Variant calling in NGS experiments

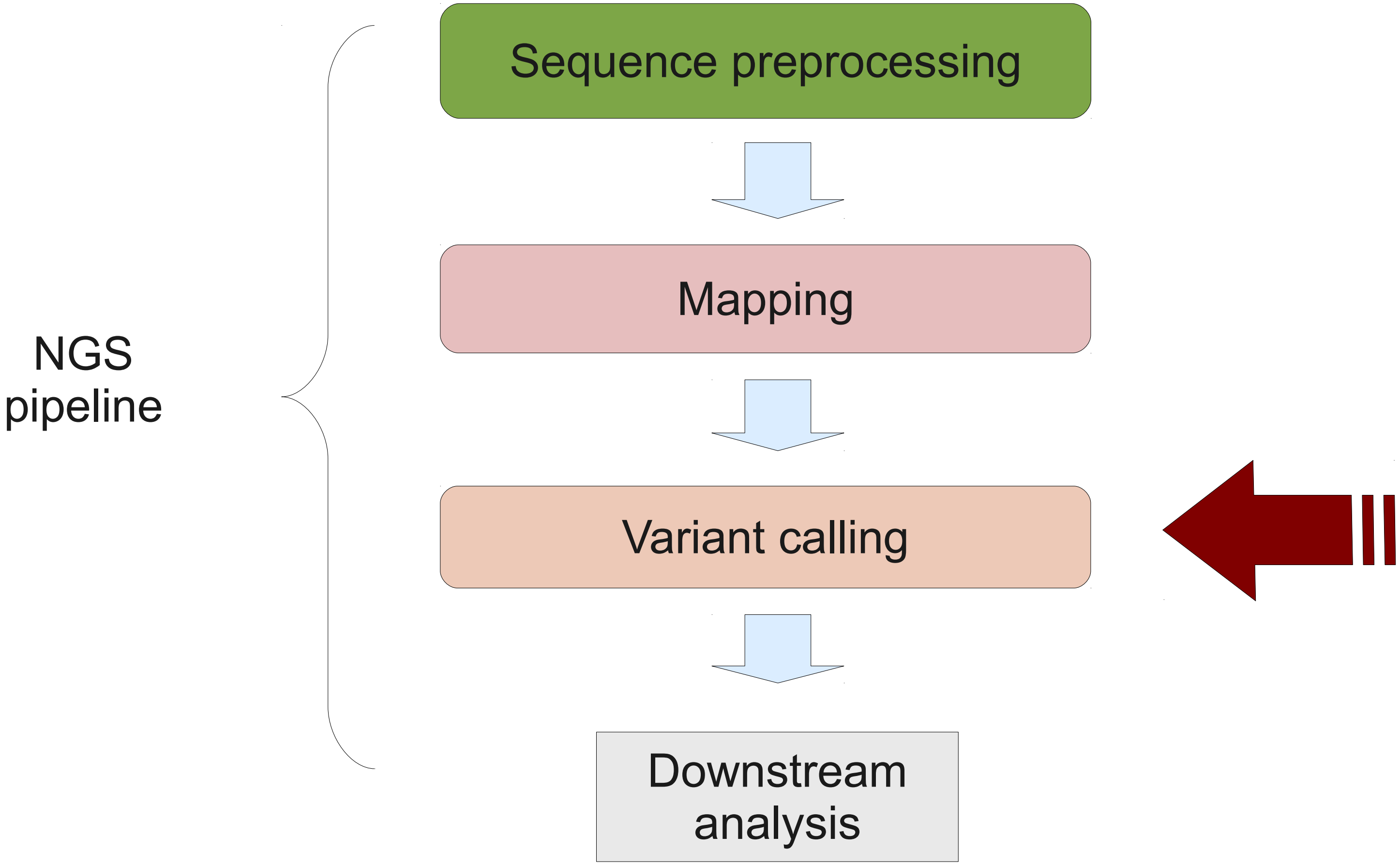
Jorge Jiménez
jjimeneza@cipf.es
BIER - CIBERER
Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)

Index

1. NGS workflow
2. Variant calling
3. Methods for calling
4. SNV and indel calling
5. VCF format
6. Missing values
7. Annotation
8. Databases

NGS Sequence preprocessing

Where we are?



NGS pipeline

Sequence preprocessing

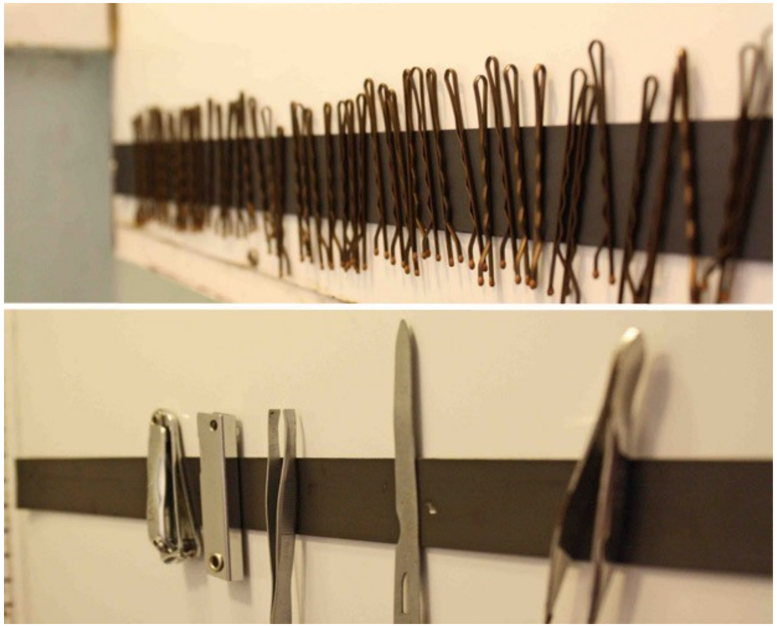
Mapping

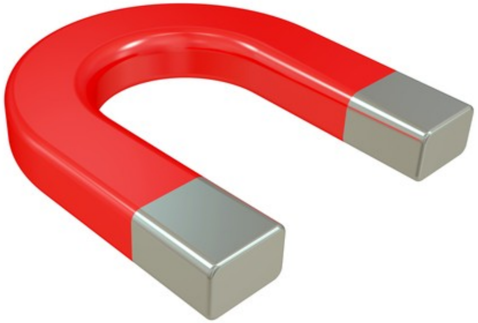
Variant calling

Downstream analysis

What is variant calling?

