IX International Course of Massive Data Analysis FOR GENOMICS

Jorge Jiménez
jjimeneza@cipf.es

Variant calling

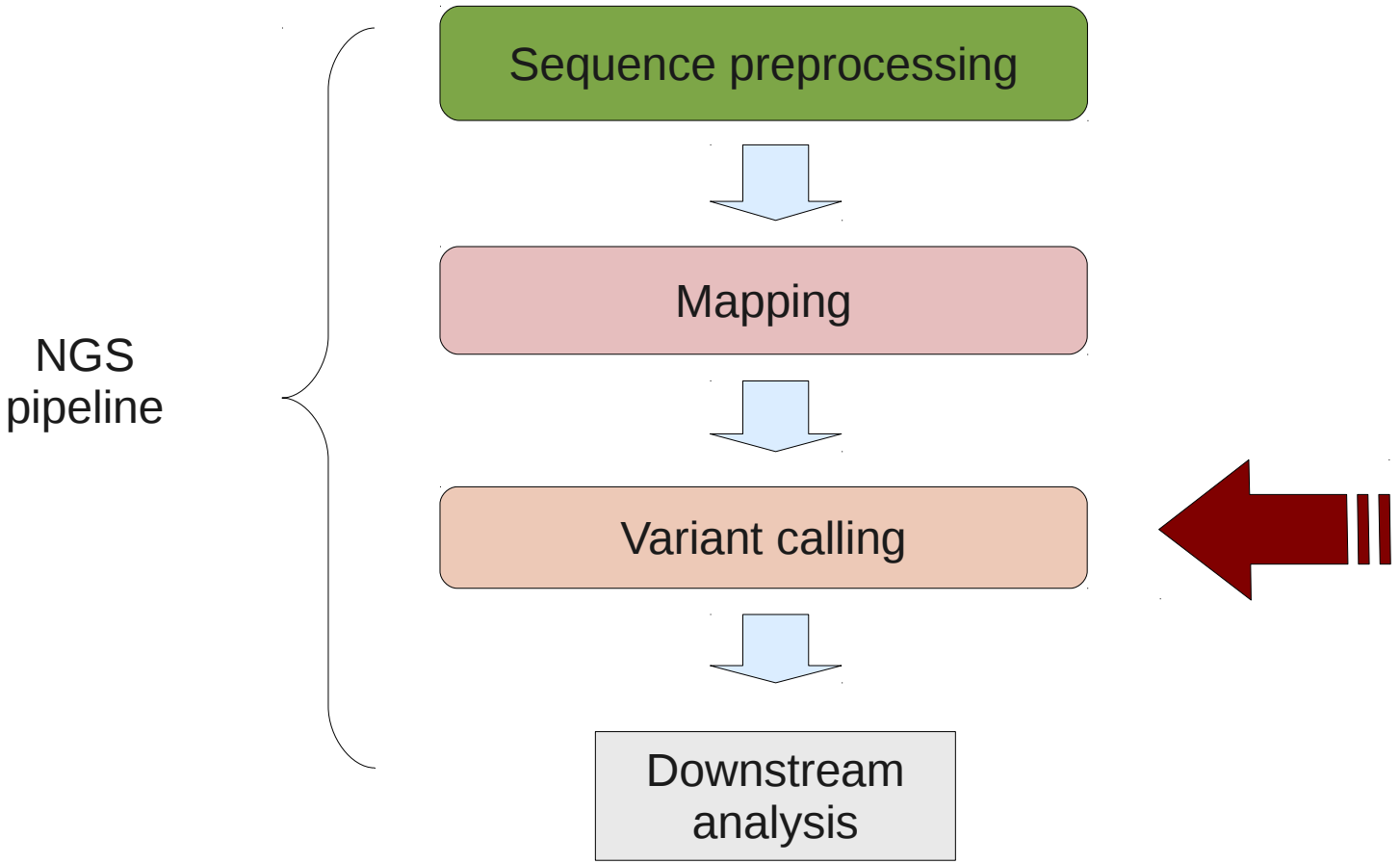
Index

Jorge Jiménez

jjimeneza@cipf.es

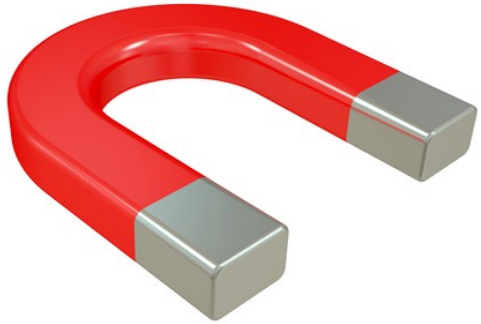
Index

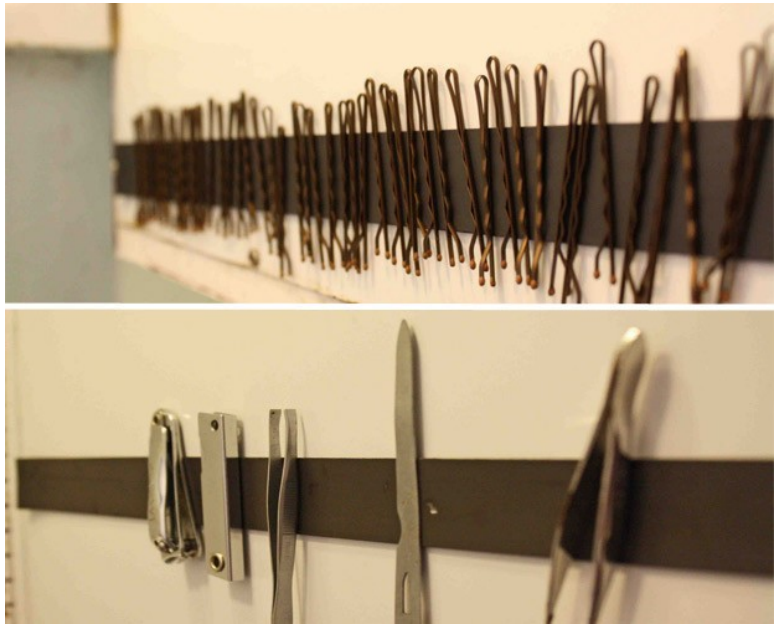
Where we are?



