NGS I : VARIANT DETECTION

Javier Perales-Patón
jperales@cnio.es

Fátima Al-Shahrour
[falshahrour@cnio.es]
Elena Piñeiro-Yáñez
[epineiro@cnio.es]
Pedro Fernandes
[pfern@igc.gulbenkian.pt]

cnio
Centro Nacional de Investigaciones Oncológicas

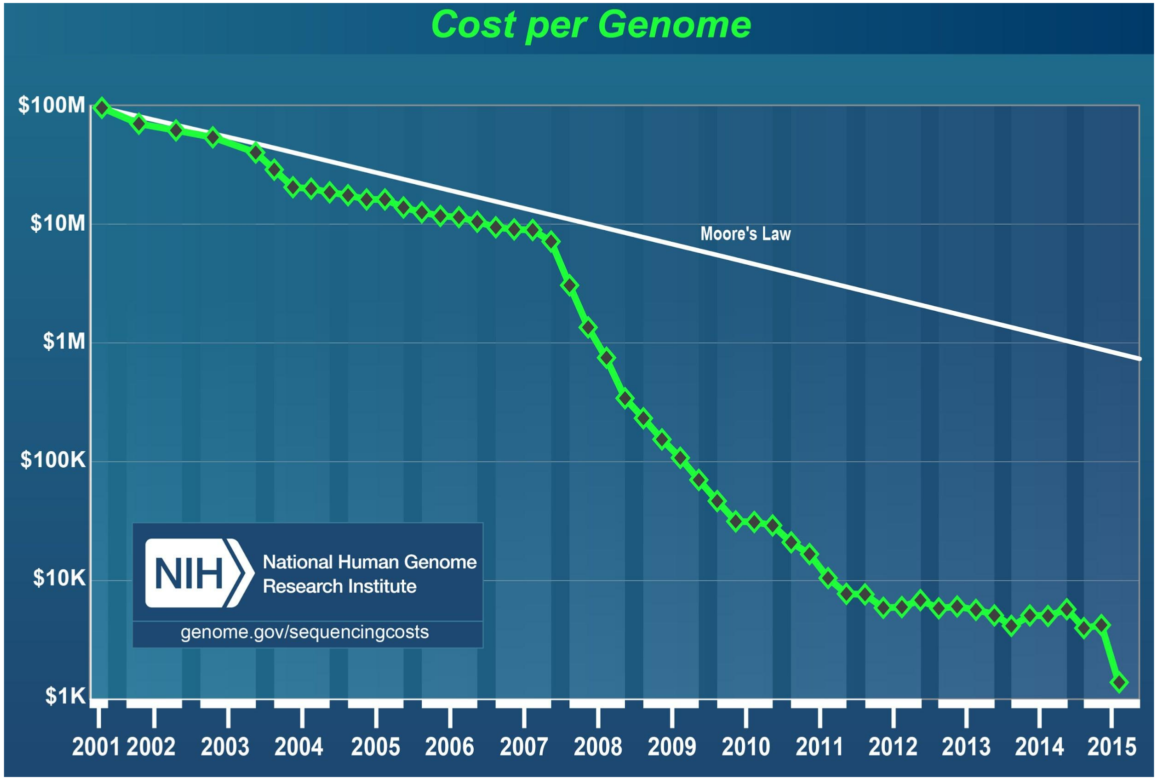
Bioinformatics Unit
CNIO. Madrid, Spain.

INSTITUTO GULBENKIAN DE CIÊNCIA

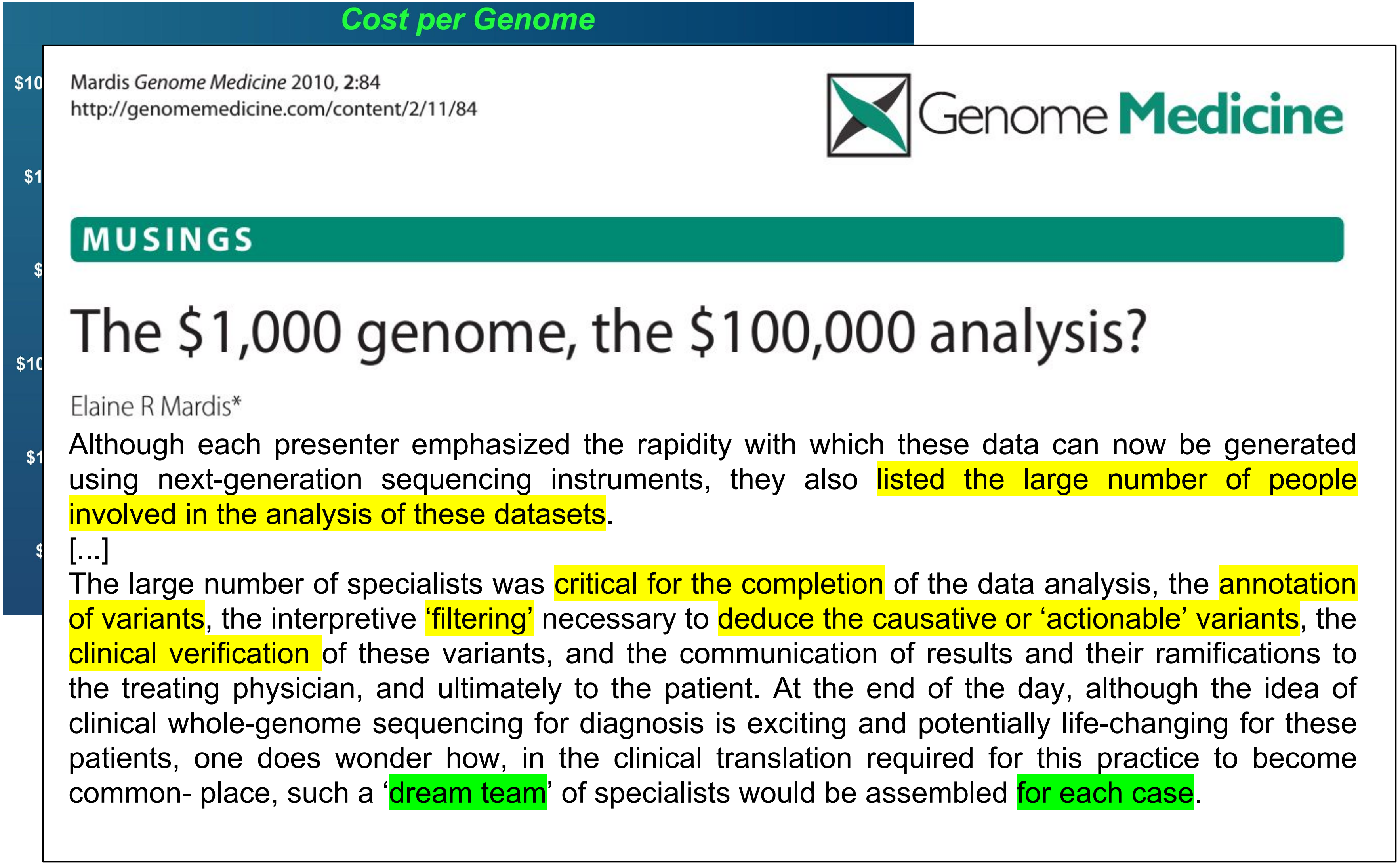
Today

09:30 - 10:00	Introduction to the course and self presentation of the participants. Personalized medicine.
11:30 - 12:30	**NGS I : Variant detection.**
14:00 - 16:00	Playing with the data and the methods.
16:30 - 18:00	Practical : Running the pipeline.

Sequencing cost has been coming down



Sequencing cost has been coming down



Cost per Genome

Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome **Medicine**

MUSINGS

The $1,000 genome, the $100,000 analysis?

