file name) ---> prepares the list of targets"
echo "----------------------------"
echo ">>> Starting a new iteration..."
echo ">>>>>> Processing file: " $file

### NOTE: check whether actual version of the GATK does support "-nt xx" for multiple parallelism in this command

java -jar $GATK -T RealignerTargetCreator -R $REFERENCE -I $INPUT/$file -known $INDELS -o $OUTPUT/$file.target_intervals.list

### ADD more GATK commands here
### java -jar $GATK walker 2 <options><infile><outfile>
### java -jar $GATK walker 3 <options><infile><outfile>
### java -jar $GATK walker ...

### Appends filename of processed sample to output file if iteration declared in the FOR-LOOP is required (if not, comment this line)
echo $file >> $CONTROL

### End of LOOP for TASK, if required (if not, comment this line)
done< $SAMPLES
```

6.3. Appendix A3. Excerpts of GATK and related commands

```
#!/bin/sh

#Task1: GATK RealignerTargetCreator
java -jar $GATK -T RealignerTargetCreator -R $REFERENCE -I $INPUT/$file -known
$INDELS -o $OUTPUT/$file.target_intervals.list
---------------------------------------------------------------------------------------------------------

#Task2: GATK IndelRealigner
java -jar $GATK -T IndelRealigner -R $REFERENCE -I $INPUT/$file -targetIntervals
$ROOT$WD/outputs_task1/$file.target_intervals.list -known $INDELS -o
$OUTPUT/$file.realigned_reads.bam
---------------------------------------------------------------------------------------------------------

#Task3: GATK BaseRecalibrator
java -jar $GATK -T BaseRecalibrator -R $REFERENCE -I
$INPUT/$file.realigned_reads.bam -knownSites $DBSNP -knownSites $INDELS -o
$OUTPUT/$file.recal_data.table
---------------------------------------------------------------------------------------------------------

#Task4: GATK BaseRecalibrator
java -jar $GATK -T BaseRecalibrator -R $REFERENCE -I
$ROOT$WD/outputs_task2/$file.realigned_reads.bam -knownSites $DBSNP -
knownSites $INDELS -BQSR $ROOT$WD/outputs_task3/$file.recal_data.table -o
$OUTPUT/$file.recal_data_post.table
---------------------------------------------------------------------------------------------------------

#Task5: AnalyzeCovariates
java -jar $GATK -T AnalyzeCovariates -R $REFERENCE -before
$INPUT1/$file.recal_data.table -after $INPUT2/$file.recal_data_post.table -plots
$OUTPUT/$file.recalibration_plots.pdf -csv my-report.csv
---------------------------------------------------------------------------------------------------------

#Task6: PrintReads
java -jar $GATK -T PrintReads -R $REFERENCE -I $INPUT/$file.realigned_reads.bam -
BQSR $INPUT2/$file.recal_data.table -o $OUTPUT/$file.recal_reads.bam
---------------------------------------------------------------------------------------------------------

#Task7: GATK HapotypeCaller
java -jar $GATK -T HaplotypeCaller -R $REFERENCE -I $INPUT/$file.recal_reads.bam
--emitRefConfidence GVCF --variant_index_type LINEAR --variant_index_parameter
128000 --dbsnp $DBSNP -L chr11 -L chr14 -L chr17 -o
$OUTPUT/$file.raw.snps.indels.g.vcf
---------------------------------------------------------------------------------------------------------