Finding A Needle In The Haystack?

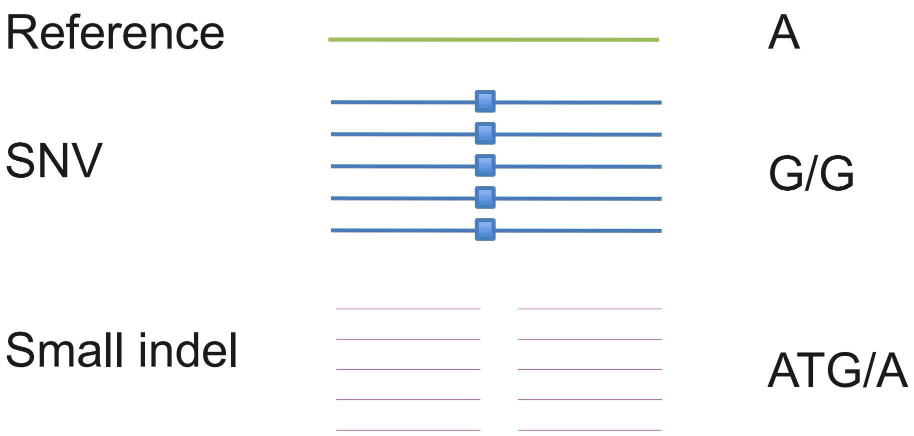




Variant types

SNV: Single nucleotide variant.

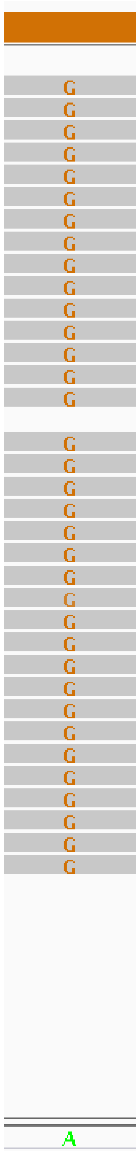
Indel: small insertion/deletion variant.

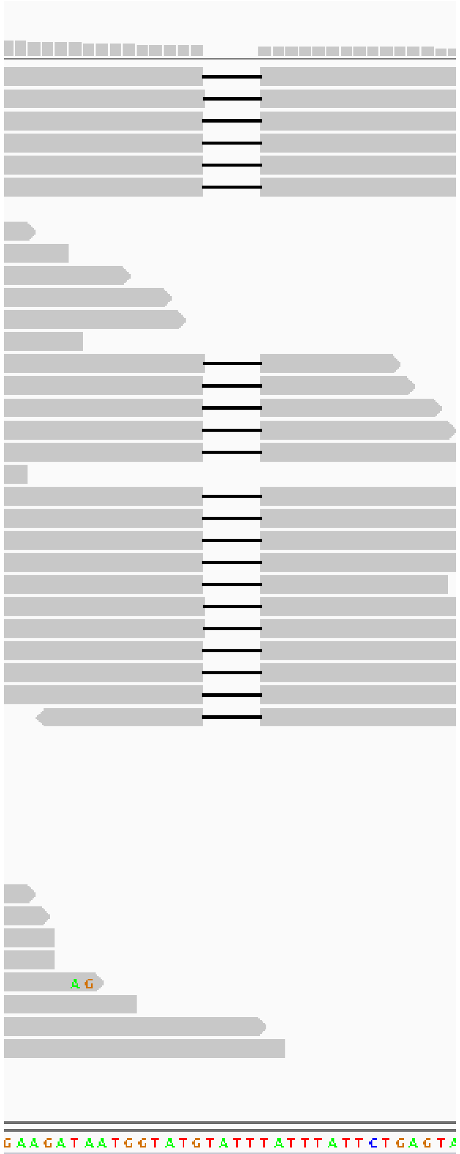


Reference A

SNV G/G

Small indel ATG/A





Genotype and variant calling – concepts

Phred Quality score:

$$Q_{\mathrm{Phred}} = -10 \log_{10} P(\mathrm{error}).$$

A score of 20 corresponds to 1% error rate in base calling

Variant calling: positions with at least one of the bases differs from reference.

Genotype calling: Process of determining the genotype of each variant.

Importance of base quality recalibration:

Obtaining well-calibrated quality scores is important, as SNP and genotype calling at a specific position in the genome depends on both the base calls and the per-base quality scores of the reads overlapping the position.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011 Jun;12(6):443-51. Review. PubMed PMID: 21587300.

Methods for calling

Early methods:

Counting the number of times each allele is observed.

Probabilistic methods:

They compute **genotype likelihood**.