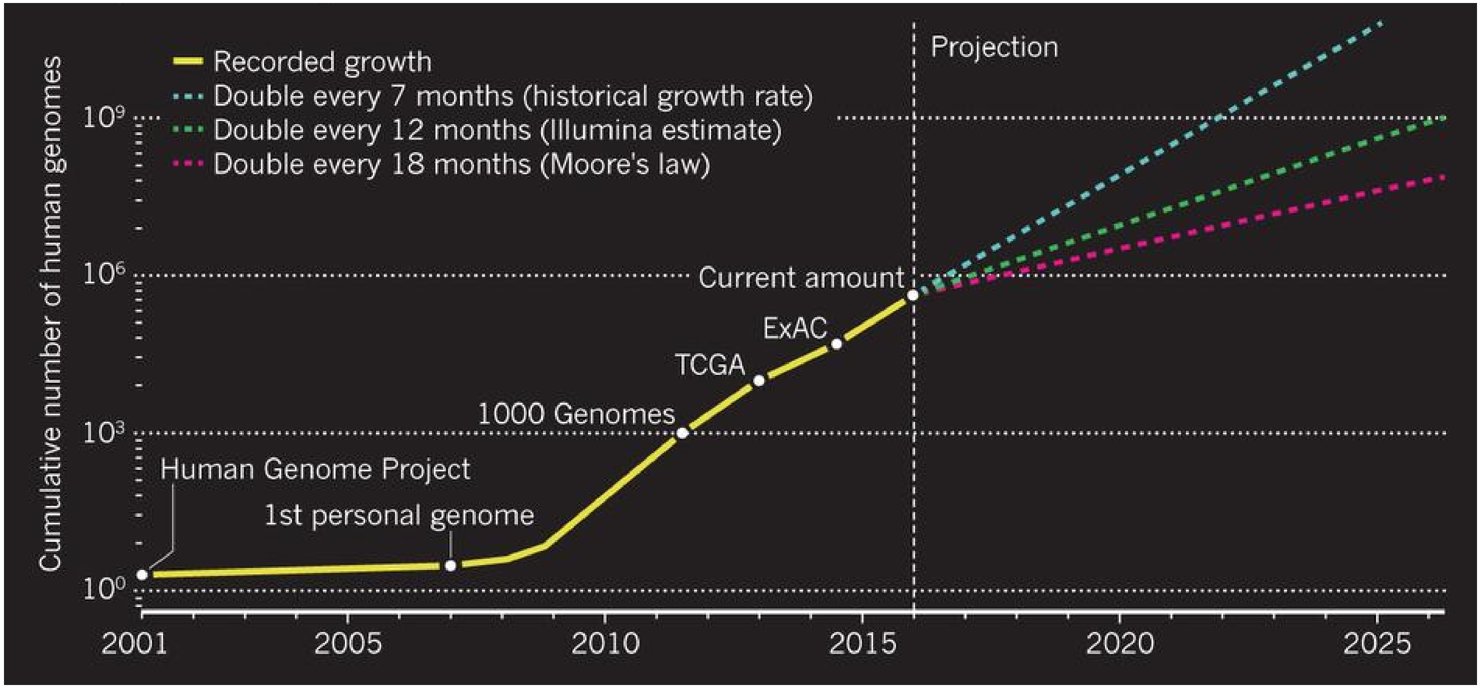
Elaine R Mardis*

Although each presenter emphasized the rapidity with which these data can now be generated using next-generation sequencing instruments, they also listed the large number of people involved in the analysis of these datasets.

[...]

The large number of specialists was critical for the completion of the data analysis, the annotation of variants, the interpretive ‘filtering’ necessary to deduce the causative or ‘actionable’ variants, the clinical verification of these variants, and the communication of results and their ramifications to the treating physician, and ultimately to the patient. At the end of the day, although the idea of clinical whole-genome sequencing for diagnosis is exciting and potentially life-changing for these patients, one does wonder how, in the clinical translation required for this practice to become common- place, such a ‘dream team’ of specialists would be assembled for each case.

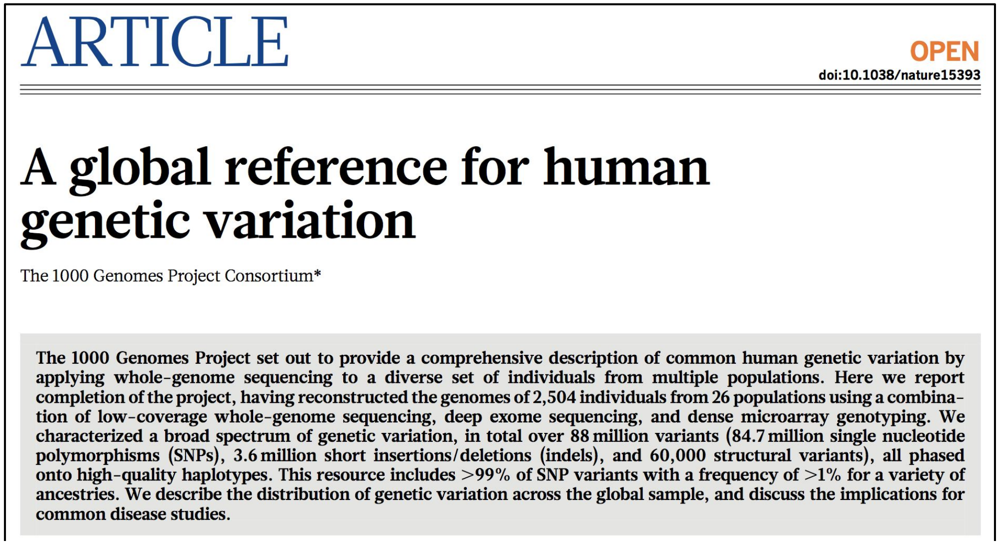
DNA sequencing soars



+ 1000 Genomes Project : hundreds of genomes.
+ TCGA : thousands (genome & exomes).
+ ExAC : > 60,000 exomes.

Stephens ZD *et al.* **Big Data: Astronomical or Genomical?**. PLoS Biol. 2015 Jul 7;13(7)
Eisenstein M. **Big data: The power of petabytes**. Nature. 2015 Nov 5;527(7576):S2-4.

1000 Genomes Project



ARTICLE

OPEN

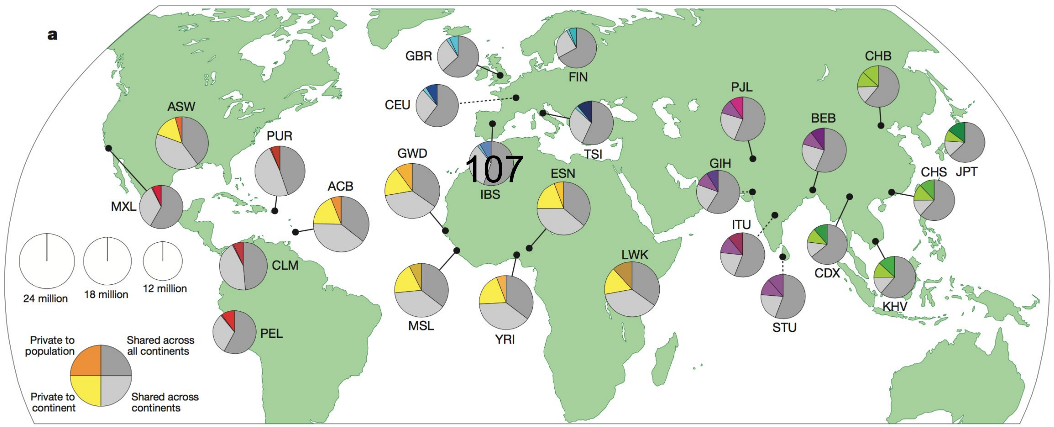
doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
END OF THE BEGINNING
Final phase of 1000 Genomes Project maps human genetic variation in open-access resource PAGES 52, 68 & 75



Phase 3	WGS	WExS
Raw bases	89 Tb	18 Tb
Samples	2,504	2,504
Region	Genome	Exome
Mean Depth	8.45x	75x
SNPs	85M	1.5M
Indels	3.6M	22K
Structural Variants	60K	6.5K
Het. Concordance (SNPs)	99.4%	99.8%

http://www.1000genomes.org/about#ProjectSamples ; Phase 1 n=1092 → Phase 3 n=2504

The objective of the Variant Detection:

Identify the most likely **genotype for each genomic position from the individual**.

- - -

In Cancer genomics, if there is a **matched-normal sample** to be compared against the tumour sample:

+ Identify somatic variants (i.e. only in tumour sample).

+ Identify copy-number alterations (large genomic aberrations).

Some concepts

- What is a genetic **variant** ?

  Genetic differences in individuals as compared to a reference genome (built from a population).

  Nomenclature:

  - **First level**: Genomic position and nucleotide change.