#Task8: GATK CombineGCVFs
#BATCH 1
```

java -jar $GATK/GenomeAnalysisTK.jar -T CombineGVCFs -R $REFERENCE --variant $INPUT/NGS_1_DNA_Targeted.merged.onTarget.bam.rmdup.bam.raw.snps.indels.g.vcf --variant $INPUT/NGS_2_DNA_Targeted.merged.onTarget.bam.rmdup.bam.raw.snps.indels.g.vcf --variant $INPUT/NGS_3_DNA_Targeted.merged.onTarget.bam.rmdup.bam.raw.snps.indels.g.vcf --variant $INPUT/NGS_4
... etc ... and repeat for each batch of 50 files

Then combine all small cohorts into onle single gVCF file
java -jar $GATK/GenomeAnalysisTK.jar -T CombineGVCFs -R $REFERENCE --variant $INPUT/ipf_cohort1.g.vcf --variant $INPUT/ipf_cohort2.g.vcf --variant $INPUT/ipf_cohort3.g.vcf --variant $INPUT/ipf_cohort4.g.vcf --variant $INPUT/ipf_cohort5.g.vcf -o $OUTPUT/ipf_cohort_12345.g.vcf
---

#Task9: GATK GenotypeGVCFs
java -jar $GATK/GenomeAnalysisTK.jar -T GenotypeGVCFs -R $REFERENCE --variant $INPUT/ipf_cohort.g.vcf --max_alternate_alleles 64 -o $OUTPUT/ipf_cohort_genotyped.vcf
---

#Task10: GATK VariantRecalibrator
java -jar $GATK -T VariantRecalibrator -R $REFERENCE -input $INPUT/ipf_cohort_genotyped.vcf -resource:hapmap,known=false,training=true,truth=true,prior=15.0 $HAPMAP -resource:omni,known=false,training=true,truth=false,prior=12.0 $OMNI -resource:dbsnp,known=true,training=false,truth=false,prior=6.0 $DBSNP -an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an ReadPosRankSum -an InbreedingCoeff -mode SNP -L chr11 -L chr14 -L chr17 --maxGaussians 4 -recalFile $OUTPUT/ipf_cohort_genotyped.vcf.recalibrate_SNP.recal -tranchesFile $OUTPUT/ipf_cohort_genotyped.vcf.recalibrate_SNP.tranches -rscriptFile $OUTPUT/ipf_cohort_genotyped.vcf.recalibrate_SNP_plots.R

#Task11: GATK ApplyRecalibration
java -jar $GATK -T ApplyRecalibration -R $REFERENCE -input $INPUT1/ipf_cohort_genotyped.vcf -mode SNP --ts_filter_level 99.0 -recalFile $INPUT2/ipf_cohort_genotyped.vcf.SNP.recal -tranchesFile $INPUT2/ipf_cohort_genotyped.vcf.SNP.tranches -o $OUTPUT/recalibrated_snps_raw_indels.vcf
---

#Task12: GATK VariantRecalibrator
java -jar $GATK -T VariantRecalibrator -R $REFERENCE -input $INPUT/recalibrated_snps_raw_indels.vcf -resource:mills,known=true,training=true,truth=true,prior=12.0 $MILLS -an QD -an DP -an FS -an SOR -an MQRankSum -an ReadPosRankSum -an InbreedingCoeff -mode

INDEL -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 -L chr11 -L chr14 -L chr17 --maxGaussians 4 -recalFile $OUTPUT/ipf_cohort_genotyped.vcf.recalibrate_INDEL.recal -tranchesFile $OUTPUT/ipf_cohort_genotyped.vcf.recalibrate_INDEL.tranches -rscriptFile $OUTPUT/ipf_cohort_genotyped.vcf.recalibrate_INDEL_plots.R

#Task13: GATK ApplyRecalibration

java -jar $GATK -T ApplyRecalibration -R $REFERENCE -input $INPUT1/recalibrated_snps_raw_indels.vcf -mode INDEL --ts_filter_level 99.0 -recalFile $INPUT2/ipf_cohort_genotyped.vcf.recalibrate_INDEL.recal -tranchesFile $INPUT2/ipf_cohort_genotyped.vcf.recalibrate_INDEL.tranches -o $OUTPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.vcf

---

#Task14: snpEff annotation

hg19 as reference

java -Xmx4g -jar $SNPEFF/snpEff.jar hg19 -v -noLog $INPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.vcf> $OUTPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.snpEff_ann_hg19.vcf

GRCh38.79 as reference

java -Xmx4g -jar $SNPEFF/snpEff.jar GRCh38.79 -v -noLog $INPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.vcf> $OUTPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.snpEff_ann.vcf

---

#Task15: GATK VariantAnnotator

java -jar $GATK -R $REFERENCE -T VariantAnnotator -A SnpEff -V $INPUT1/ipf_cohort_genotyped.vcf.recalibrated_variants.vcf -snpEffFile $INPUT2/ipf_cohort_genotyped.vcf.recalibrated_variants.snpEff_ann_GRCh38.79.vcf -L chr11 -L chr14 -L chr17 -o $OUTPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.snpEff_ann_GRCh38.79.GATK_VR.vcf --dbsnp $DBSNP

#Task16: ANNOVAR annotation

$ANNOVAR/table_annovar.pl $INPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.vcf $ANNOVAR/humandb/ -buildver hg19 -out $OUTPUT/ipf_cohort_genotyped.vcf.recalibrated_variants.vcf_bufferSize10000_annovar -remove -protocol refGene,cytoBand,genomicSuperDups,esp6500siv2_all,1000g2014oct_all,1000g2014oct_afr,1000g2014oct_eas,1000g2014oct_eur,snp138,ljb26_all -operation g,r,r,f,f,f,f,f,f,f -nastring . -vcfinput

---

#Task17: BEDtools file perations