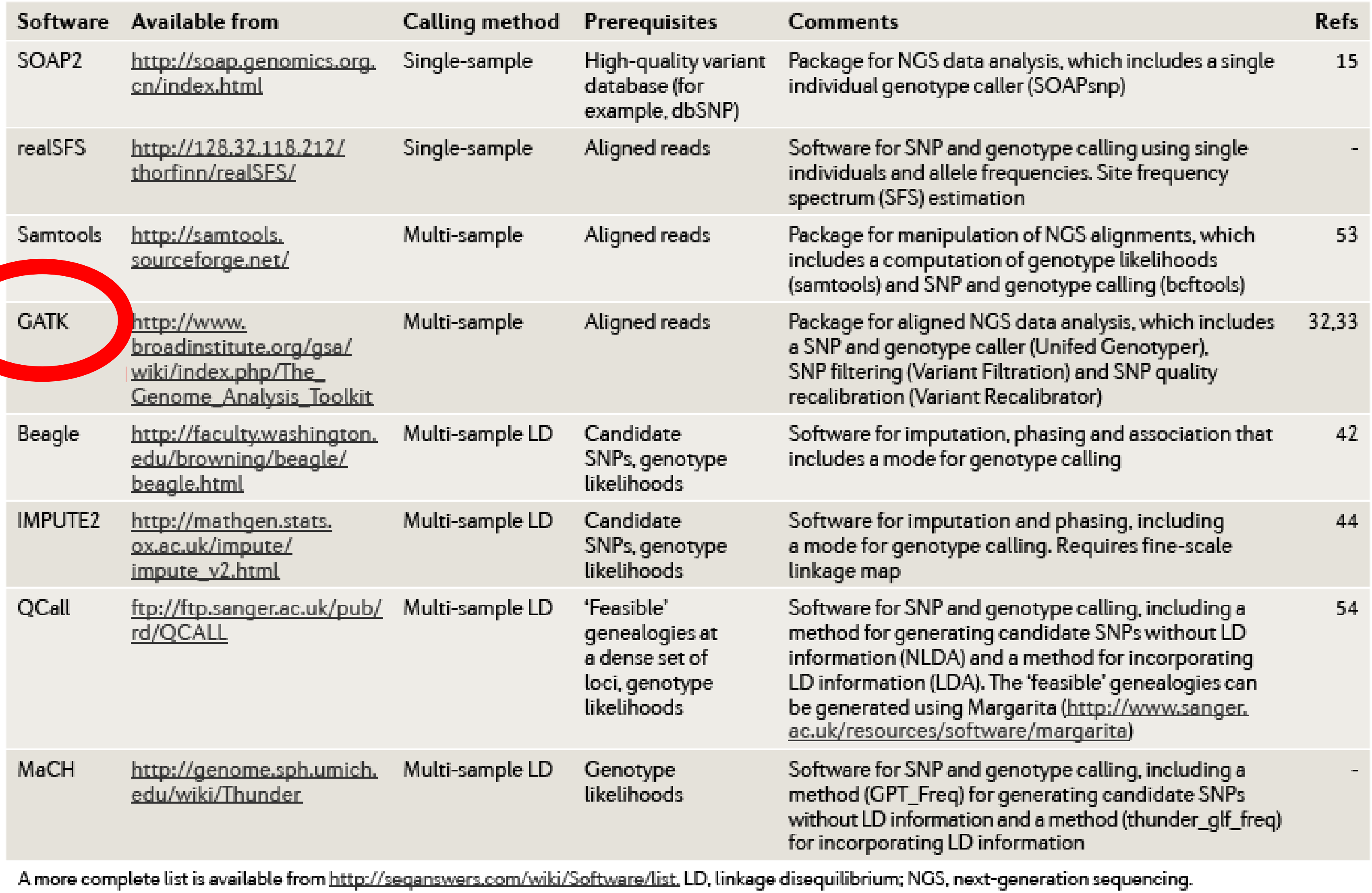
Advantages:

- Provide statistical measures of uncertainty.

- Lead to higher accuracy of genotype calling.

- Provide a natural framework for incorporating information: AF, LD.

Calling algorithms



Software	Available from	Calling method	Prerequisites	Comments	Refs
SOAP2	http://soap.genomics.org.cn/index.html	Single-sample	High-quality variant database (for example, dbSNP)	Package for NGS data analysis, which includes a single individual genotype caller (SOAPsnp)	15
realSFS	http://128.32.118.212/thorfinn/realSFS/	Single-sample	Aligned reads	Software for SNP and genotype calling using single individuals and allele frequencies. Site frequency spectrum (SFS) estimation	-
Samtools	http://samtools.sourceforge.net/	Multi-sample	Aligned reads	Package for manipulation of NGS alignments, which includes a computation of genotype likelihoods (samtools) and SNP and genotype calling (bcftools)	53
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	Multi-sample	Aligned reads	Package for aligned NGS data analysis, which includes a SNP and genotype caller (Unifed Genotyper), SNP filtering (Variant Filtration) and SNP quality recalibration (Variant Recalibrator)	32,33
Beagle	http://faculty.washington.edu/browning/beagle/beagle.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation, phasing and association that includes a mode for genotype calling	42
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Multi-sample LD	Candidate SNPs, genotype likelihoods	Software for imputation and phasing, including a mode for genotype calling. Requires fine-scale linkage map	44
QCall	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	Multi-sample LD	'Feasible' genealogies at a dense set of loci, genotype likelihoods	Software for SNP and genotype calling, including a method for generating candidate SNPs without LD information (NLDA) and a method for incorporating LD information (LDA). The 'feasible' genealogies can be generated using Margarita (http://www.sanger.ac.uk/resources/software/margarita)	54
MaCH	http://genome.sph.umich.edu/wiki/Thunder	Multi-sample LD	Genotype likelihoods	Software for SNP and genotype calling, including a method (GPT_Freq) for generating candidate SNPs without LD information and a method (thunder_glf_freq) for incorporating LD information	-

A more complete list is available from http://seqanswers.com/wiki/Software/list. LD, linkage disequilibrium; NGS, next-generation sequencing.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011 Jun;12(6):443-51. Review. PubMed PMID: 21587300.

Why GATK?

- Probabilistic method: Bayesian estimation of the most likely genotype.

- Calculates many parameters for each position of the genome.

- SNP and indel calling.

- Used in many NGS projects, including the 1000 Genomes Project, The Cancer Genome Atlas, etc.

- Base quality recalibration.

- Uses standard input and output files.

- Many tools for manage VCF files.

Indel calling

- Many available softwares like dindel, samtools, frebayes, ...

- Sequence aligners are often unable to perfectly map reads containing insertions or deletions.

- Indel-containing reads can be either less unmapped or arranged in gapless alignments.

- Mismatches in a particular read can interfere with the gap.

- Indel detection becomes difficult with so many missing reads.