Jorge Jiménez
jjimeneza@cipf.es

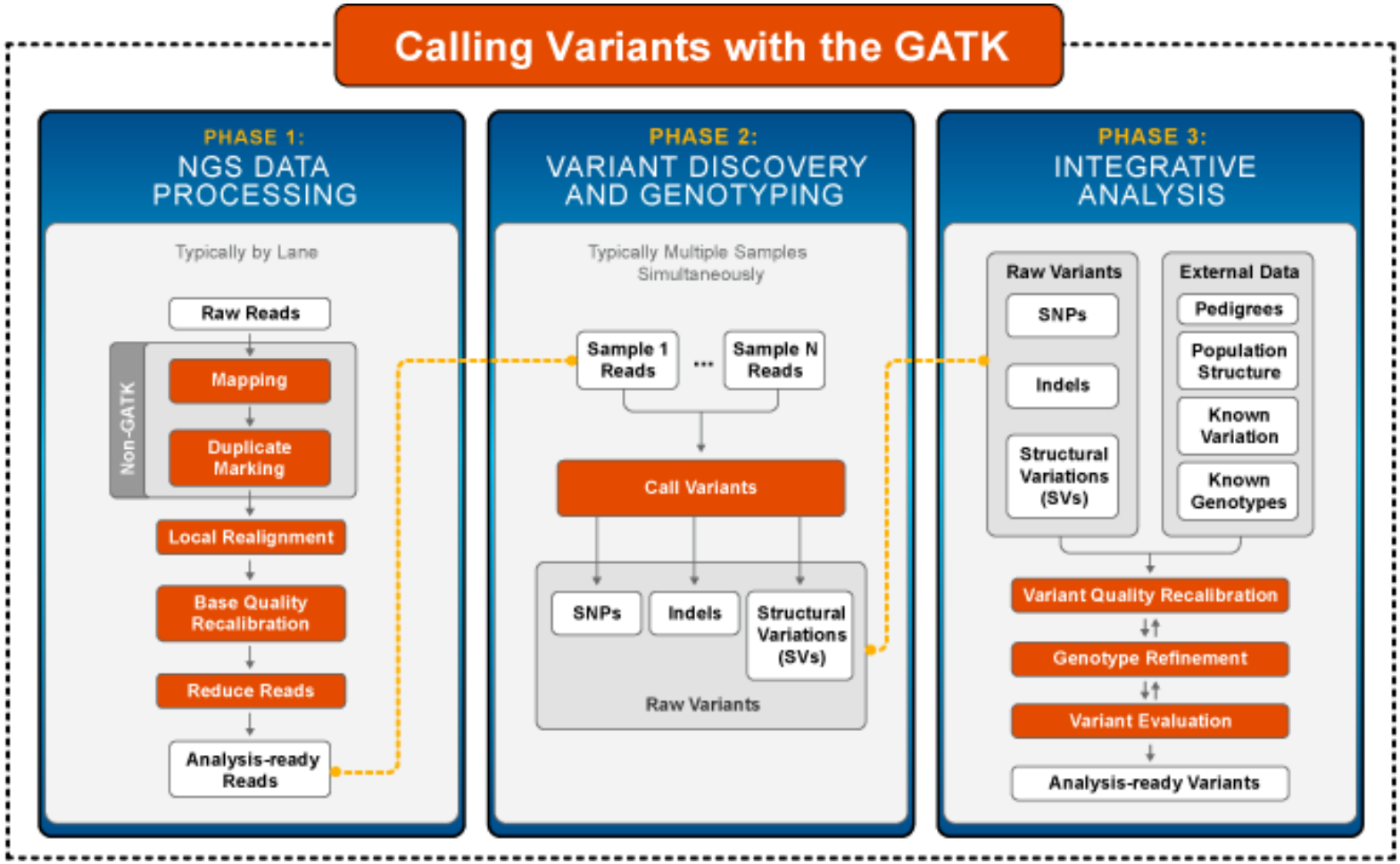
Finding A Needle In The Haystack?