    Chromosome Name:Genomic position (coordinates):Reference allele>Alternative allele

    chr12:25398284-25398284:C>T

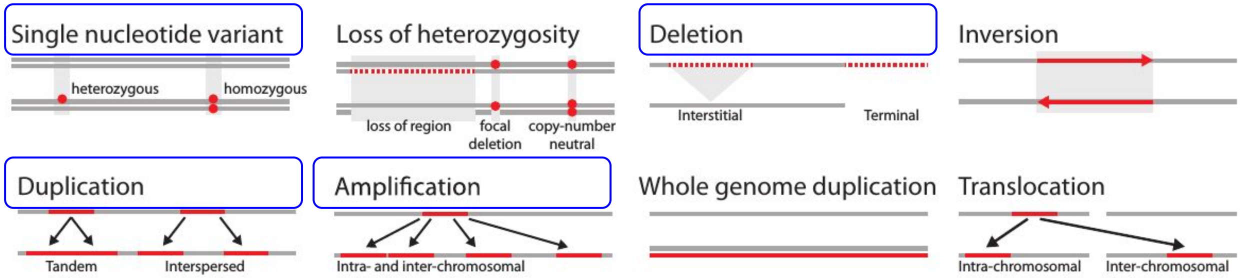
- Classes of variants :
  - **Germline** : inherited. E.g. a SNP, or a SNV related to a rare disease.
  - **Somatic** : acquired within a cell lineage. E.g. Cancer mutations.

- Polymorphism

  common variant in a given population (SNP, Single Nucleotide Polymorphism).

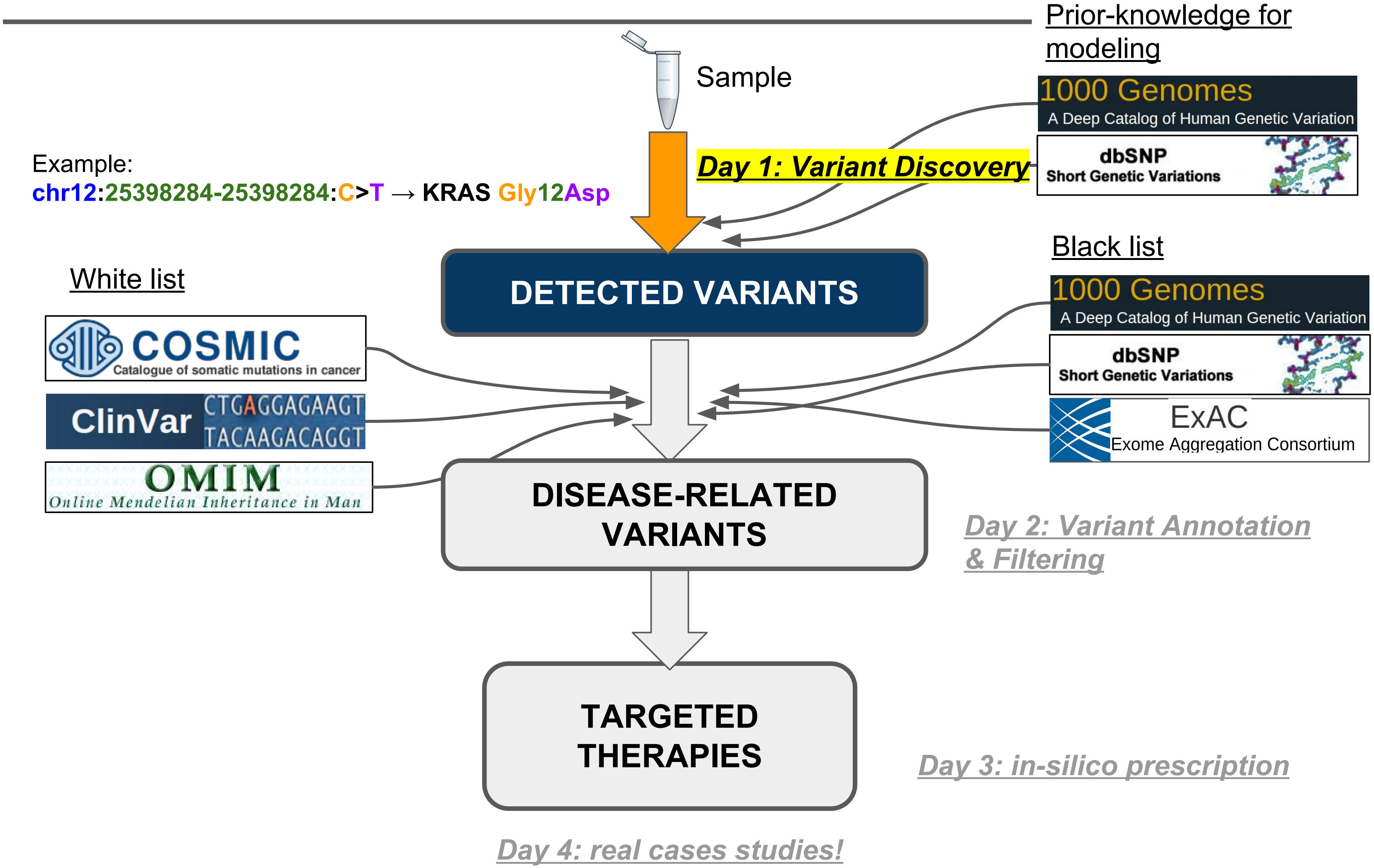
  Present in at least 1% in a population.

- Types of genomic variants:



Variants need a context

Example:
chr12:25398284-25398284:C>T → KRAS Gly12Asp



Sample

Prior-knowledge for modeling

1000 Genomes
A Deep Catalog of Human Genetic Variation
dbSNP
Short Genetic Variations

Day 1: Variant Discovery

DETECTED VARIANTS

White list

COSMIC
Catalogue of somatic mutations in cancer
ClinVar
CTGAGGAGAAGT
TACAAGACAGGT
OMIM
Online Mendelian Inheritance in Man