```
$BEDTOOLS/bin/multiIntersectBed -i $INPUT/*.bed >
$OUTPUT/target_intervals.bed -header
```

#Task18: BEDtools and VCFtools file operations

```
$VCFTOOLS/vcftools --gzvcf $INPUT1 --chr 11 --from-bp 1212759 --to-bp 1431096 --
recode --recode-INFO-all --out $OUTPUT/ALL1KGP_chr11.genotypes
gzip -9 -k $OUTPUT/ALL1KGP_chr11.genotypes.recode.vcf
$VCFTOOLS/vcftools --gzvcf $INPUT2 --chr 14 --from-bp 47308642 --to-bp 48144657 -
-recode --recode-INFO-all --out $OUTPUT/ALL1KGP_chr14.genotypes
gzip -9 -k $OUTPUT/ALL1KGP_chr14.genotypes.recode.vcf
$VCFTOOLS/vcftools --gzvcf $INPUT3 --chr 17 --from-bp 43672512 --to-bp 44836908 -
-recode --recode-INFO-all --out $OUTPUT/ALL1KGP_chr17.genotypes
gzip -9 -k $OUTPUT/ALL1KGP_chr17.genotypes.recode.vcf
```

#Task19: VCFtools and command line file operations

```
(zcat $INPUT1 | head -250 | grep ^#; zcat $INPUT1 | grep -v ^#; zcat $INPUT2 | grep
-v ^#; zcat $INPUT3 | grep -v ^# ) | gzip -c > $OUTPUT/ALL1KGP_chr-11-14-
17.genotypes.vcf.gz

grep CEU $GENOME/integrated_call_samples_v3.20130502.ALL.panel | cut -f1 >
$OUTPUT/CEU.samples.list
grep GBR $GENOME/integrated_call_samples_v3.20130502.ALL.panel | cut -f1 >
$OUTPUT/GBR.samples.list
grep FIN $GENOME/integrated_call_samples_v3.20130502.ALL.panel | cut -f1 >
$OUTPUT/FIN.samples.list
grep IBS $GENOME/integrated_call_samples_v3.20130502.ALL.panel | cut -f1 >
$OUTPUT/IBS.samples.list
grep TSI $GENOME/integrated_call_samples_v3.20130502.ALL.panel | cut -f1 >
$OUTPUT/TSI.samples.list
cat $OUTPUT/*.samples.list >
$OUTPUT/1KGP_phase3_CEU_FIN_GRB_IBS_TSI_individuals.txt