Jorge Jiménez
jjimeneza@cipf.es

Variant Calling pipeline



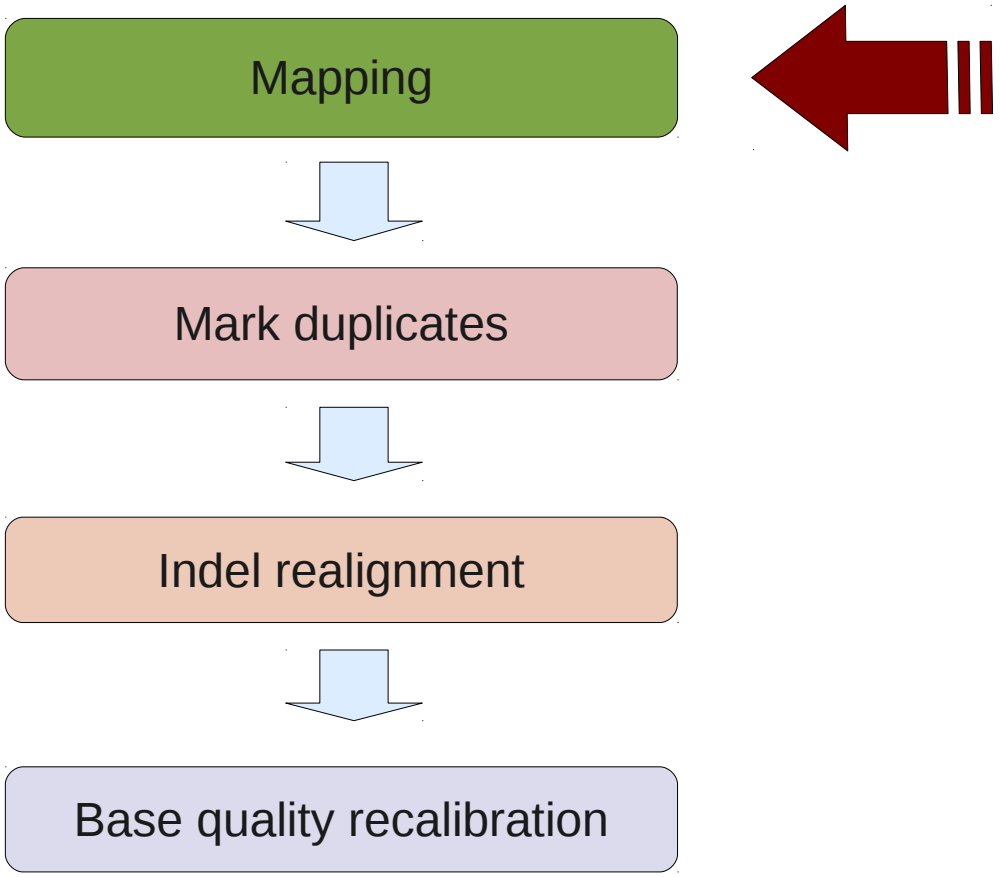
Jorge Jiménez
jjimeneza@cipf.es

Index

Jorge Jiménez

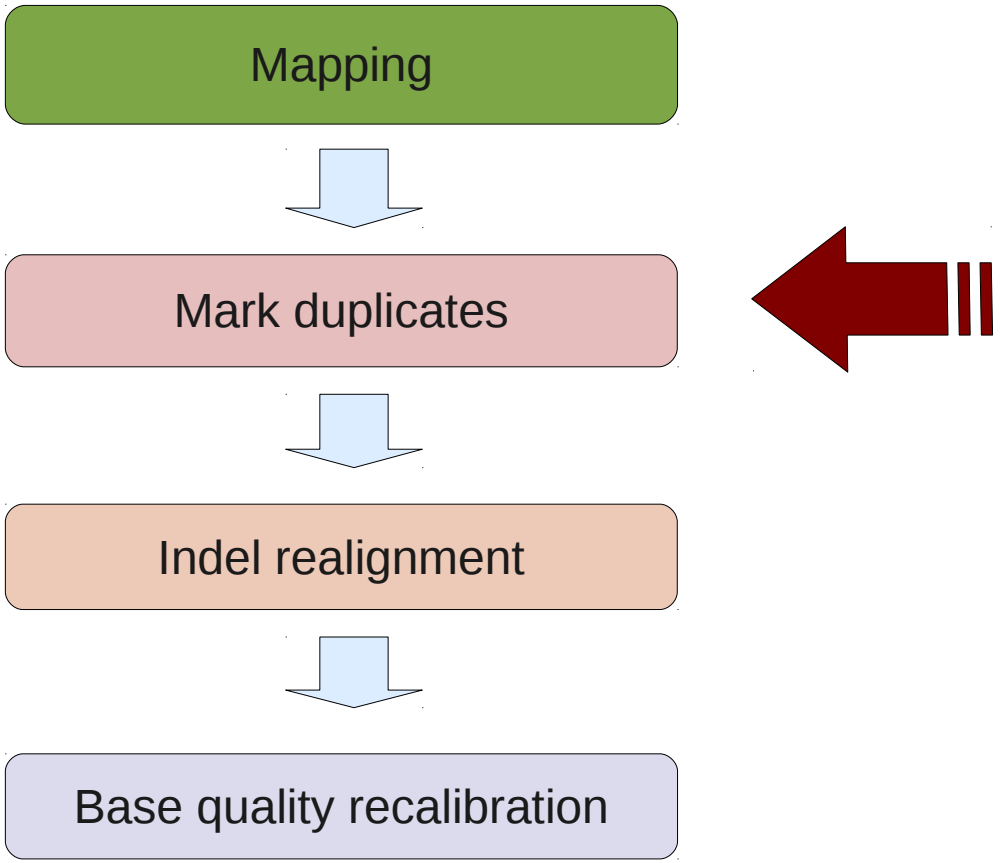
jjimeneza@cipf.es

Alignment processing



Jorge Jiménez
jjimeneza@cipf.es

Alignment processing



Jorge Jiménez
jjimeneza@cipf.es

Marking duplicates

All second-generation sequencing platforms are NOT single molecule sequencing

- PCR amplification step in library preparation
- Can result in duplicate DNA fragments in the final library prep.
- PCR-free protocols do exist – require large volumes of input DNA

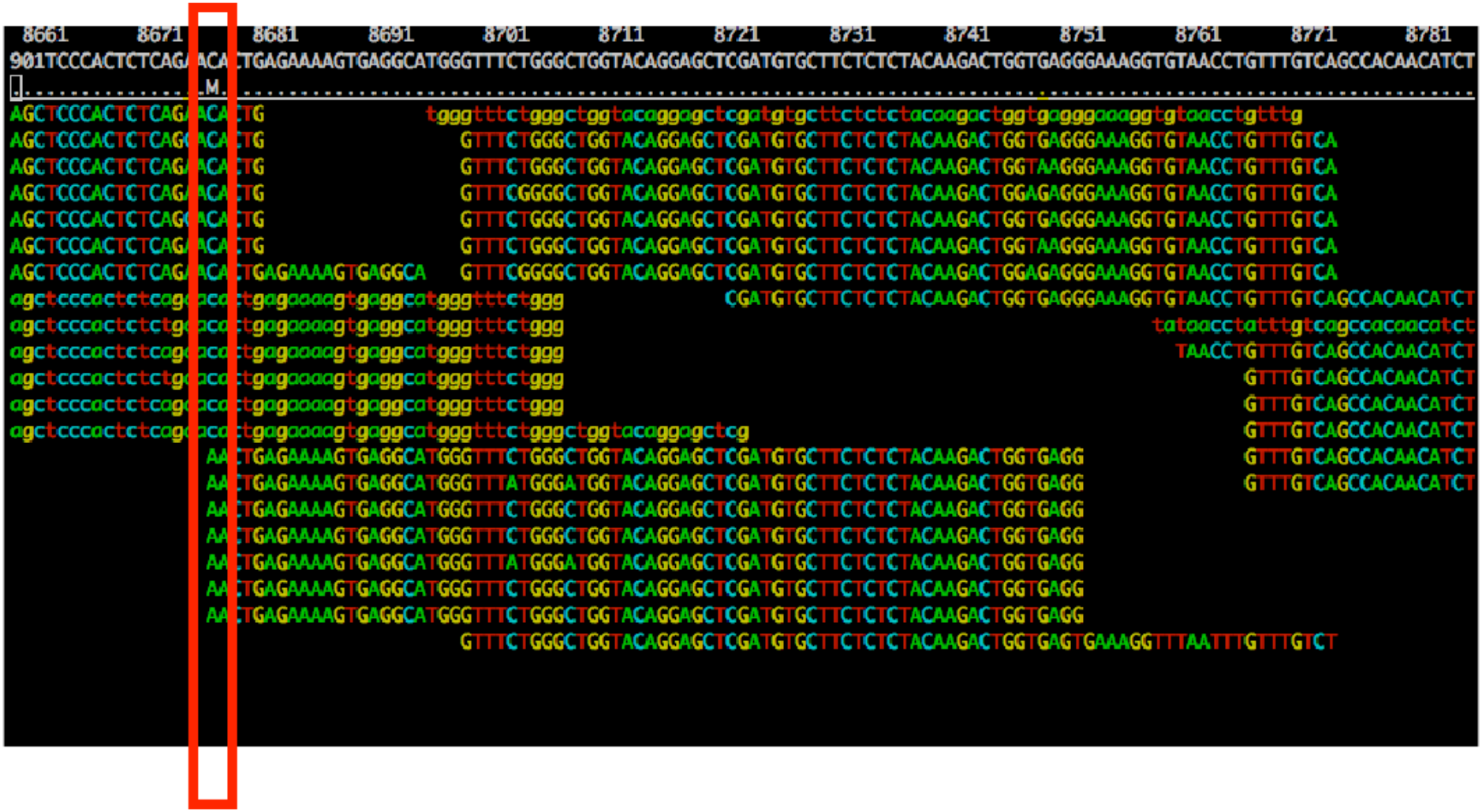
Generally low number of duplicates in good libraries (<3%)