Black list

1000 Genomes
A Deep Catalog of Human Genetic Variation
dbSNP
Short Genetic Variations
ExAC
Exome Aggregation Consortium

DISEASE-RELATED VARIANTS

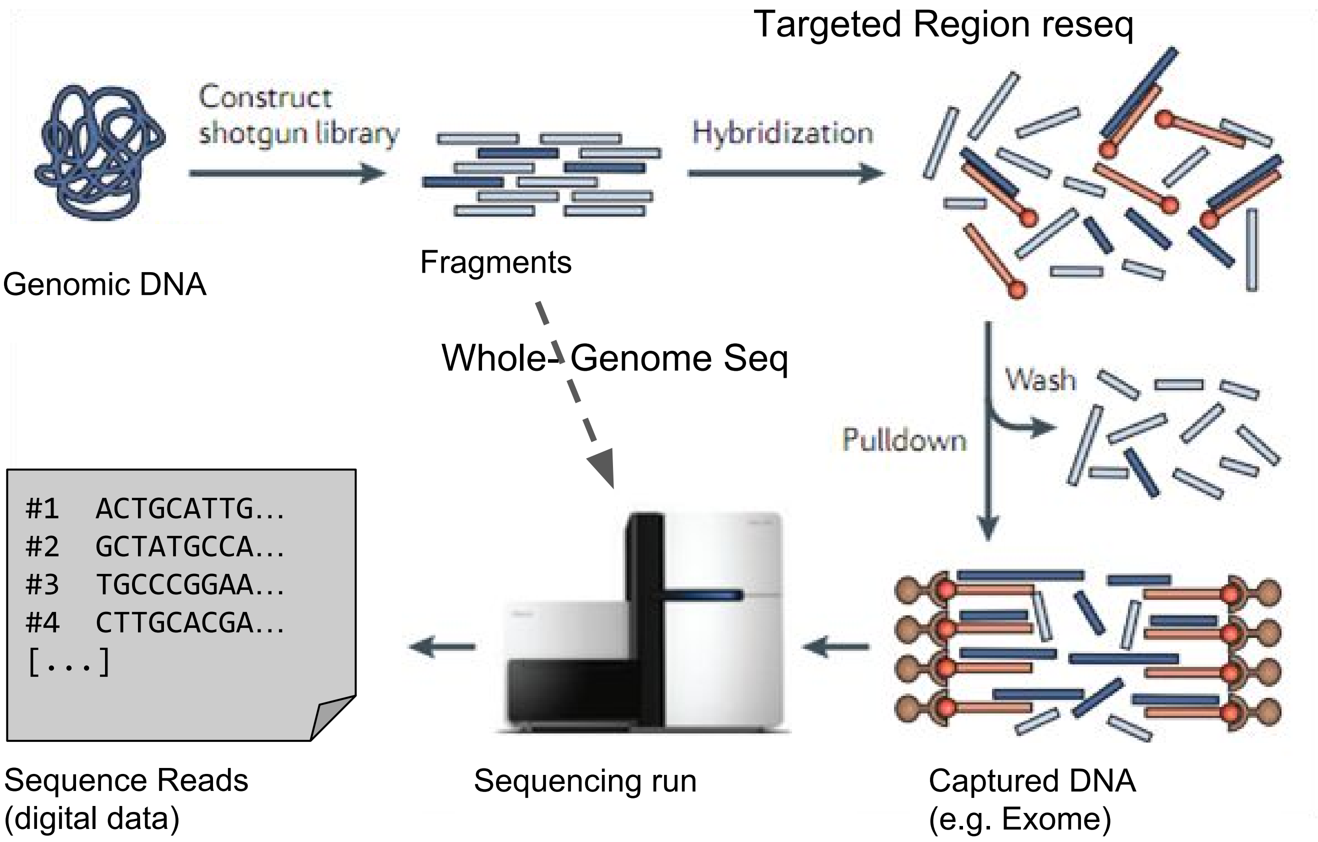
Day 2: Variant Annotation & Filtering

TARGETED THERAPIES

Day 3: in-silico prescription

Day 4: real cases studies!

DNA Sequencing data generation



Targeted Region reseq

Genomic DNA

Fragments

Whole- Genome Seq

```
#1  ACTGCATTG...
#2  GCTATGCCA...
#3  TGCCCGGAA...
#4  CTTGCACGA...
[...]
```

Sequence Reads
(digital data)

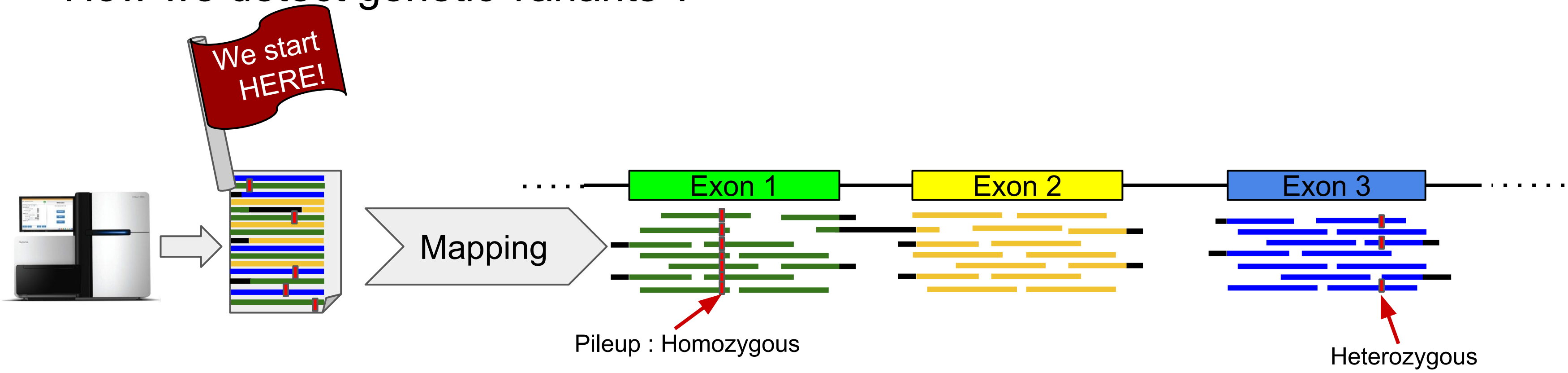
Sequencing run

Captured DNA
(e.g. Exome)

The sequence reads belong from the ends of the original fragment.

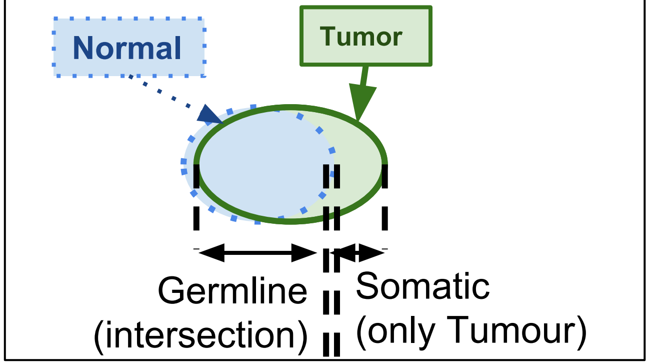
Fundamentals of variant detection

- How we detect genetic variants ?

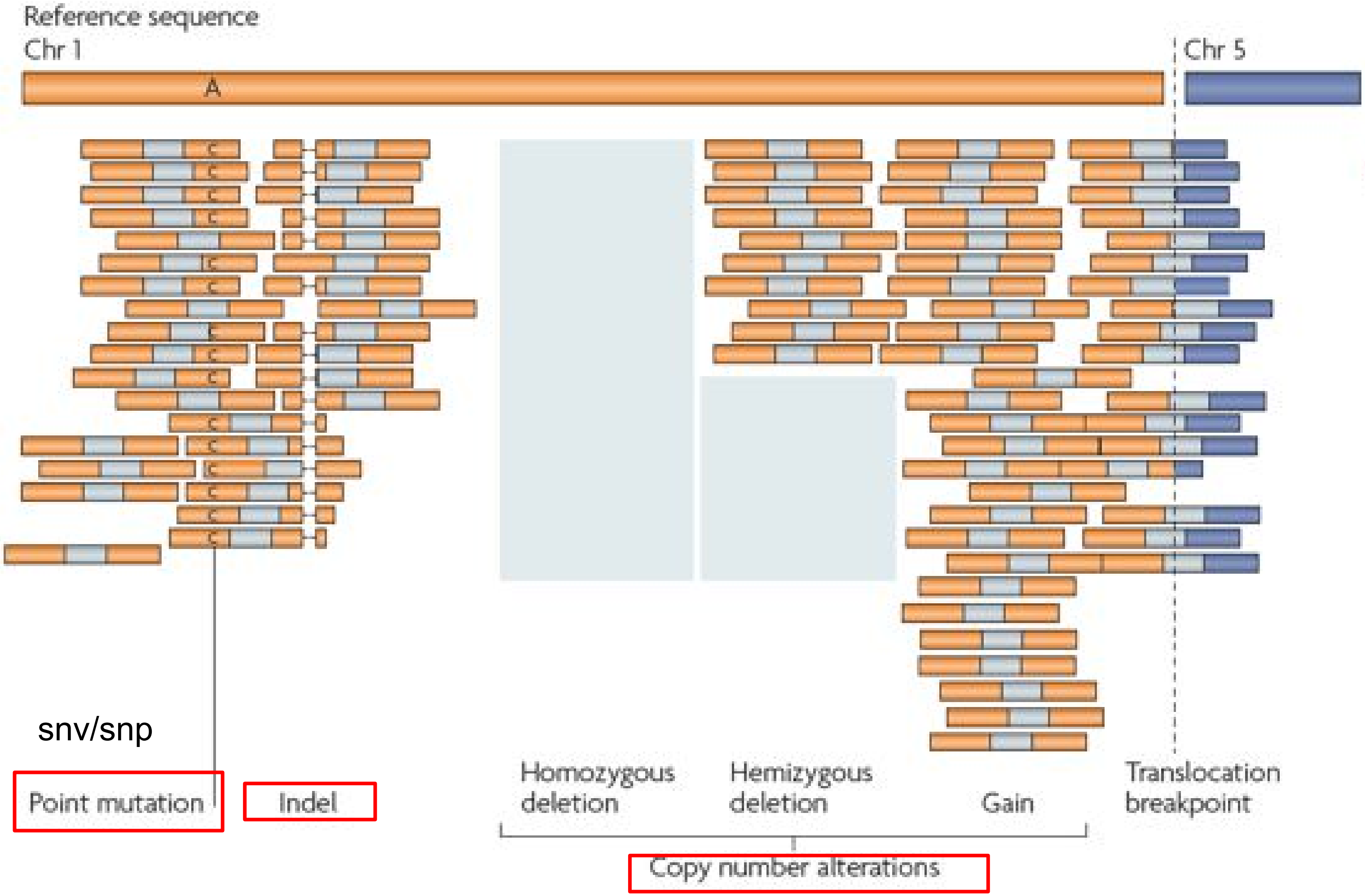


Are these differences true calls?
Statistical method to estimate the most likely genotype.