$VCFTOOLS/vcftools --gzvcf $INPUT/ALL1KGP_chr-11-14-
17.genotypes.controls.vcf.gz --out $OUTPUT/ALL1KGP_chr-11-14-
17.genotypes.controls.vcf.gz --plink
```

#Task20: GATK CombineVariants

```
java -jar $GATK -T CombineVariants -R $REFERENCE --variant cases.vcf --variant
controls.vcf -o cases_controls.vcf -genotypeMergeOptions UNIQUIFY
```

6.4. Appendix A4. Approximate computing times and number of generated files for each of the GATK and non-GATK steps integrated in the bioinformatics pipeline

# Task	GATK walker / other tools	Estimated computing time	Output file type	# files
1	RealignerTargetCreator	3 d	target_intervals.list	192
2	IndelRealigner	6 h	realigned_reads.bam	384
3	BaseRecalibrator	18 h	recal_data.table	192
4	BaseRecalibrator	16 h	recal_data_post.table	192
5	AnalyzeCovariates	~ 1.5 h	recalibration_plots.pdf	192
6	PrintReads	13 h	recal_reads.bam	384
7	HaplotypeCaller	3 d	raw.snps.indels.g.vcf	384
8	CombineGVCFs	< 1 h	cohort.g.vcf	2
9	GenotypeGVCFs	< 1 h	cohort_genotyped.vcf	2
10	VariantRecalibrator	< 1h	cohort_genotyped.vcf.recalibrate_SNP.recal cohort_genotyped.vcf.recalibrate_SNP.tranches cohort_genotyped.vcf.recalibrate_SNP _plots.R.pdf	6
11	ApplyRecalibration	< 1h	recalibrated_snps_raw_indels.vcf	2
12	VariantRecalibrator	< 1h	cohort_genotyped.vcf.recalibrate_indels.recal cohort_genotyped.vcf.recalibrate_indels.tranches cohort_genotyped.vcf.recalibrate_indels _plots.R.pdf	5
13	ApplyRecalibration	< 1h	cohort_genotyped.vcf.recalibrated_variants.vcf	2
14	snpEff	< 1h	Annotated VCF file (with snpEff)	6
15	VariantAnnotator	< 1h	Annotated.VCF.snpEff_GRCh38.79.GATK_VR	4
16	ANNOVAR	< 1h	Annotated VCF file (with ANNOVAR)	5

# Task	GATK walker / other tools	Estimated computing time	Output file type	# files
17	BEDtools	~ 2 h	bam.rmdup.bam.bed	197
18	BEDtools VCFtools		ALL1KGP_chr11.genotypes.recode.vcf ALL1KGP_chr14.genotypes.recode.vcf ALL1KGP_chr17.genotypes.recode.vcf	9
19	VCFtools	< 1 h	ALL1KGP_chr-11-14-17.5pops.genotypes.vcf ALL1KGP_chr-11-14-17.26pops.genotypes.vcf	4
20	CombineVariants VCFtools	< 1 h	Cases and controls VCF	6

7. REFERENCES

[1] Gross T.J. and Hunninghake G.W. *Idiopathic pulmonary fibrosis*. N Engl J Med. August 2001; 345(7): 517-525. doi: 10.1097/MAJ.0b013e31821a9d8e.

[2] Hambly N, Shimbori C, Kolb M. *Molecular classification of idiopathic pulmonary fibrosis: Personalized medicine, genetics and biomarkers*. Respirology. 2015 Oct;20(7):1010-1022 doi: 10.1111/resp.12569.

[3] Ley B, Collard HR, King TE Jr. *Clinical course and prediction of survival in idiopathic pulmonary fibrosis*. Am. J. Respir. Crit.Care Med. 2011; 183: 431–40.

[4] Elisabeth Bendstrup, Athol U. Wells. *AIR 2014 – The Year in IPF*. Sarcoidosis Vasculitis and Diffuse Lung Diseases 2015; 32; Suppl. 1: 3.

[5] Fingerlin T. E. et al. *Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis*. Nat Genet. 2013 June; 45(6): 613–620. doi:10.1038/ng.2609.

[6] Peljto A.L. et al. *Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis*. JAMA. 2013; 309(21): 2232-2239. doi: 10.1001/jama.2013.5827.

[7] Seibold M.A. et al. *A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis*, N Engl J Med. April 2011; 364(16): 1503-1512. doi: 10.1056/NEJMoa1013660.

[8] Noth I. et al. *Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study*. Lancet Respir Med. 2013 June; 1(4): 309–317. doi: 10.1016/S2213-2600(13)70045-6.

[9] Mathai SK, Schwartz DA. *Taking the "I" out of IPF*. Eur Respir J. 2015 Jun;45(6):1539-41. doi: 10.1183/09031936.00052715.