- Align reads to the reference genome
- Identify read-pairs where the outer ends map to the same position on the genome and remove all but 1 copy
    - Samtools: samtools rmdup or samtools rmdupse
    - Picard/GATK: MarkDuplicates

Can result in false SNP calls

- Duplicates manifest themselves as high read depth support

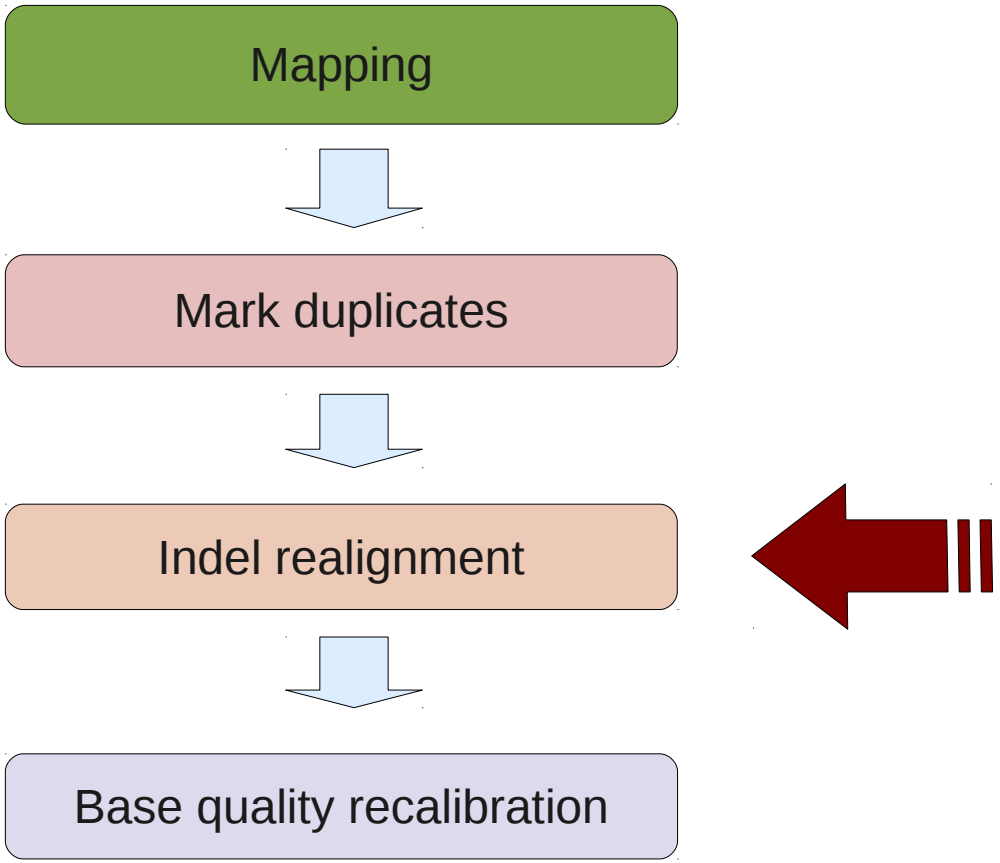
Duplicates and false SNPs



Jorge Jiménez
jjimeneza@cipf.es

Variant calling

Alignment processing



Jorge Jiménez
jjimeneza@cipf.es

Indel realignment

Short indels can pose difficulties for alignment programs

Realignment algorithm

- Input set of known indel sites and a BAM file
- At each site, model the indel haplotype and the reference haplotype
- Given the information on a known indel
    -Which scenario are the reads more likely to be derived from?
- New BAM file produced with read cigar lines modified where indels have been introduced by the realignment process

Software: GATK

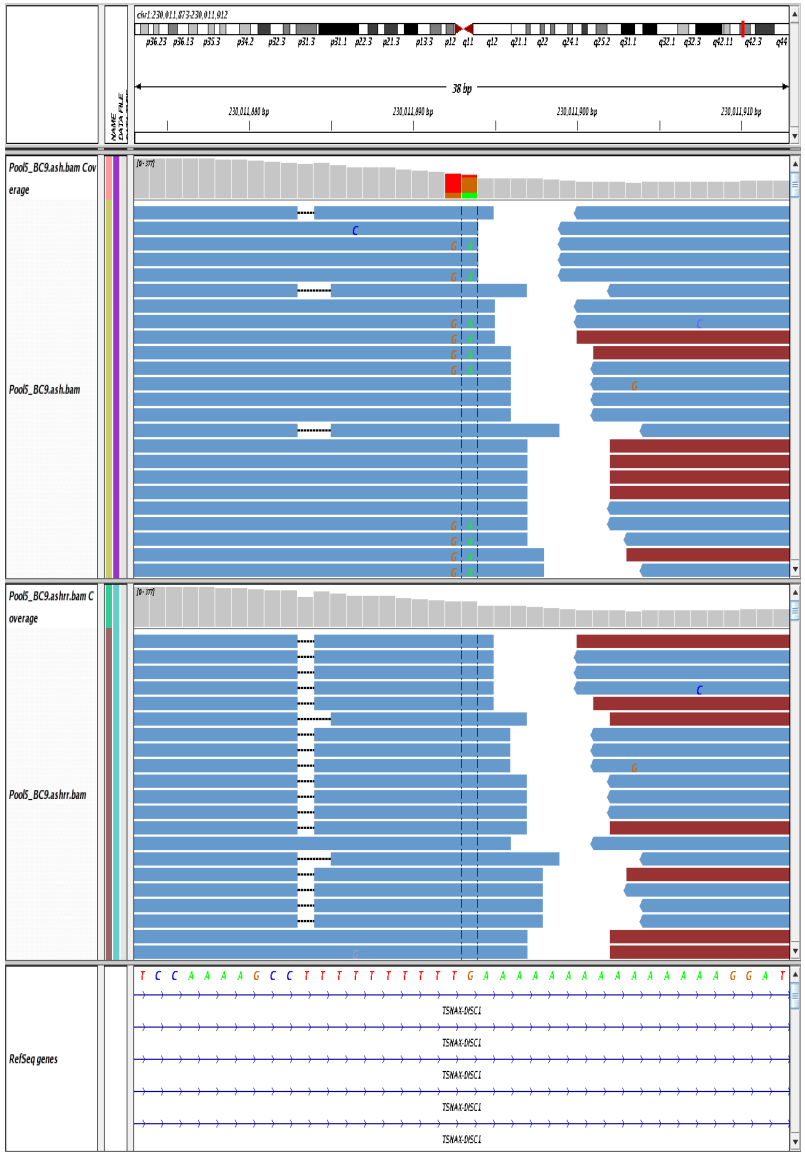
What sites?