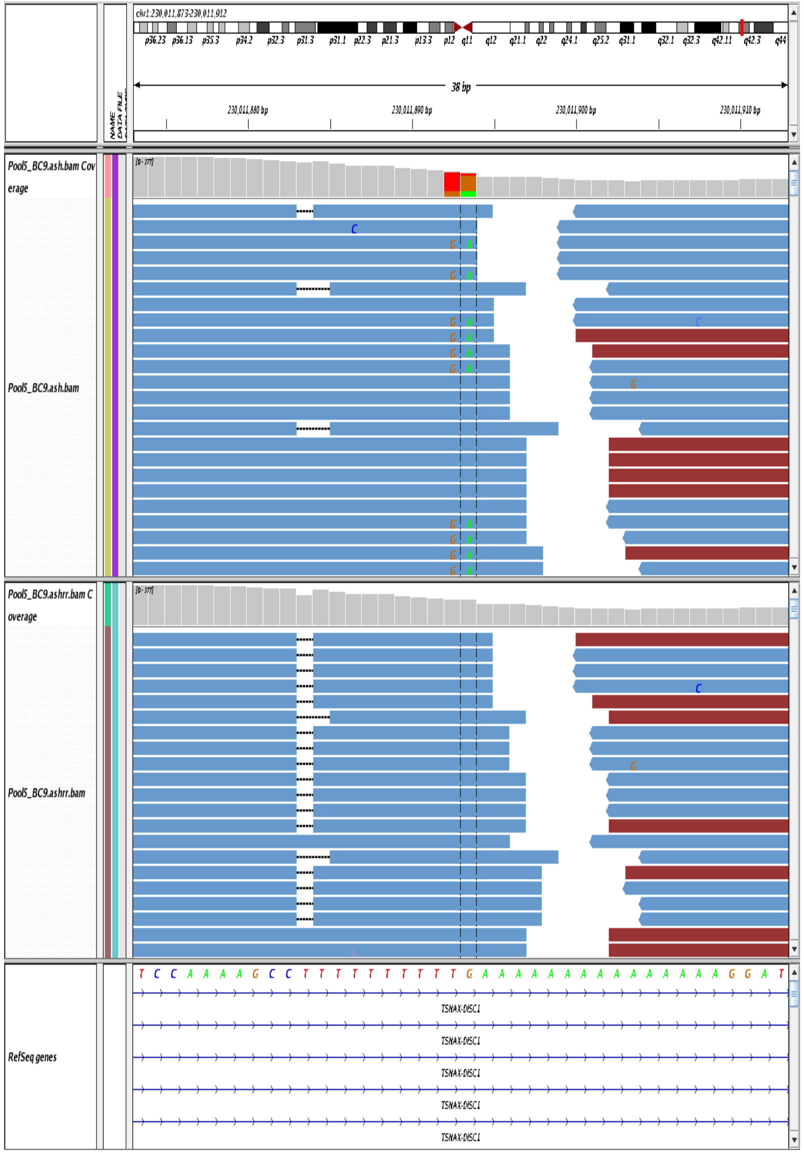
- Artifacts introduced by the gapless alignments cause the appearance of false positive SNPs (usually in clusters) → Local realignment

GATK

Local realignment

Local realignment of all reads at a specific location simultaneously to minimize mismatches to the reference genome

Reduces erroneous SNPs refines location of INDELS


chr1:230,011,873-230,011,912
p36.23 p36.13 p35.3 p34.2 p32.3 p31.3 p31.1 p22.3 p21.3 p13.3 p12 q11 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3 q44
38 bp
230,011,880 bp
230,011,890 bp
230,011,900 bp
230,011,910 bp
NAME
DATA FILE
Pool5_BC9.ash.bam Coverage
[0 - 377]
Pool5_BC9.ash.bam
Pool5_BC9.ashrr.bam Coverage
[0 - 377]
Pool5_BC9.ashrr.bam
RefSeq genes
T C C A A A A G C C T T T T T T T T T T G A A A A A A A A A A A A A A A G G A T
TSNAX-DISC1
TSNAX-DISC1
TSNAX-DISC1
TSNAX-DISC1
TSNAX-DISC1

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. PMID: 21478889

Calling all bases: missing values

We want to know the genotype of all the bases of the exome.

Calling:

- SNVs + all sites of capture kit

- indels

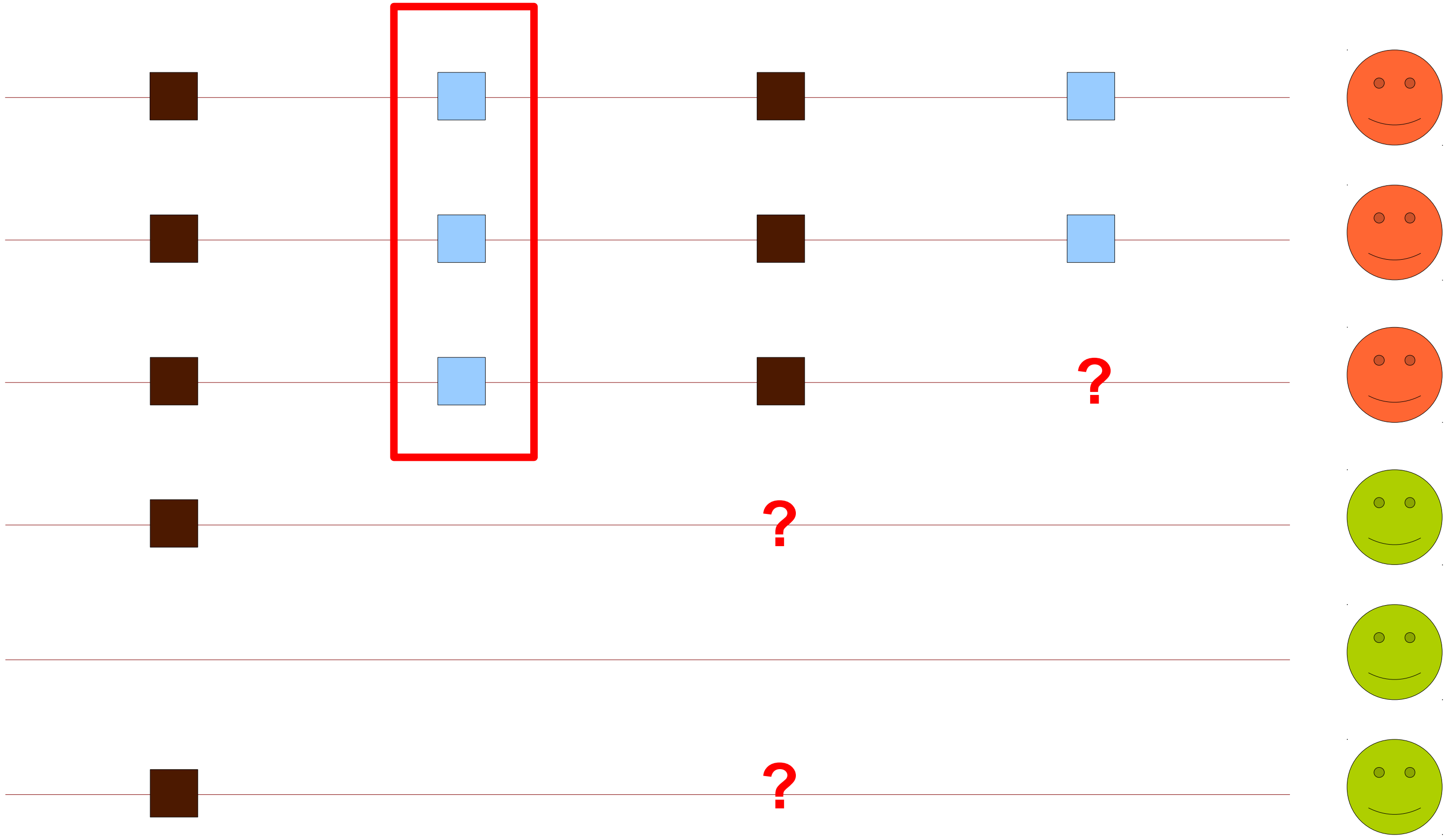
Two types of missing values:

No coverage: ./. → Not sequenced base

Filtered: -/- → Low quality base

We do not know the genotype of these bases

Missing values



False positives and negatives

FALSE POSITIVES:

Error rate: 1/10,000 bases

False positives can be decreased by increasing the number of control and cases samples.