[10] Zhang Y., Noth I., Garcia J. G. N., and Kamiski N. *A Variant in the Promoter of MUC5B and Idiopathic Pulmonary Fibrosis*. N Engl J Med. April 21; 364(16): 1576–1577. doi: 10.1056/NEJMc1013504.

[11] Stock CJ, et al..*Mucin 5B promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or sarcoidosis*. Thorax. 2013;68:436–41. doi: 10.1136/thoraxjnl-2012-201786.

[12] Elisabetta Renzoni, Veeraraghavan Srihari, and Piersante Sestini. *Pathogenesis of idiopathic pulmonary fibrosis: review of recent findings*. F1000Prime Rep. 2014; 6: 69. doi: 10.12703/P6-69.

[13] Yang IV et al. *Expression of cilium-associated genes defines novel molecular subtypes of idiopathic pulmonary fibrosis*. Thorax. 2013;68:1114–21. doi: 10.1136/thoraxjnl-2012-202943.

[14] Coghlan MA et al. *Sequencing of idiopathic pulmonary fibrosis-related genes reveals independent single gene associations*. BMJ Open Respir Res. 2014 Dec 10;1(1):e000057. doi: 10.1136/bmjresp-2014-000057.

[15] Ganesh Raghu et al. *An Official ATS/ERS/JRS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-based Guidelines for Diagnosis and Management*, American Journal of Respiratory and Critical Care Medicine, Vol. 183, No. 6 (2011), pp. 788-824. doi: 10.1164/rccm.2009-040GL

[16] Fernando García-Alcalde, Konstantin Okonechnikov, José Carbonell, Luis M. Cruz, Stefan Götz, Sonia Tarazona, Joaquín Dopazo, Thomas F. Meyer, and Ana Conesa. *Qualimap: evaluating next-generation sequencing alignment data*. Bioinformatics 28, no. 20 (2012): 2678-2679. doi: 10.1093/bioinformatics/bts503.

[17] McKenna A, et al. *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110.

[18] DePristo MA et al. *A framework for variation discovery and genotyping using next-generation DNA sequencing data*. Nat Genet. 2011 May;43(5):491-8. doi: 10.1038/ng.806.

[19] Van der Auwera GA et al. *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*. 2013 Current Protocols in Bioinformatics 43:11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.

[20] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, WGS500 Consortium, Andrew O M Wilkie, Gil McVean & Gerton Lunter. *Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications*. Nature Genetics 46, 912–918 (2014) doi: 10.1038/ng.3036.

[21] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3*. 2012 Apr-Jun;6(2):80-92. doi: 10.4161/fly.19695.

[22] Kai Wang, Mingyao Li, and Hakon Hakonarson. *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data*. Nucleic Acids Res. 2010 Sep; 38(16): e164. doi: 10.1093/nar/gkq603.

[23] Chaolong Wang et al. *Ancestry Estimation and Control of Population Stratification for Sequence-based Association Studies*. Nature Genetics 46, 409–415 (2014). doi:10.1038/ng.2924.

[24] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. *The Sequence alignment/map (SAM) format and SAMtools*. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.

[25] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group. *The Variant Call Format and VCFtools*. Bioinformatics. 2011 Aug 1; 27(15): 2156–2158. doi: 10.1093/bioinformatics/btr330.

[26] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007). *PLINK: a toolset for whole-genome association and population-based linkage analysis*. American Journal of Human Genetics, 81. doi: 10.1086/519795.