- Previously published indel sites, dbSNP, 1000 genomes, generate a rough/ high confidence indel set

Indel realignment

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

Reduces erroneous SNP and refines location of INDELS

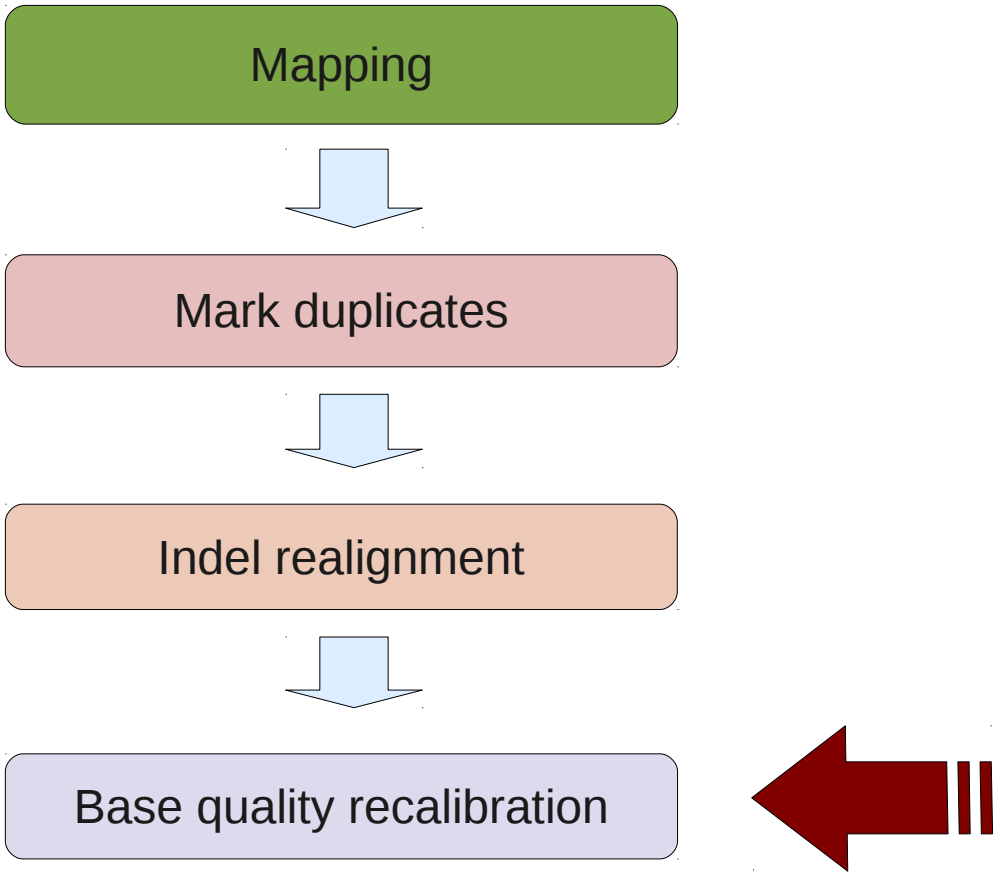


DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8. PMID: 21478889

Jorge Jiménez

jjimeneza@cipf.es

Alignment processing



Jorge Jiménez
jjimeneza@cipf.es

Base quality recalibration

Aim:

- The reported quality score is closer to its actual probability of mismatching the reference genome
- This tool attempts to correct for variation in quality with machine cycle and sequence context.

It analyzes the covariation among several features of a base:

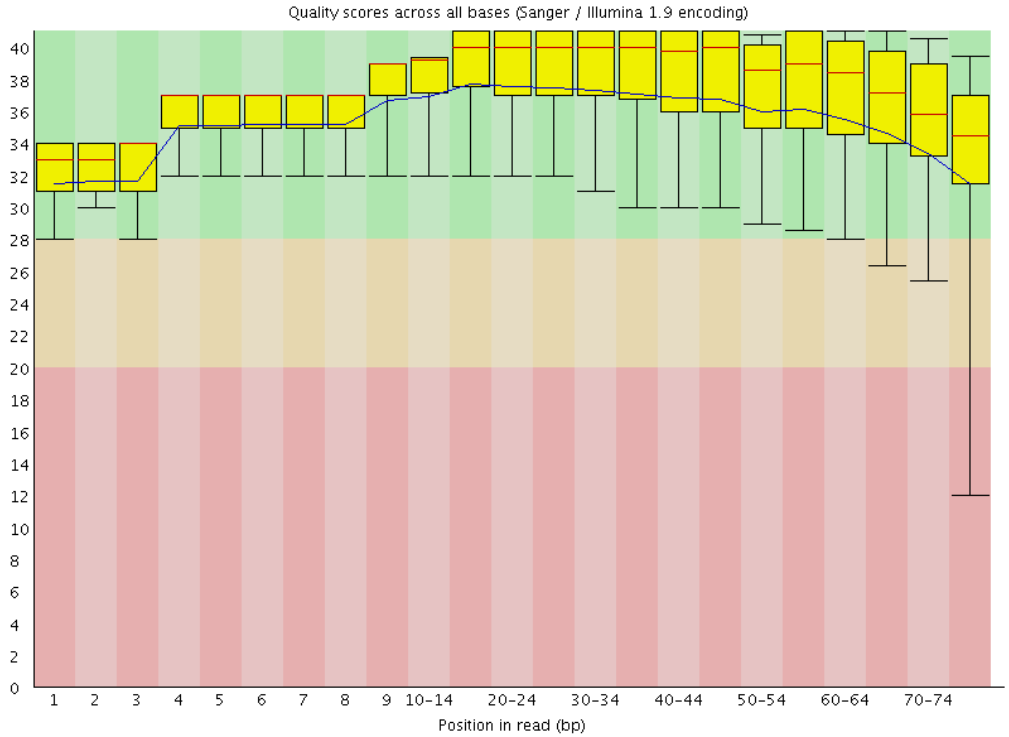
- Reported/original quality score
- The position within the read
- The preceding and current nucleotide (sequencing chemistry effect) observed by the sequencing machine
- Probability of mismatching the reference genome

These covariates are then subsequently applied through a piecewise tabular correction to recalibrate the quality scores of all reads in a BAM file