Genes: MUC4, MUC16, ORF, ...

FALSE NEGATIVES:

Variants not sequenced → missing values

Errors in variants

VCF format

```
##[HEADER LINES]
#CHROM  POS       ID          REF   ALT   QUAL      FILTER  INFO          FORMAT          NA12878
1       873762    .           T     G     5231.78   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
1       877664    rs3828047   A     G     3931.66   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
1       899282    rs28548431  C     T     71.77     PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:25.92:103,0,26
1       974165    rs9442391   T     C     29.84     LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:60.91:61,0,255
```

INFO fields:

```
AC=1;AF=0.50;AN=2;DP=315;Dels=0.00;HRun=2;HaplotypeScore=15.11;MQ=91.05;MQ0=15;QD=16.61;SB=-1533.02;VQSLOD=-1.5473
```

QD: QualByDepth. Variant confidence. Low scores are indicative of false positives calls and artifacts.

ReadPosRankSum: Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele.

FS: Fisher's Exact Test to detect strand bias.

MQ: Root Mean Square of the mapping quality of the reads across all samples.

HaplotypeScore: Consistency of the site with two (and only two) segregating haplotypes.

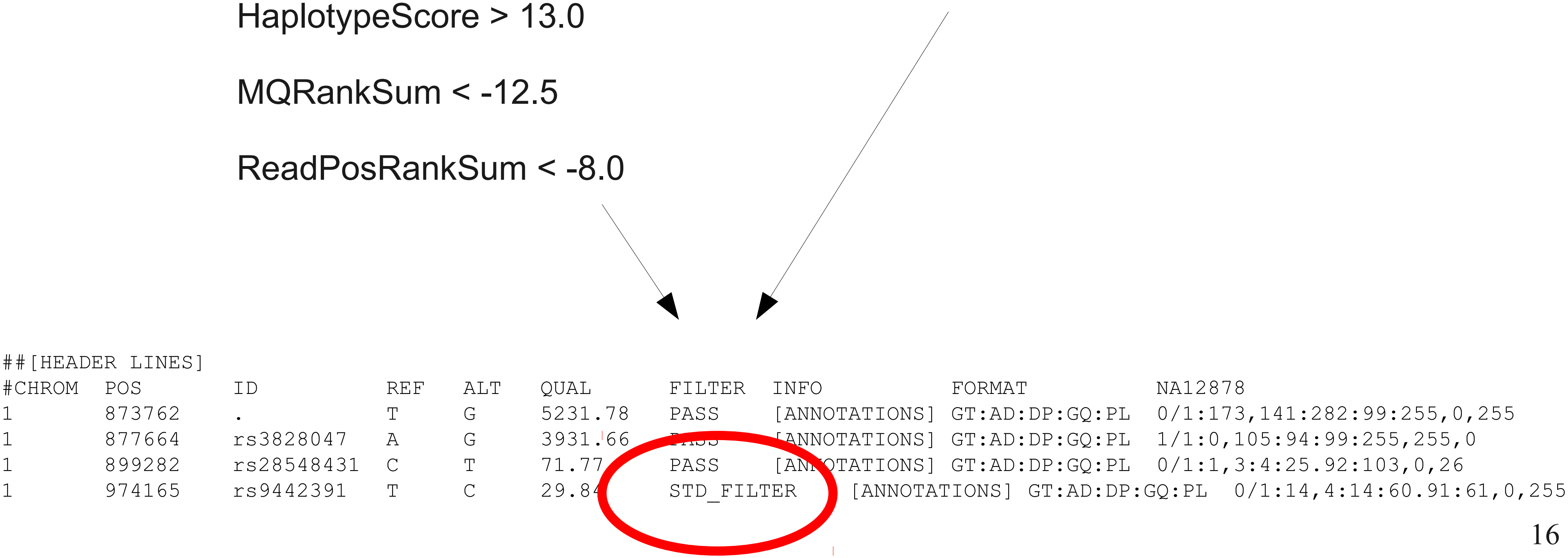
MQRankSum: Mann-Whitney Rank Sum Test for mapping qualities.

More info: http://www.broadinstitute.org/gsa/wiki/index.php/Understanding_the_Unified_Genotyper%27s_VCF_files

Filtering SNVs and indels

Filtering parameters for SNVs and indels are different.

SNVs	**indels**
QD < 2.0	QD < 2.0
MQ < 40.0	FS > 200.0
FS > 60.0	ReadPosRankSum < -20.0
HaplotypeScore > 13.0	
MQRankSum < -12.5	
ReadPosRankSum < -8.0	



```
##[HEADER LINES]
#CHROM   POS      ID          REF    ALT    QUAL       FILTER   INFO           FORMAT           NA12878
1        873762   .           T      G      5231.78    PASS     [ANNOTATIONS]  GT:AD:DP:GQ:PL   0/1:173,141:282:99:255,0,255
1        877664   rs3828047   A      G      3931.66    PASS     [ANNOTATIONS]  GT:AD:DP:GQ:PL   1/1:0,105:94:99:255,255,0
1        899282   rs28548431  C      T      71.77      PASS     [ANNOTATIONS]  GT:AD:DP:GQ:PL   0/1:1,3:4:25.92:103,0,26
1        974165   rs9442391   T      C      29.84      STD_FILTER     [ANNOTATIONS] GT:AD:DP:GQ:PL   0/1:14,4:14:60.91:61,0,255
```

Annotation

Definition:

An annotation is a note added by way of explanation or commentary.

Annotation of the variants:

- gene → RefSeq

- consequence type.

- dbSNP version 135.

- alelle frequency of alternative allele in 1000 genomes project.

- alelle frequency of alternative allele in NHLBI GO Exome Sequencing Project.

- PolyPhen.

- SIFT.

- conservation.

- disease associated to the variant.

- associated disease to the gene.

- GO terms.