[27] Aron R. Quinlan and Ira M. Hall. *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics (2010) 26 (6): 841-842. doi: 10.1093/bioinformatics/btq033.

[28] Nab Raj Roshyara, and Markus Scholz. *fcGENE: A Versatile Tool for Processing and Transforming SNP Datasets*. PLoS One. 2014; 9(7): e97589. doi: 10.1371/journal.pone.0097589.

[29] Nick Patterson, Alkes L Price, and David Reich. *Population Structure and Eigenanalysis*. PLoS Genet. 2006 Dec; 2(12): e190. doi: 10.1371/journal.pgen.0020190.

[30] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick & David Reich. *Principal components analysis corrects for stratification in genome-wide association studies*. Nature Genetics 38, 904 - 909 (2006). doi: 10.1038/ng1847.

[31] Qi Y, Liu X, Liu CG, Wang B, Hess KR, Symmans WF, Shi W, Pusztai L. *Reproducibility of Variant Calls in Replicate Next Generation Sequencing Experiment.* PLoS One. 2015 Jul 2;10(7):e0119230. doi: 10.1371/journal.pone.0119230. eCollection 2015.

[32] Altmann A, Weber P, Bader D, Preuss M, Binder EB, Müller-Myhsok B. *A beginners guide to SNP calling from high-throughput DNA-sequencing data*. Hum Genet. 2012 Oct;131(10):1541-54. doi: 10.1007/s00439-012-1213-z.

[33] Li H, Durbin R. *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics. 2009;25:1754–60. doi: 10.1093/bioinformatics/btp324.

[34] Li H, Ruan J, Durbin R. *Mapping short DNA sequencing reads and calling variants using mapping quality scores*. Genome Res. 2008 Nov;18(11):1851-8. doi: 10.1101/gr.078212.108.

[35] Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. *SNP detection for massively parallel whole-genome resequencing*. Genome Res. 2009 Jun;19(6):1124-32. doi: 10.1101/gr.088013.108.

[36] Ratan A, Olson TL, Loughran TP Jr, Miller W. *Identification of indels in next-generation sequencing data. BMC Bioinformatics*. 2015 Feb 13;16:42. doi: 10.1186/s12859-015-0483-6.

[37] Wang, J., Scofield, D., Street, N. & Ingvarsson, P. (2015). *Variant calling using NGS data in European aspen (Populus tremula)* (1ed.). In: Sablok, G., Kumar, S., Ueno, S., Kuo, J., Varotto, C. (Eds.) (Ed.), Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches: Springer International Publishing Switzerland: Springer.

[38] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. *An integrated map of genetic variation from 1,092 human genomes*. Nature. 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632.

[39] Michael E. Weale. *Quality Control for Genome-Wide Association Studies*. Chapter 19, in Genetic Variation, Methods in Molecular Biology Volume 628, 2010, pp 341-372.

[40] Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., et al. *Population structure, differential bias and genomic control in a large-scale, case-control association study*. Nat Genet. 2005 Nov;37(11):1243-6.

[41] Silverberg MS, et al. *Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study*. Nat Genet. 2009;41:216. doi: 10.1038/ng.275.

[42] Fisher SA, et al. *Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease*. Nat Genet. 2008;40:710. doi: 10.1038/ng.145.

[43] Wellcome Trust Case Control Consortium. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*. Nature. 2007 Jun 7;447(7145):661-78.

[44] Morris AP, Zeggini E. *An evaluation of statistical approaches to rare variant analysis in genetic association studies*. Genet Epidemiol. 2010;34:188. doi: 10.1002/gepi.20450.

[45] Jian W., Sanjay S. *Testing Hardy-Weinberg Proportions in a Frequency-Matched Case-Control Genetic Association Study*. PLoS One. 2011; 6(11): e27642. doi: 10.1371/journal.pone.0027642.