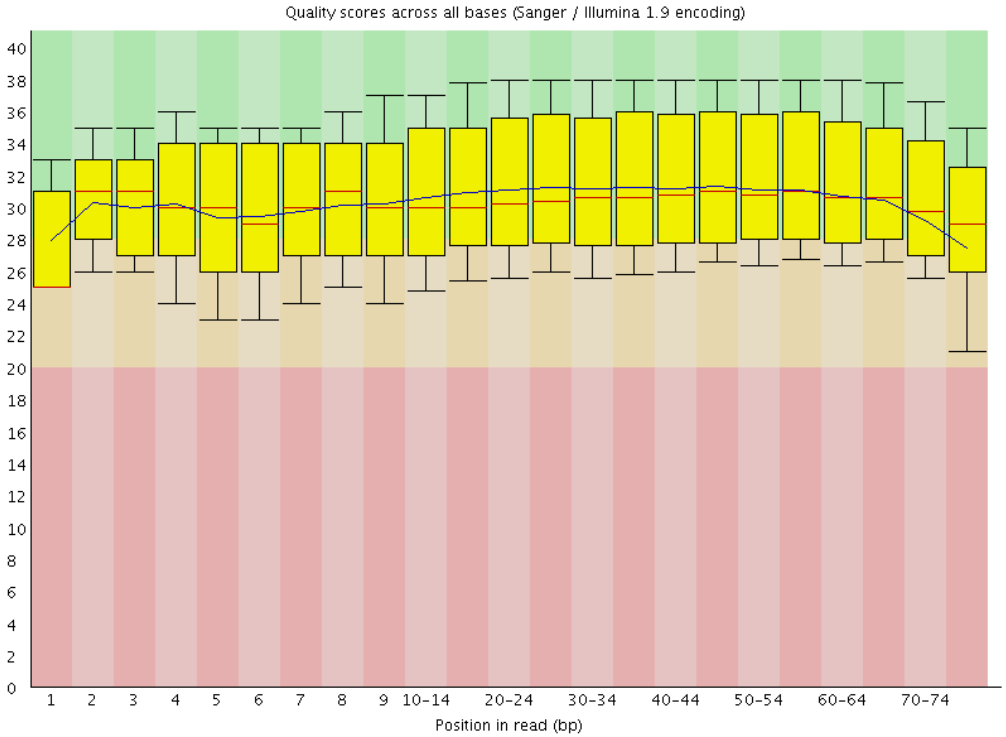
Requires a reference genome and a catalog of variable sites

Base quality recalibration

Before



After



Phred Quality score:

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

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

Jorge Jiménez
jjimeneza@cipf.es

Index

Jorge Jiménez

jjimeneza@cipf.es

SNP calling

SNP – single nucleotide polymorphisms/variant

- Examine the bases aligned to position and look for differences
- Sequence context of the SNP e.g. homopolymer run

Two steps:

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

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

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.

SNP calling

Factors to consider when calling SNPs

- Base call qualities of each supporting base
- Proximity to:
  - Small indel
  - Homopolymer run (>4-5bp for 454 and >10bp for illumina)
- Mapping qualities of the reads supporting the SNP
  Low mapping qualities indicates repetitive sequence
- Read length
- Paired reads
- Sequencing depth

Type of analysis:

- Variants present in a population
- Rare variants
- Somatic variants
- Pooled samples

Variant calling software

SNV callers

- GATK
- Samtools
- Beagle
- Soap2
- Impute 2
- Varscan2
- Strelka
- MuTect

Somatic

(...)

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.

- Indel realignment

- Uses standard input and output files.

- Many tools for manage VCF files.

- Multi-sample calling

http://www.broadinstitute.org/gatk/

Samtools

- Estimation of the most likely genotype.

- Manage of VCF and BAM files.

- 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.

- Uses standard input and output files.

- Multi-sample calling

http://samtools.sourceforge.net/

Variant quality score recalibration

Aim:
To assign a well-calibrated probability to each variant call in a call set.

The tool develops a continuous, covarying estimate of the relationship between SNP call annotations (QD, SB, Hrun, HaplotypeScore, for example) and the the probability that a SNP is a true genetic variant versus a sequencing or data processing artifact.

The model is determined adaptively based on "known sites" (HapMap 3 sites and Omni 2.5M SNP chip array) and evaluates the probability that each call is real.

The score that gets added to the INFO field of each variant is called the VQSLOD. It is the log odds ratio of being a true variant versus being false under the trained Gaussian mixture model.

Requires a reference genome and a catalog of variable sites

Jorge Jiménez
jjimeneza@cipf.es

Index

Jorge Jiménez

jjimeneza@cipf.es

Indel calling

Small insertions and deletions observed in the alignment of the read relative to the reference genome

- BAM format
  - I or D character in CIGAR denote indel in the read

Simple method

- Call indels based on the I or D events in the BAM file
  - Samtools varFilter

Factors to consider when calling indels

- Misalignment of the read
- Alignment scoring - often cheaper to introduce multiple SNPs than an indel
- Sufficient flanking sequence either side of the read
- Homopolymer runs either side of the indel
- Length of the reads
- Homozygous or heterozygous

Indel calling

Simple models for calling indels based on the initial alignments show high false positives and negatives

More sophisticated algorithms been developed

- E.g. Dindel, GATK

Example Algorithm overview

- Scan for all I or D operations across the input BAM file
- Foreach I or D operation
  - Create new haplotype based on the indel event
  - Realign the reads onto the alternative reference
  - Count the number of reads that support the indel in the alternative reference
  - Make the indel call

Very computationally intensive if testing every possible indel

Indel calling software

Indel callers

- GATK
- Samtools
- Dindel

- Varscan2
- GATK
- Strelka

Somatic

Index

Jorge Jiménez

jjimeneza@cipf.es

VCF: variant calling format

HEADER

Arbitrary number of meta-information lines
Starting with characters '##'
Column definition line starts with single '#'

DATA

- Chromosome (CHROM)
- Position of the start of the variant (POS)
- Unique identifiers of the variant (ID)
- Reference allele (REF)
- Comma separated list of alternate non-reference alleles (ALT)
- Phred-scaled quality score (QUAL)
- Site filtering information (FILTER)
- User extensible annotation (INFO)
- Information for sample (FORMAT)
- Values for sample (ID of sample)

VCF variant calling format

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Jorge Jiménez
jjimeneza@cipf.es

Index

Jorge Jiménez

jjimeneza@cipf.es

Structural variations

Large DNA rearrangements (>100bp)

Frequent causes of disease

- Referred to as genomic disorders
- Mendelian diseases or complex traits such as behaviors
- Prevalent in cancer genomes

Many types of genomic structural variation (SV)

Insertions, deletions, copy number changes, inversions, translocations & complex events

Comparative genomic hybridization (CGH) traditionally used to for copy number discovery