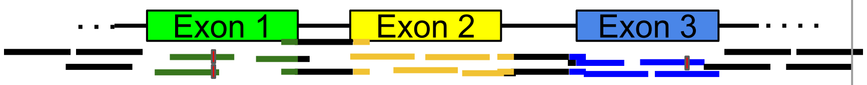
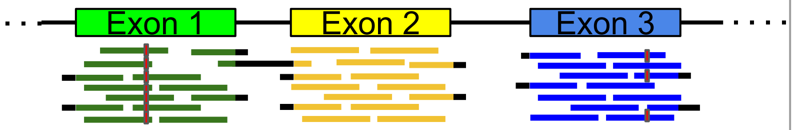
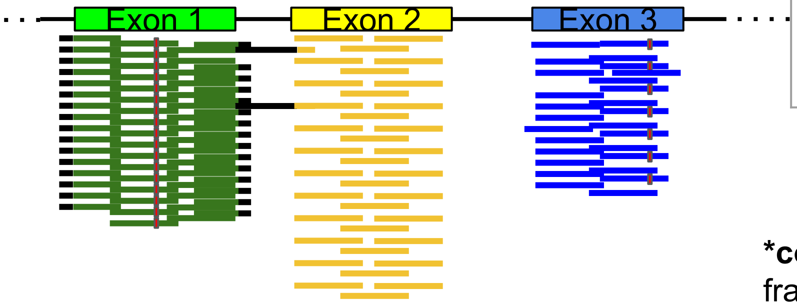
Discern somatic mutations by comparison:



Different types of variants detected by mapping reads

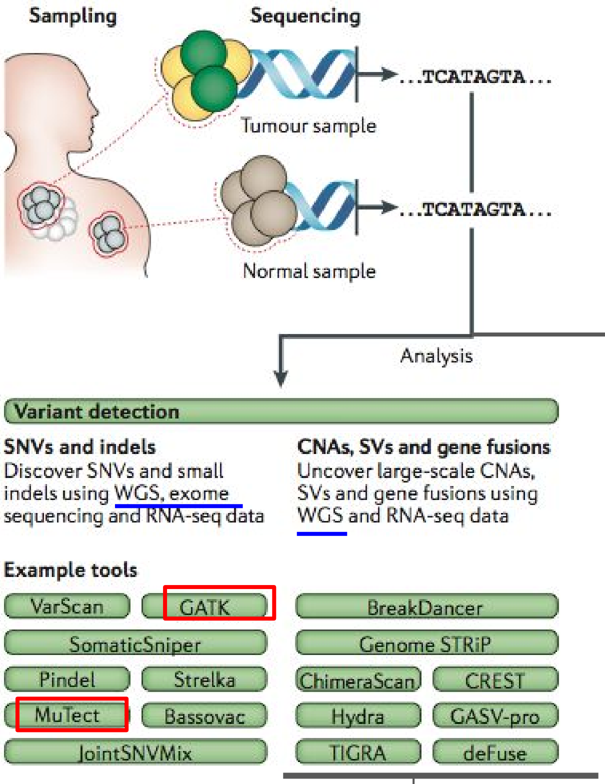


Whole-Genome, and targeted resequencing

	# bp pos seq	**Type of variants discovered**	**Avg Coverage per pos**	**Cost**
Whole-Genome Sequencing 	~100 Gb	**- coding variants*,** intronic and regulatory sites. - **Structural variants** - **CNA** #Variants= 3M - 4M.	30x	High
Whole-Exome Sequencing 	~32Mb 50Mb	- **coding variants***. - Some intronic and regulatory sites. - **CNA** (challenging). **#**Variants= 20k - 60k.	20x - 80x	Low
Panel of genes by amplicon/PCR approach 	*ND*	Depends on the design - Particular **coding variants*** - **CNA** (challenging) # variants = *ND*	1000x - 5000x	Low

***coding variants**: missense, stop gained, stop lost, frameshift, splice region...

Methods for Variant Detection



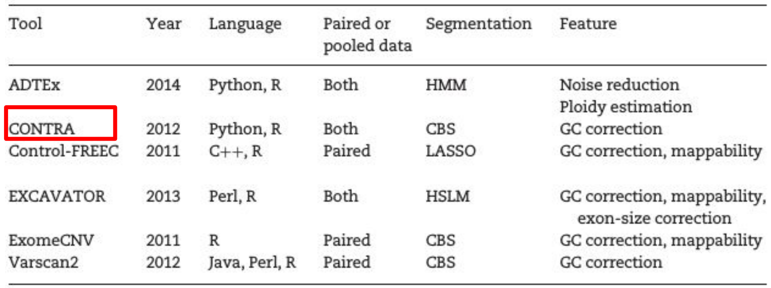
SNVs and indels
Discover SNVs and small indels using WGS, exome sequencing and RNA-seq data

CNAs, SVs and gene fusions
Uncover large-scale CNAs, SVs and gene fusions using WGS and RNA-seq data

Example tools

Several Methods have been published.

More on CNAs



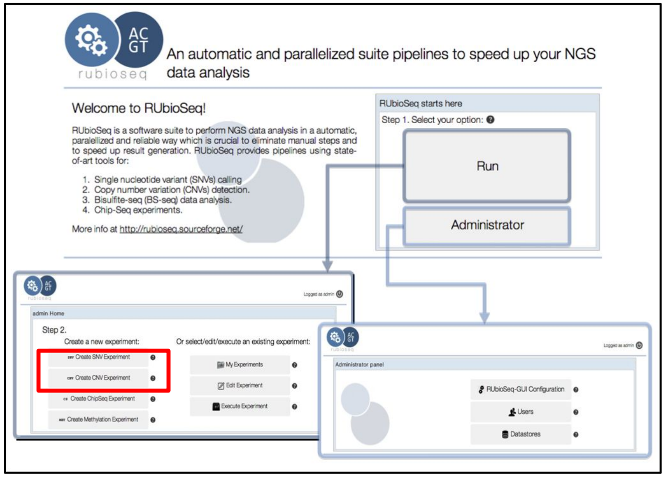
Tool	Year	Language	Paired or pooled data	Segmentation	Feature
ADTEx	2014	Python, R	Both	HMM	Noise reduction Ploidy estimation
CONTRA	2012	Python, R	Both	CBS	GC correction
Control-FREEC	2011	C++, R	Paired	LASSO	GC correction, mappability
EXCAVATOR	2013	Perl, R	Both	HSLM	GC correction, mappability, exon-size correction
ExomeCNV	2011	R	Paired	CBS	GC correction, mappability
Varscan2	2012	Java, Perl, R	Paired	CBS	GC correction

Only WGS

Appropriate methods for Whole-Exome seq

Further reading:
Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Nat Rev Genet – (2014). doi:10.1038/nrg3767
Nam J.N. et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics (2015)

Our proposed Variant Calling Pipeline



An automatic and parallelized suite pipelines to speed up your NGS data analysis

Welcome to RUbioSeq!

RUbioSeq is a software suite to perform NGS data analysis in a automatic, paralellized and reliable way which is crucial to eliminate manual steps and to speed up result generation. RUbioSeq provides pipelines using state-of-art tools for:

1. Single nucleotide variant (SNVs) calling
2. Copy number variation (CNVs) detection.
3. Bisulfite-seq (BS-seq) data analysis.
4. Chip-Seq experiments.

More info at http://rubioseq.sourceforge.net/

RUbioSeq starts here

Step 1. Select your option:

Logged as admin

admin Home

Step 2.