Variant

Medina I, De Maria A, Bleda M, et al. "VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing." Nucleic Acids Res.. 2012;40(Web Server issue):W54-8.

- It can annotate single nucleotide variants (SNVs) and insertions/deletions.

- Very fast.

- web server.

- Input: VCF file.

- Annotation of all the transcripts.

- (...)

http://variant.bioinfo.cipf.es/

Annovar

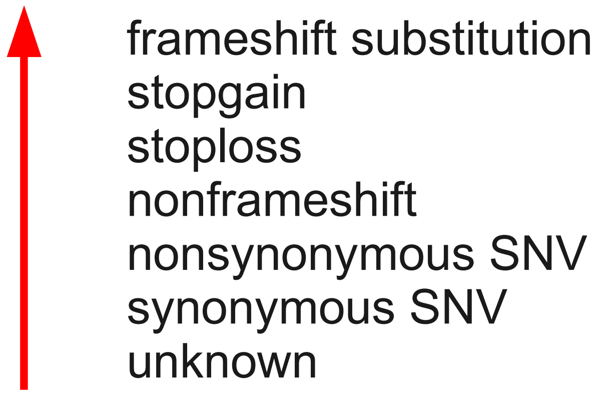
Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.

- It can annotate single nucleotide variants (SNVs) and insertions/deletions.

- Gene-based, region-based and filter-based annotation of genetic variants.

- Versatile annotator: custom annotations.

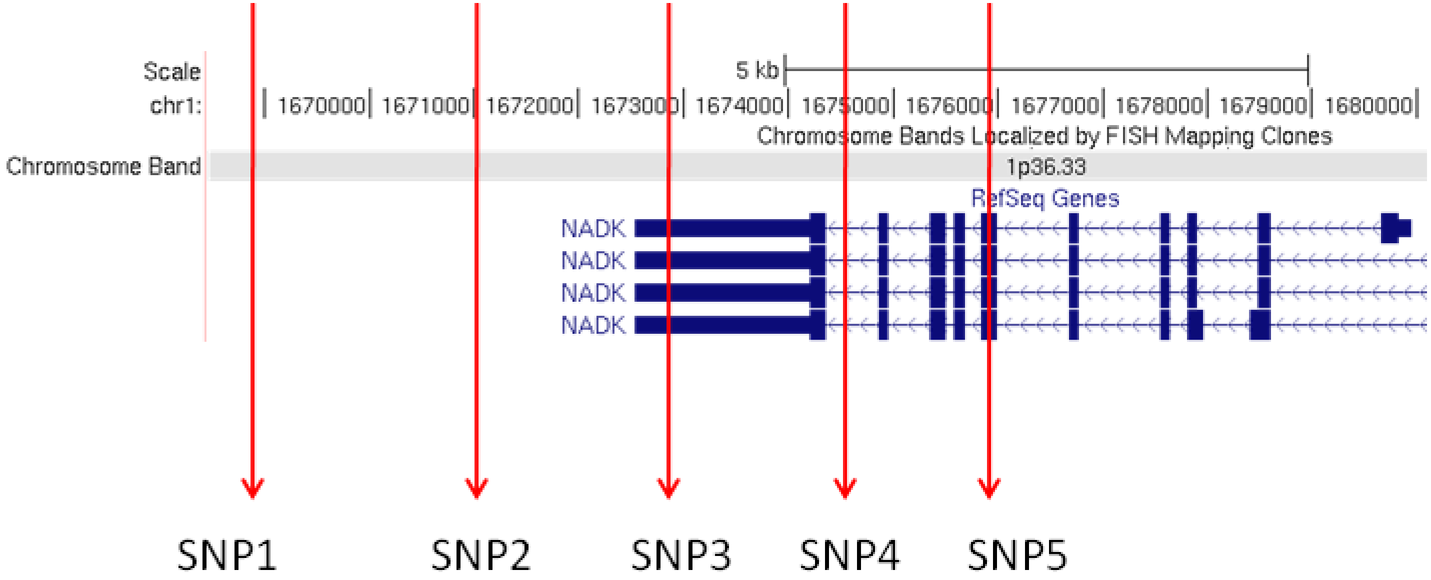
- Only print the worst consequence.



frameshift substitution
stopgain
stoploss
nonframeshift
nonsynonymous SNV
synonymous SNV
unknown

http://www.openbioinformatics.org/annovar/

Consequence type



frameshift substitution
stopgain
stoploss
nonframeshift
nonsynonymous SNV
synonymous SNV
unknown

exonic
splicing
ncRNA
UTR5
UTR3
intronic
upstream
downstream
intergenic

SIFT/PolyPhen

Assigns a “functional importance” score to SNV

SIFT:

Score < 0.5 → no benign variants

Score >= 0.5 → benign variants

http://sift.jcvi.org/

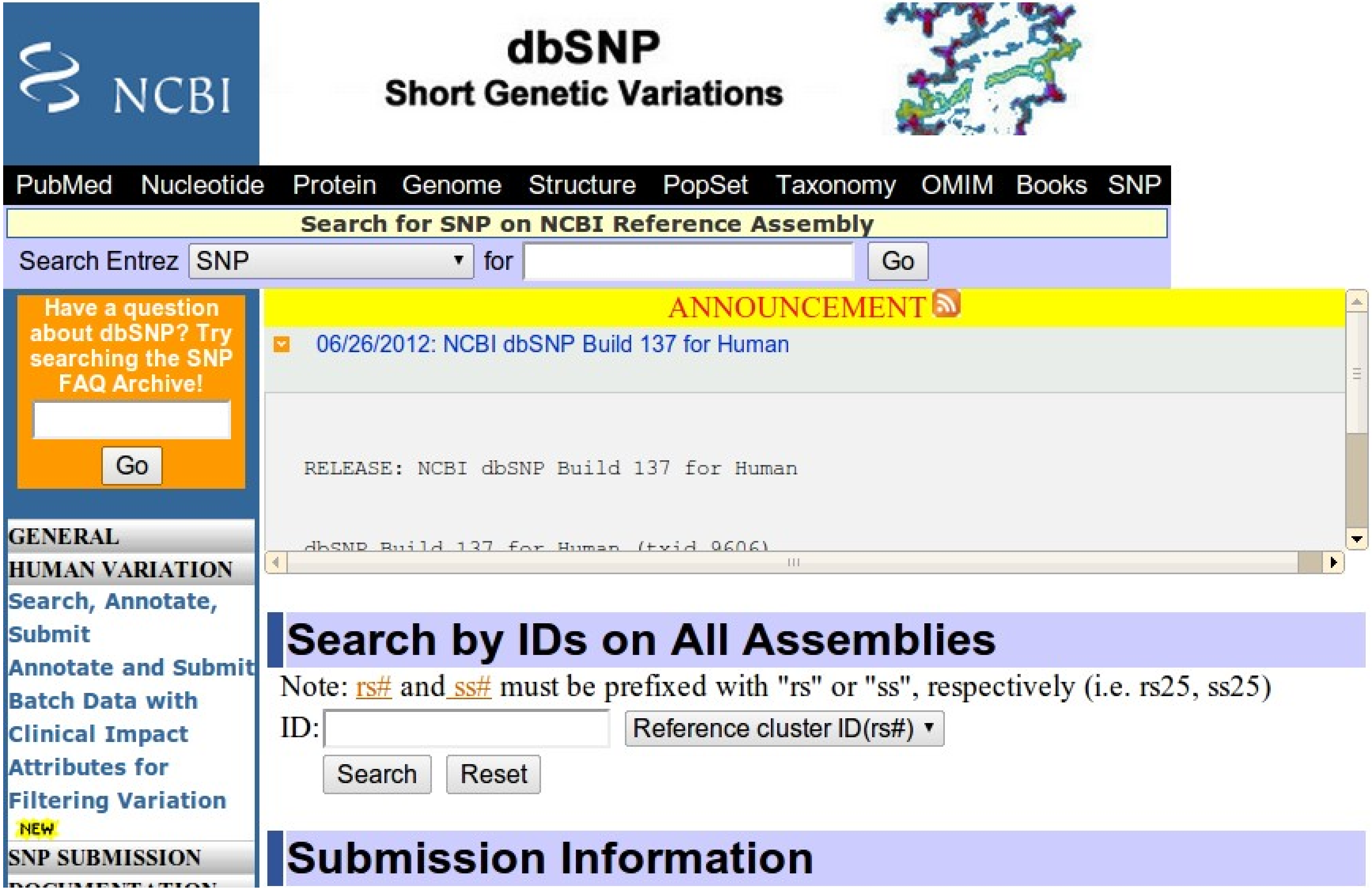
PolyPhen:

0 → benign variants

1 → no benign variants

http://genetics.bwh.harvard.edu/pph2/

dbSNP



dbSNP

Short Genetic Variations

NCBI
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

ANNOUNCEMENT

06/26/2012: NCBI dbSNP Build 137 for Human

```
RELEASE: NCBI dbSNP Build 137 for Human

dbSNP Build 137 for Human (txid 9606)
```

GENERAL
HUMAN VARIATION
Search, Annotate, Submit
Annotate and Submit
Batch Data with Clinical Impact
Attributes for Filtering Variation
NEW
SNP SUBMISSION
DOCUMENTATION

Search by IDs on All Assemblies

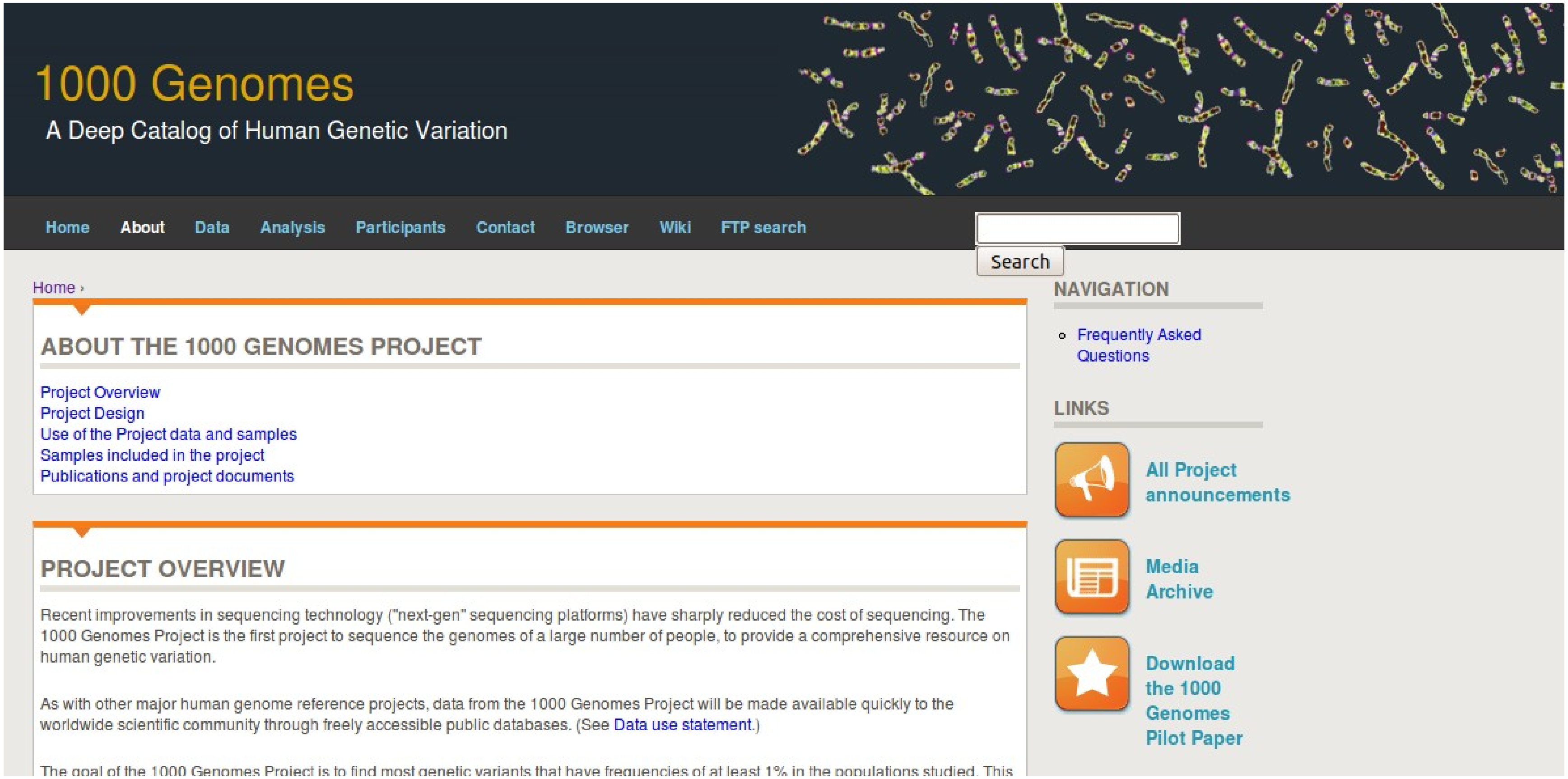
Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