- CNVs of 1–50 kb in size have been under-ascertained

Next-gen sequencing revolutionised field of SV discovery

- Parallel sequencing of ends of large numbers of DNA fragments
- Examine alignment distance of reads to discover presence of genomic rearrangments
- Resolution down to ~100bp

Structural variations

Several types of structural variations

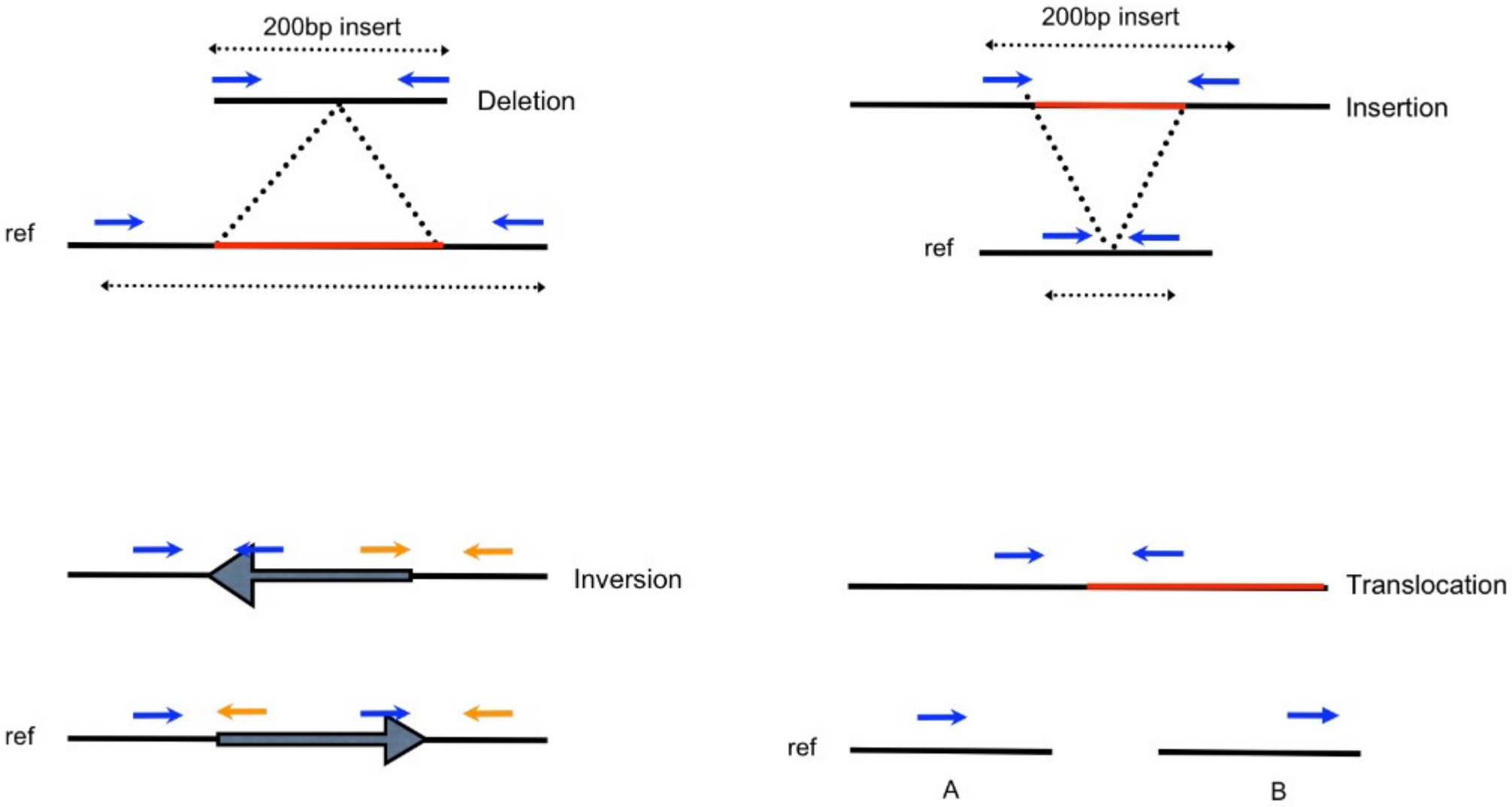
- Large Insertions/deletions
- Inversions
- Translocations
- Copy number variations

Read pair information used to detect these events

- Paired end sequencing of either end of DNA fragment
- Observe deviations from the expected fragment size
- Presence/absence of mate pairs
- Read depth to detect copy number variations
- Several SV callers published recently