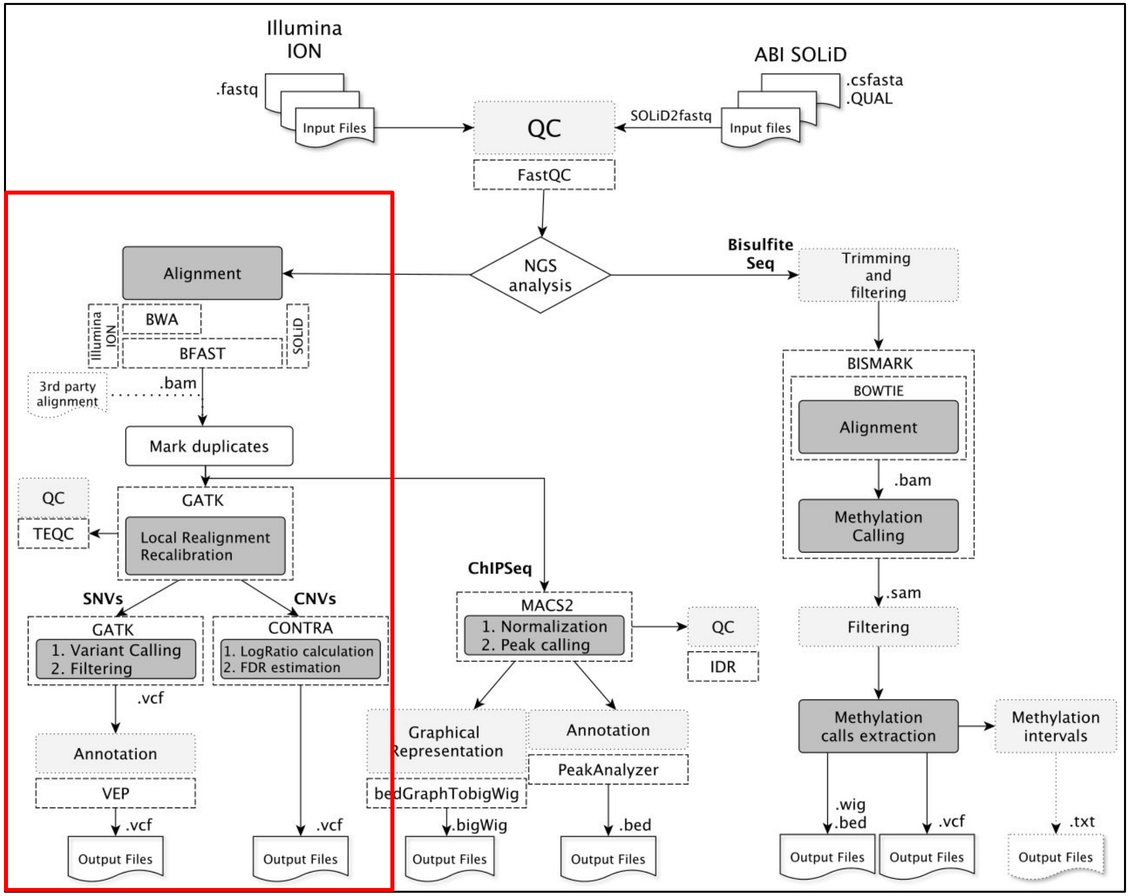
Create a new experiment:

Or select/edit/execute an existing experiment:

Administrator panel

Graphic User Interface

http://rubioseq.bioinfo.cnio.es/



workflow schema

Developed by the ***Bioinformatics Unit*** at the **Spanish National Cancer Research Centre** (Madrid, Spain).

Rubio-Camarillo *et al*. Comput Methods Programs Biomed. 2017 Jan;138:73-81

Rubio-Camarillo *et al*. Bioinformatics (2013) 29 (13), 1687-1689

What is Crucial in Variant calling

- For clinical practices, the use of **gold standard methods** and **reproducible analysis** are mandatory.

- The analysis is based on the comparison against the **reference genome** :
  A single consensus sequence for the whole genome. It was built up from a high quality set of representative samples of the specie (from different populations).
  It is the first-line comparison during analysis.

  By **Genome Reference Consortium (GRC)** (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/)

  - **Human assemblies (Versions):**
    + `GRCh37/hg19` : former version. Released in 2012. It is still the preference for analysis.
    + `CRCh38/hg38` : current version (Sep. 2017). Released in 2014. More accurate, comprehensive (includes Haplotypes) and sophisticated.

  "CRCh38 is here now, but still waiting."

- We must know what **regions along the genome** were sequenced in the experiment ? that is, the Sequencing library.

Bundle of files for Variant Detection

1. **Raw sequencing data** from the patient's sample.
2. **Genome Reference** (standard 1000 Genomes, fasta).
3. List of **Target beats or intervals** of genomic regions sequenced by the Library protocol.

4. **dbSNP** (VCF file) for a recent dbSNP release (build 138, it includes the 1000 Genomes).
5. HapMap genotypes and sites VCFs
6. **OMNI 2.5 genotypes for 1000 Genomes samples** (VCF).
7. The current best set of **known indels** to be used for local realignment); use both files:
   - 1000G_phase1.indels.b37.vcf (currently from the 1000 Genomes Phase I indel calls)
   - Mills_and_1000G_gold_standard.indels.b37.sites.vcf

Q: How you can get this bundle of files?

A: you could get them from the **Broad Institute's FTP**

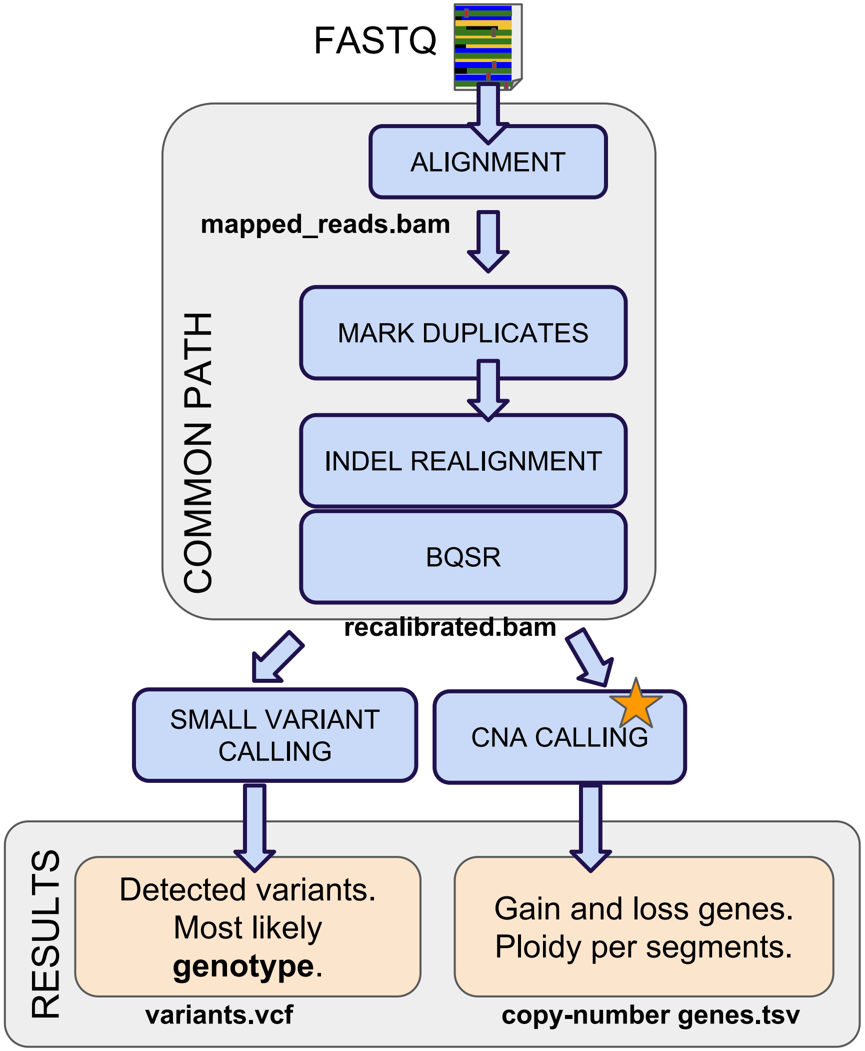
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/

You also need the Intervals file from your NGS provider (Illumina, Ion Torrent,...)

For this workshop, we got these files for you.