ID: Reference cluster ID(rs#)

Search Reset

Submission Information

1000 genomes project



1000 Genomes

A Deep Catalog of Human Genetic Variation

Home ›

ABOUT THE 1000 GENOMES PROJECT

- Project Overview
- Project Design
- Use of the Project data and samples
- Samples included in the project
- Publications and project documents

PROJECT OVERVIEW

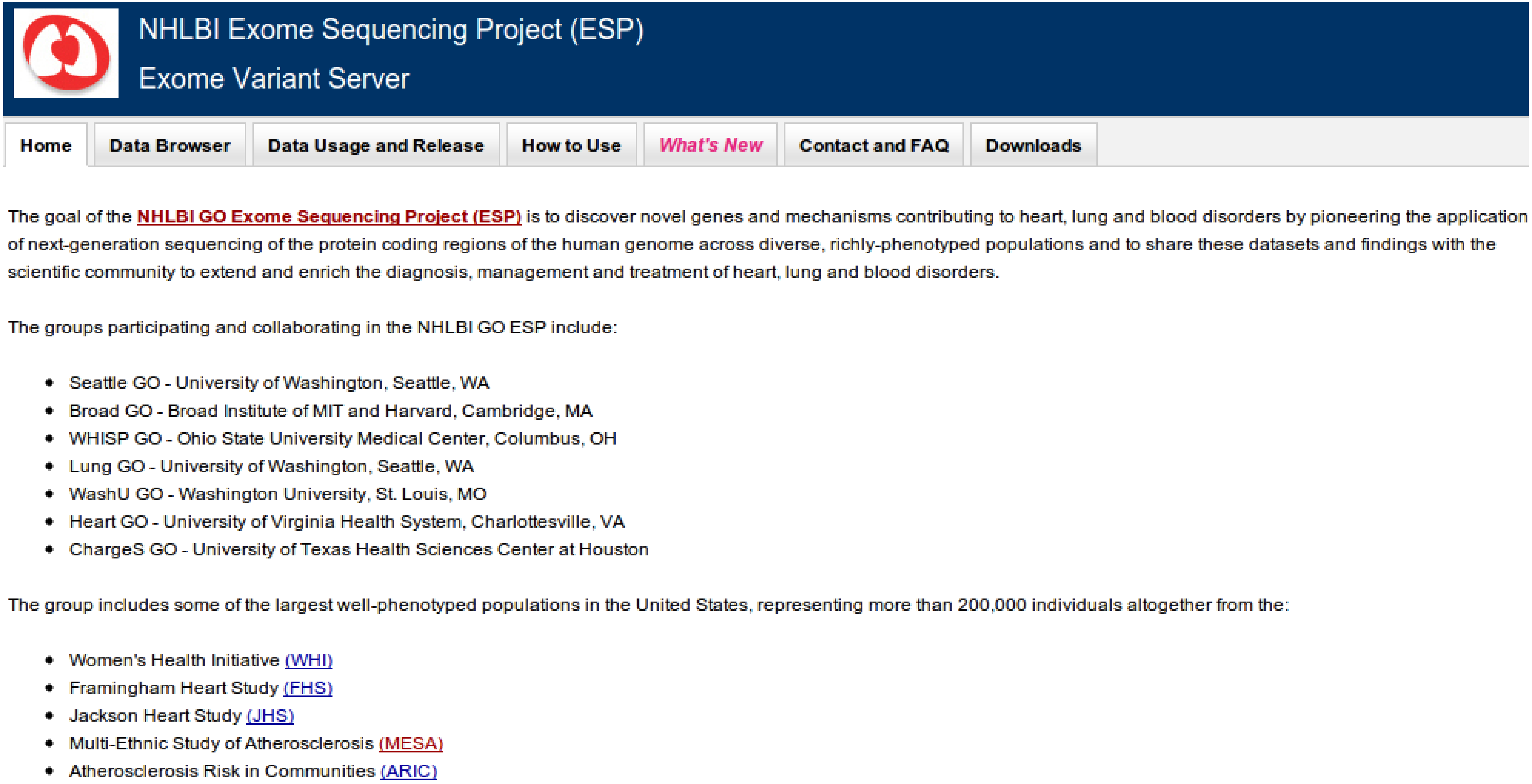
Recent improvements in sequencing technology ("next-gen" sequencing platforms) have sharply reduced the cost of sequencing. The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation.

As with other major human genome reference projects, data from the 1000 Genomes Project will be made available quickly to the worldwide scientific community through freely accessible public databases. (See Data use statement.)

The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied. This

NAVIGATION
Frequently Asked Questions
LINKS
All Project announcements
Media Archive
Download the 1000 Genomes Pilot Paper

NHLBI Exome Sequencing Project (ESP)



The goal of the NHLBI GO Exome Sequencing Project (ESP) is to discover novel genes and mechanisms contributing to heart, lung and blood disorders by pioneering the application of next-generation sequencing of the protein coding regions of the human genome across diverse, richly-phenotyped populations and to share these datasets and findings with the scientific community to extend and enrich the diagnosis, management and treatment of heart, lung and blood disorders.

The groups participating and collaborating in the NHLBI GO ESP include:

- Seattle GO - University of Washington, Seattle, WA
- Broad GO - Broad Institute of MIT and Harvard, Cambridge, MA
- WHISP GO - Ohio State University Medical Center, Columbus, OH
- Lung GO - University of Washington, Seattle, WA
- WashU GO - Washington University, St. Louis, MO
- Heart GO - University of Virginia Health System, Charlottesville, VA
- ChargeS GO - University of Texas Health Sciences Center at Houston

The group includes some of the largest well-phenotyped populations in the United States, representing more than 200,000 individuals altogether from the:

- Women's Health Initiative (WHI)
- Framingham Heart Study (FHS)
- Jackson Heart Study (JHS)
- Multi-Ethnic Study of Atherosclerosis (MESA)
- Atherosclerosis Risk in Communities (ARIC)

TO DO

- Decrease number of false positives

Statistical analisys to detect errors

Statistical analysis to improve filters

- Improve indel calling

- Improve annotation

- (...)

Questions?