Run several callers and produce large set of partially overlapping calls

Structural variations types



Structural variations software

Breakdancer

- Insertions, deletions, inversions, translocations

http://gmt.genome.wustl.edu/breakdancer/current/

Pindel

- Insertions and deletions

https://trac.nbic.nl/pindel/

Genome STRIP

- Calling across low coverage populations + genotyping

http://www.broadinstitute.org/gsa/wiki/index.php/Genome_STRiP

RetroSeq

- Mobile element insertion discovery

https://github.com/tk2/RetroSeq

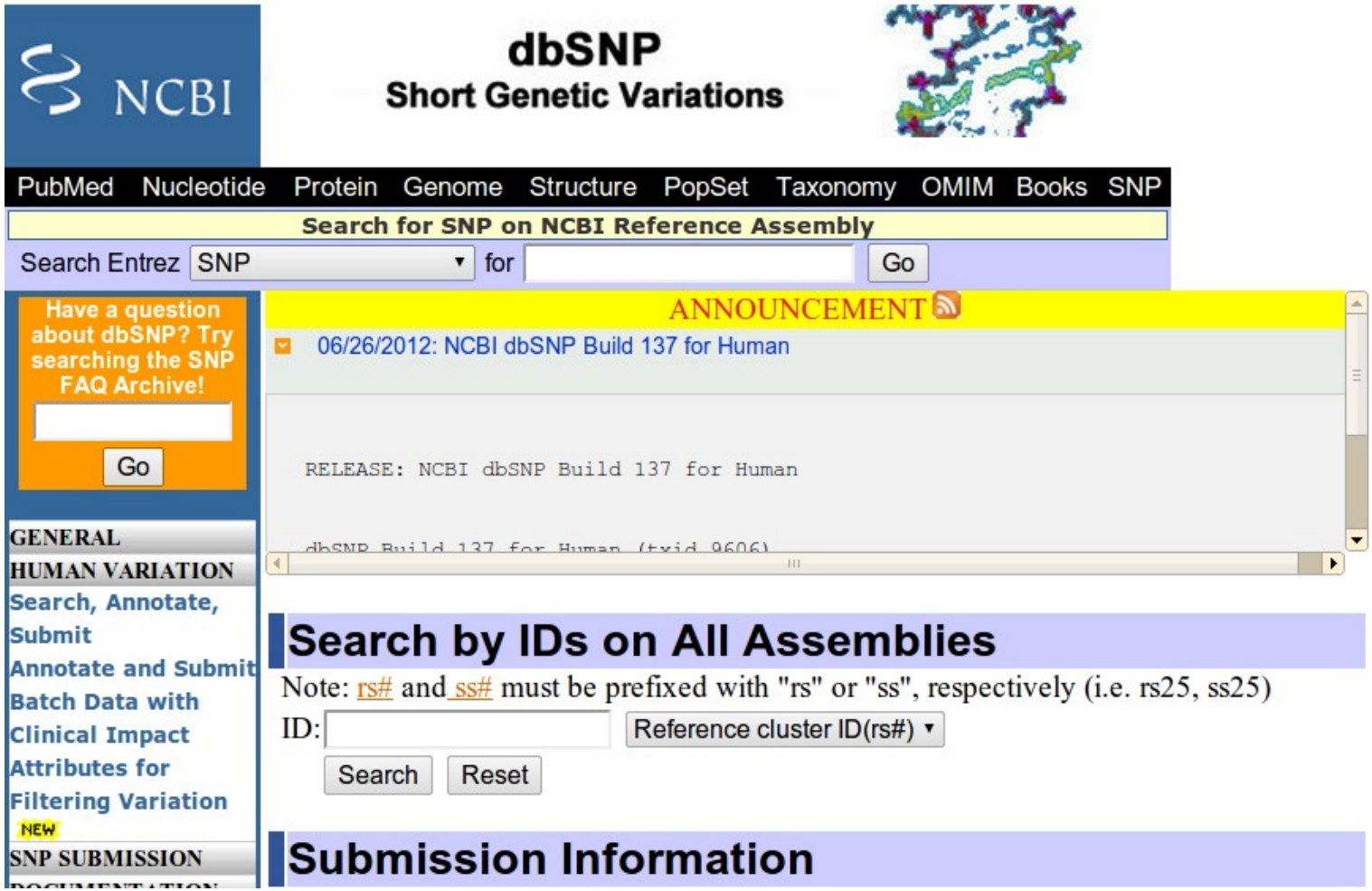
Breakpoint assembly

- Tigra: http://genome.wustl.edu/software/tigra_sv
- SVMerge: http://svmerge.sourceforge.net/

Integrating Calls

- SVMerge: http://svmerge.sourceforge.net/

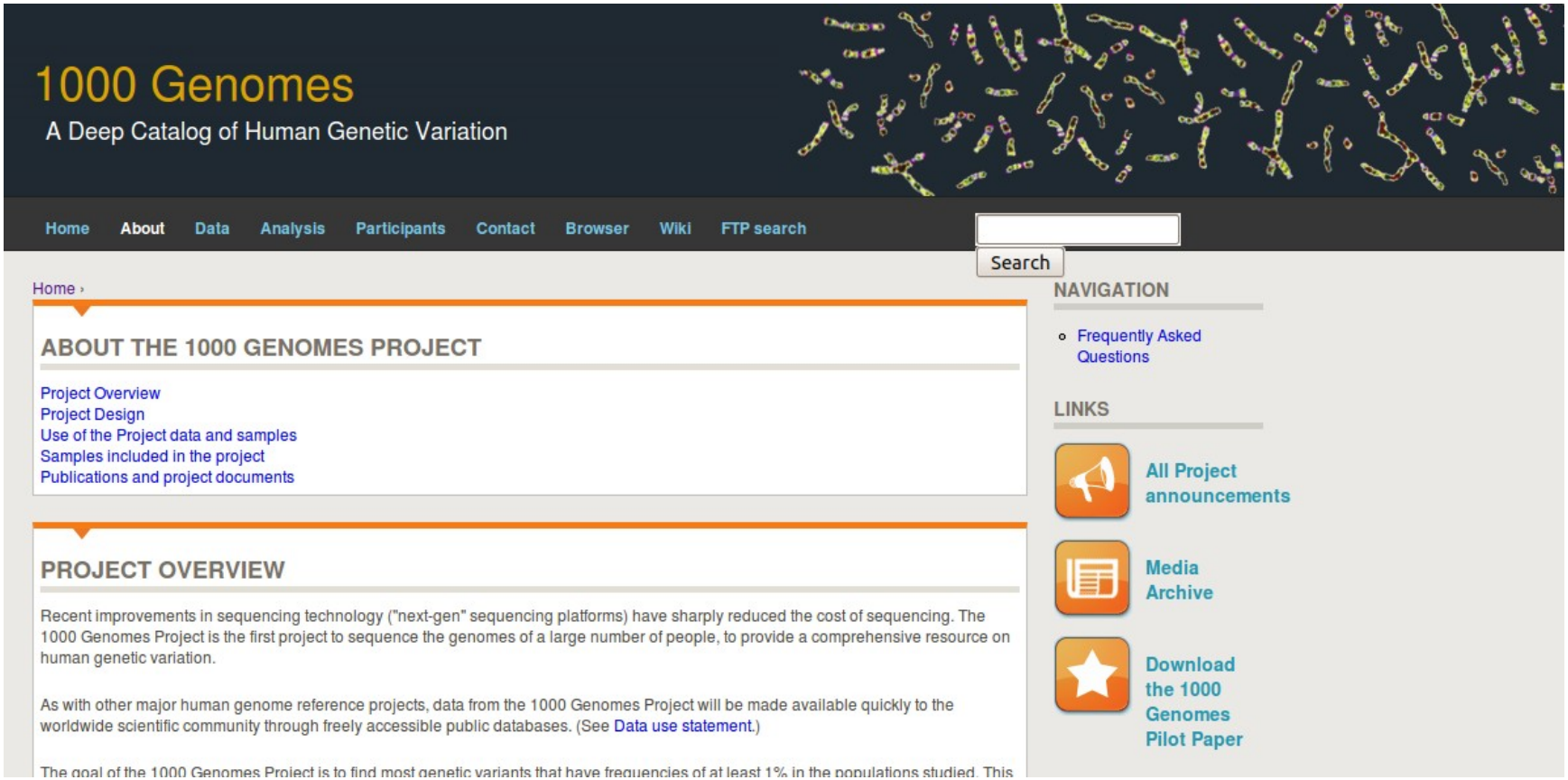
dbSNP



http://www.ncbi.nlm.nih.gov/projects/SNP/

Jorge Jiménez
jjimeneza@cipf.es

1000 genomes project



http://www.1000genomes.org/

Jorge Jiménez
jjimeneza@cipf.es

NHLBI Exome Sequencing project

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)

http://evs.gs.washington.edu/EVS/

Questions?

Jorge Jiménez
jjimeneza@cipf.es