[46] Balding DJ. *A tutorial on statistical methods for population association studies*. Nat Rev Genet. 2006 Oct;7(10):781-91.

[47] Salanti G, Sanderson S, Higgins JP. *Obstacles and opportunities in meta-analysis of genetic association studies*. Genet Med. 2005 Jan;7(1):13-20.

[48] Nielsen DM, Ehm MG, Weir BS. *Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus*. Am J Hum Genet. 1998 Nov;63(5):1531-40.

[49] Astle W, Balding DJ. *Population structure and cryptic relatedness in genetic association studies*. Statistical Science. 2009;24:451–471. doi: 10.1214/09-sts307.

[50] Carl A. Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P. Morris, and Krina T. Zondervan. *Data quality control in genetic case-control association studies*. Nat Protoc. 2010 Sep; 5(9): 1564–1573. doi: 10.1038/nprot.2010.116.

[51] Shriner D. *Investigating population stratification and admixture using eigenanalysis of dense genotypes*. Heredity (Edinb). 2011 Oct;107(5):413-20. doi: 10.1038/hdy.2011.26. Epub 2011 Mar 30.

[52] Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. *Basic statistical analysis in genetic case-control studies*. Nat Protoc. 2011 Feb;6(2):121-33. doi: 10.1038/nprot.2010.182.

[53] Tiwari, H.K., Barnholtz-Sloan, J., Wineinger, N., Padilla, M.A., Vaughan, L.K. and Allison, D.B. *Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles*. Hum Hered. 2008;66(2):67-86. doi: 10.1159/000119107.

[54] Setakis E, Stirnadel H, Balding DJ. *Logistic regression protects against population structure in genetic association studies*. Genome Res. 2006 Feb;16(2):290-6. Genome Res. 2006 Feb; 16(2): 290–296. doi: 10.1101/gr.4346306.

[55] Devlin B, and Roeder K. *Genomic control for association studies*. Biometrics. 1999 Dec;55(4):997-1004.

[56] Tian, C., Gregersen, P.K. and Seldin, M.F. *Accounting for ancestry: population substructure and genome-wide association studies*. Hum Mol Genet. 2008 Oct 15;17(R2):R143-50. doi: 10.1093/hmg/ddn268.

[57] Vittinghoff E, McCulloch CE, Glidden DV, and Shiboski SC. *Linear and Non-Linear Regression Methods in Epidemiology and Biostatistics*. Handbook of Statistics 01/2007; 27:148-186. doi: 10.1016/S0169-7161(07)27005-1.

[58] Chaolong Wang, Zachary A. Szpiech, James H. Degnan, Mattias Jakobsson, Trevor J. Pemberton, John A. Hardy, Andrew B. Singleton, and Noah A. Rosenberg. *Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis*. Stat Appl Genet Mol Biol. 2010 Jan 1; 9(1): Article 13. doi: 10.2202/1544-6115.1493.

[59] Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. *The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution*. Curr Protoc Bioinform. 2012;39:1.

[60] Kildegaard HF, Baycin-Hizal D, Lewis NE, Betenbaugh MJ. *The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology*. Curr Opin Biotechnol. 2013 Dec;24(6):1102-7. doi: 10.1016/j.copbio.2013.02.007.

[61] Furey TS. *ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions*. Nat Rev Genet. 2012 Dec;13(12):840-52. doi: 10.1038/nrg3306.

[62] Robasky K, Lewis NE, Church GM. *The role of replicates for error mitigation in next-generation sequencing*. Nat Rev Genet. 2014 Jan;15(1):56-62. doi: 10.1038/nrg3655.

[63] O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, Lyon GJ. *Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing*. Genome Med. 2013 Mar 27;5(3):28. doi: 10.1186/gm432.