More info.:
https://software.broadinstitute.org/gatk/download/bundle

Point mutations and CNV Calling Process

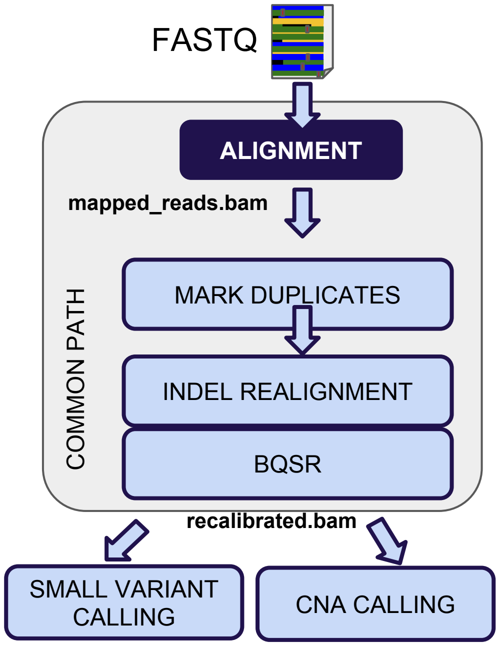


★ It requires **tumour and matched normal** sample (or a **panel of normals)**

Variant Calling Pipeline

1. Alignment

WORKFLOW:


FASTQ

COMMON PATH

ALIGNMENT

mapped_reads.bam

MARK DUPLICATES

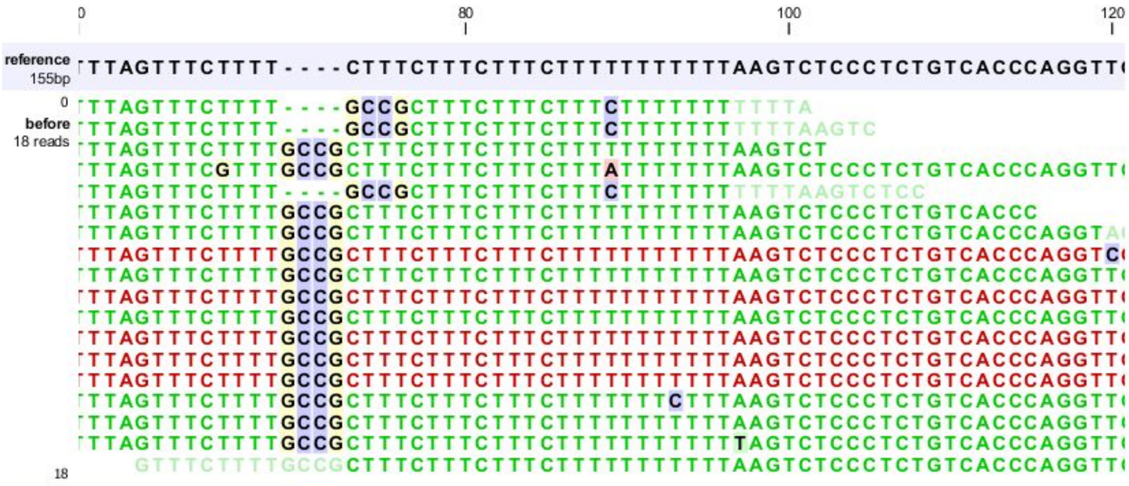
INDEL REALIGNMENT

BQSR

recalibrated.bam

SMALL VARIANT CALLING

CNA CALLING


```
           0                              80                  100                 120
reference  TTAGTTTCTTTT----CTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
155bp
0          TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTA
before     TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTC
18 reads   TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCT
           TTAGTTTCGTTTGCCGCTTTCTTTCTTTCTTTATTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTA
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTCTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTTAGTCTCCCTCTGTCACCCAGGTT
18             GTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
```

- Fast mapping on the reference genome by creating indexes. It is computationally intensive, but it is done only once.

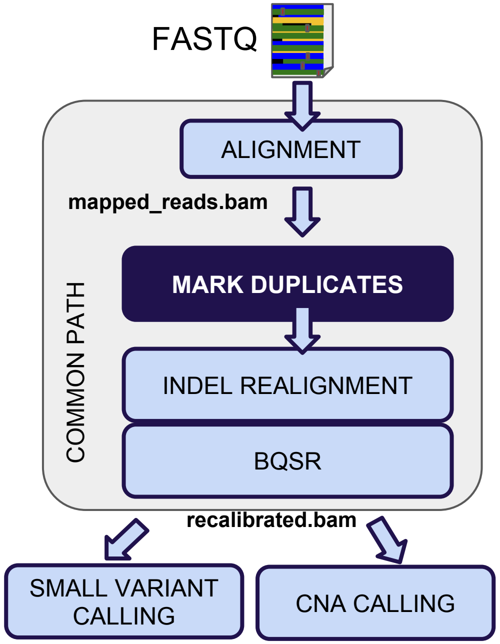
- Search for candidate sites to align a given read by using seeds (fragments of a read).

METHOD: by **BWA & Bfast+BWA**
https://github.com/lh3/bwa#citing-bwa
http://sourceforge.net/projects/bfast/files/bfast%2Bbwa/

Variant Calling Pipeline

2. Mark duplicates

WORKFLOW:

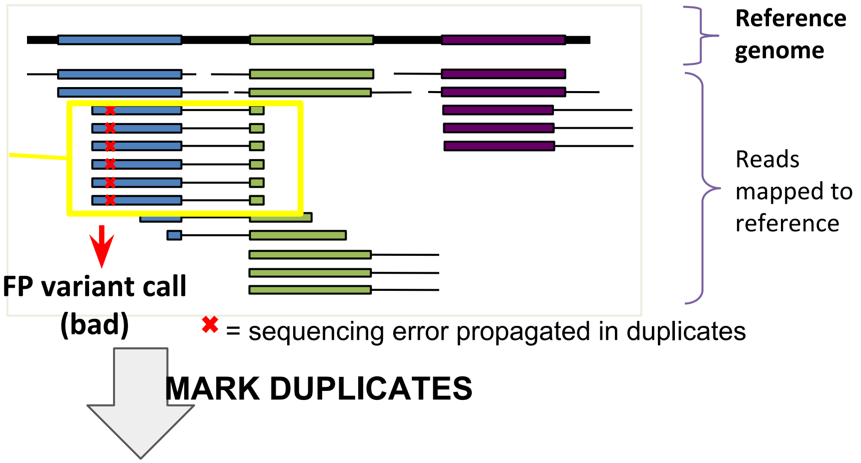


Under the hood:

- Duplicates derive from PCR amplification (library preparation): one fragment is sequenced multiple times.
- An error at the beginning of the PCR (first steps) is propagated.

- Therefore, duplicates are **worthless** for the analysis: *Duplicates are source of False Positives calls while only provide redundancy.*

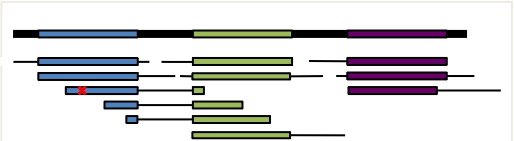
Solution: retrieve the best one, discard the duplicates:



Duplicates share the **same alignment properties** : sequence, start and end positions

✖ = sequencing error propagated in duplicates

After marking duplicates, the variant caller will only see :