[64] Ebersberger I, Metzler D, Schwarz C, Pääbo S. *Genomewide comparison of DNA sequences between humans and chimpanzees*. Am J Hum Genet. 2002 Jun;70(6):1490-7. Epub 2002 Apr 30.

[65] Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A, Brizuela L, McCombie WR, Wigler M, Hannon GJ, Hicks JB. *High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing*. Genome Res. 2009 Sep;19(9):1593-605. doi: 10.1101/gr.095190.109.

[66] Baes CF, Dolezal MA, Koltes JE, Bapst B, Fritz-Waters E, Jansen S, Flury C, Signer-Hasler H, Stricker C, Fernando R, Fries R, Moll J, Garrick DJ, Reecy JM, Gredler B. *Evaluation of variant identification methods for whole genome sequencing data in dairy cattle*. BMC Genomics. 2014 Nov 1;15:948. doi: 10.1186/1471-2164-15-948.

[67] Liu X, Han S, Wang Z, Gelernter J, Yang BZ. *Variant callers for next-generation sequencing data: a comparison study*. PLoS One. 2013 Sep 27;8(9):e75619. doi: 10.1371/journal.pone.0075619.

[68] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. *LocusZoom: regional visualization of genome-wide association scan results*. Bioinformatics. 2010 Sep 15;26(18):2336-7. doi: 10.1093/bioinformatics/btq419.

[69] Boettger L. M., Handsaker R E., Zody M. C., and McCarroll S. A. *Structural haplotypes and recent evolution of the human 17q21.31 region*. Nat Genet. 2012 Aug; 44(8): 881–885. doi: 10.1038/ng.2334.

[70] Dudbridge F, Gusnanto A. *Estimation of significance thresholds for genomewide association scans*. Genet Epidemiol. 2008 Apr;32(3):227-34. doi: 10.1002/gepi.20297.

[71] Pe'er I, Yelensky R, Altshuler D, Daly MJ. *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants*. Genet Epidemiol. 2008 May;32(4):381-5. doi: 10.1002/gepi.20303.

[72] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. *Genome-wide association studies for complex traits: consensus, uncertainty and challenges*. Nat Rev Genet. 2008 May;9(5):356-69. doi: 10.1038/nrg2344.

[73] Pal LR, Yu CH, Mount SM, Moult J. *Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease*. BMC Genomics. 2015;16 Suppl 8:S4. doi: 10.1186/1471-2164-16-S8-S4.

[74] Lucas D. Ward and Manolis Kellis. *Interpreting non-coding variation in complex disease genetics*. Nat Biotechnol. 2012 Nov; 30(11): 1095–1106. doi: 10.1038/nbt.2422.

[75] Lucas D. Ward, and Manolis Kellis. *HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants*. Nucl. Acids Res. (2012) 40 (D1): D930-D934. doi: 10.1093/nar/gkr917.

[76] Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. *Annotation of functional variation in personal genomes using RegulomeDB*. Genome Res. 2012 Sep;22(9):1790-7. doi: 10.1101/gr.137323.112.

[77] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. *Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics*. 2010 Aug 15;26(16):2069-70. doi: 10.1093/bioinformatics/btq330.

[78] ENCODE Project Consortium. *An integrated encyclopedia of DNA elements in the human genome*. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.

[79] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. *Identifying a high fraction of the human genome to be under selective constraint using GERP++*. PLoS Comput Biol. 2010 Dec 2;6(12):e1001025. doi: 10.1371/journal.pcbi.1001025.

[80] Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. *Identifying novel constrained elements by exploiting biased substitution patterns*. Bioinformatics. 2009 Jun 15;25(12):i54-62. doi: 10.1093/bioinformatics/btp190.