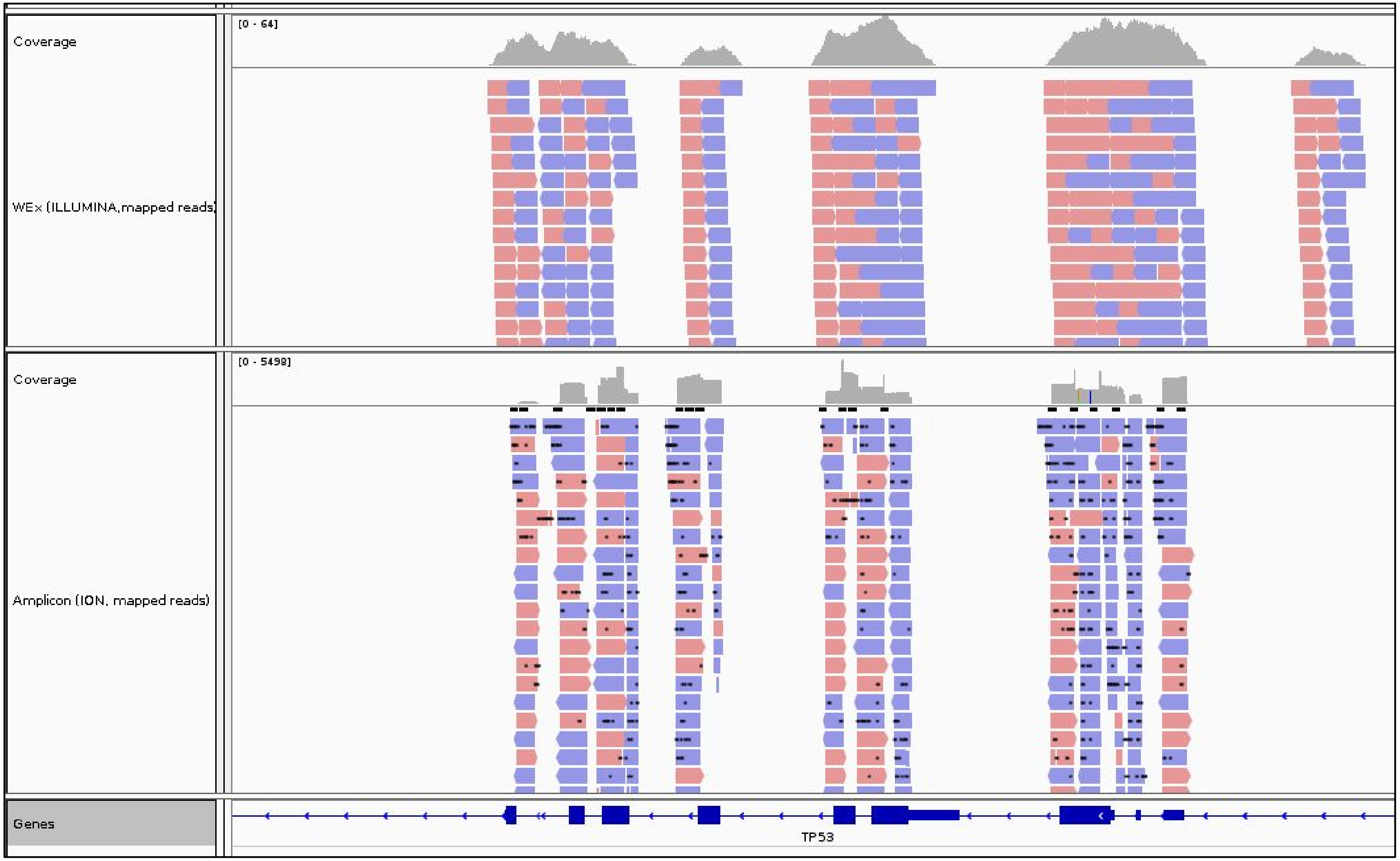
… and thus be more likely to make the right call

METHOD: by **Picard-tools**
http://broadinstitute.github.io/picard/
(alternatives : **samtools**)

Adapted from GATK

Variant Calling Pipeline

2. Mark duplicates: WEx Vs. Amplicon

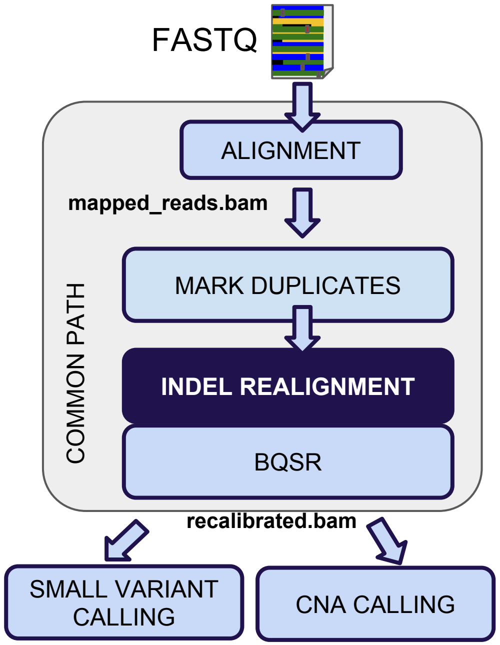


WARNING: **Do NOT mark duplicates in data derived from** amplicon techniques (**Ion Torrent**).
More info.: http://gatkforums.broadinstitute.org/discussion/5847/remove-duplicates-from-targetted-sequencing-using-amplicon-approach

Variant Calling Pipeline

3. Indel realignment

WORKFLOW:



Algorithms align reads very fast with high accuracy, but not perfectly.

> *During alignment, penalties on mismatches are much cheaper than gaps (indels). Aligners will tend to choose Mismatches at the beginning, and locate indels in the rest.*

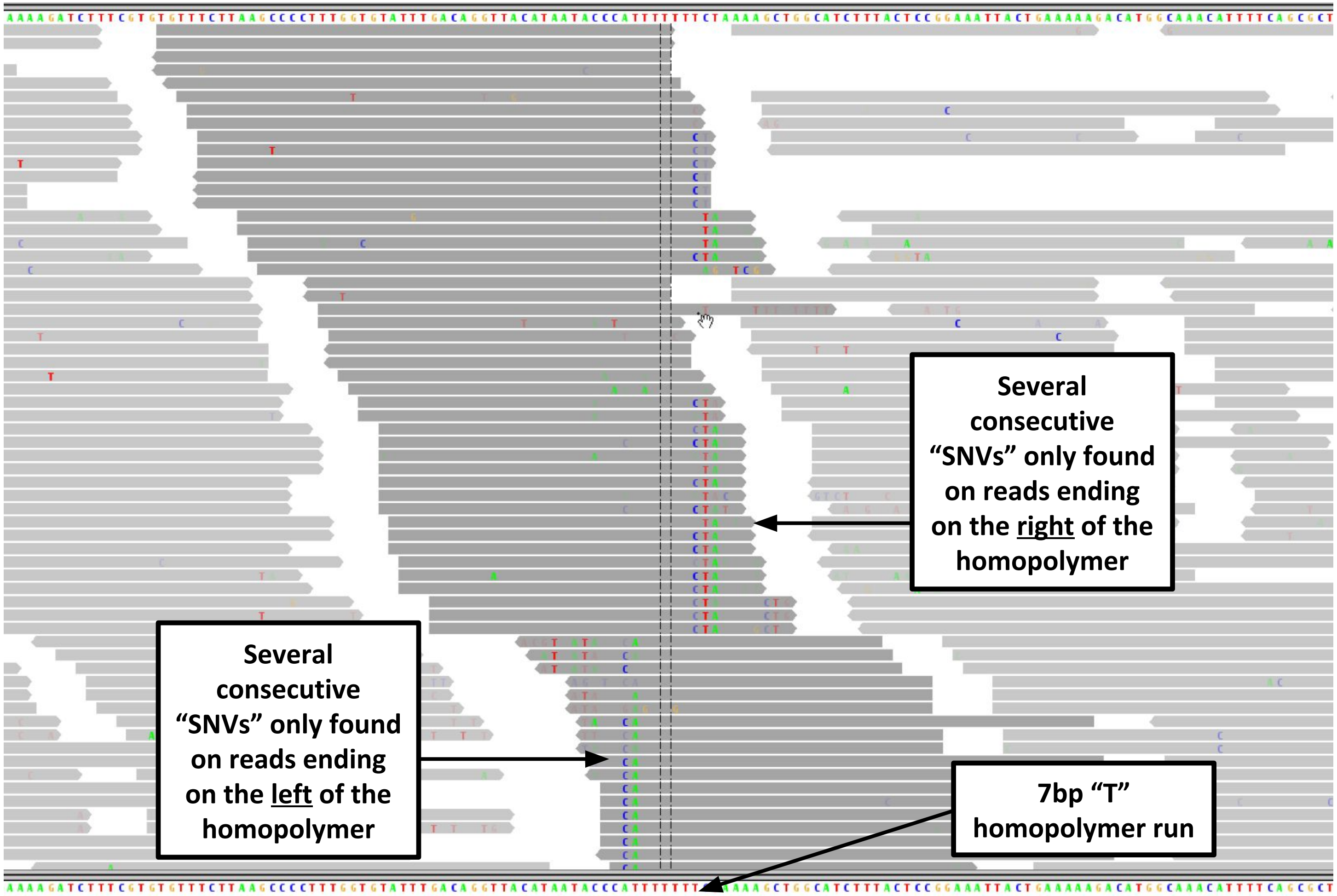
Also, there are sometimes multiple solution (alignments) for a given read. Aligners choose one randomly.

Variant calling requires the most perfect alignment as possible to avoid False Positives.

METHOD: by **GATK**
https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php

Variant Calling Pipeline

3. Indel realignment



Several consecutive "SNVs" only found on reads ending on the right of the homopolymer

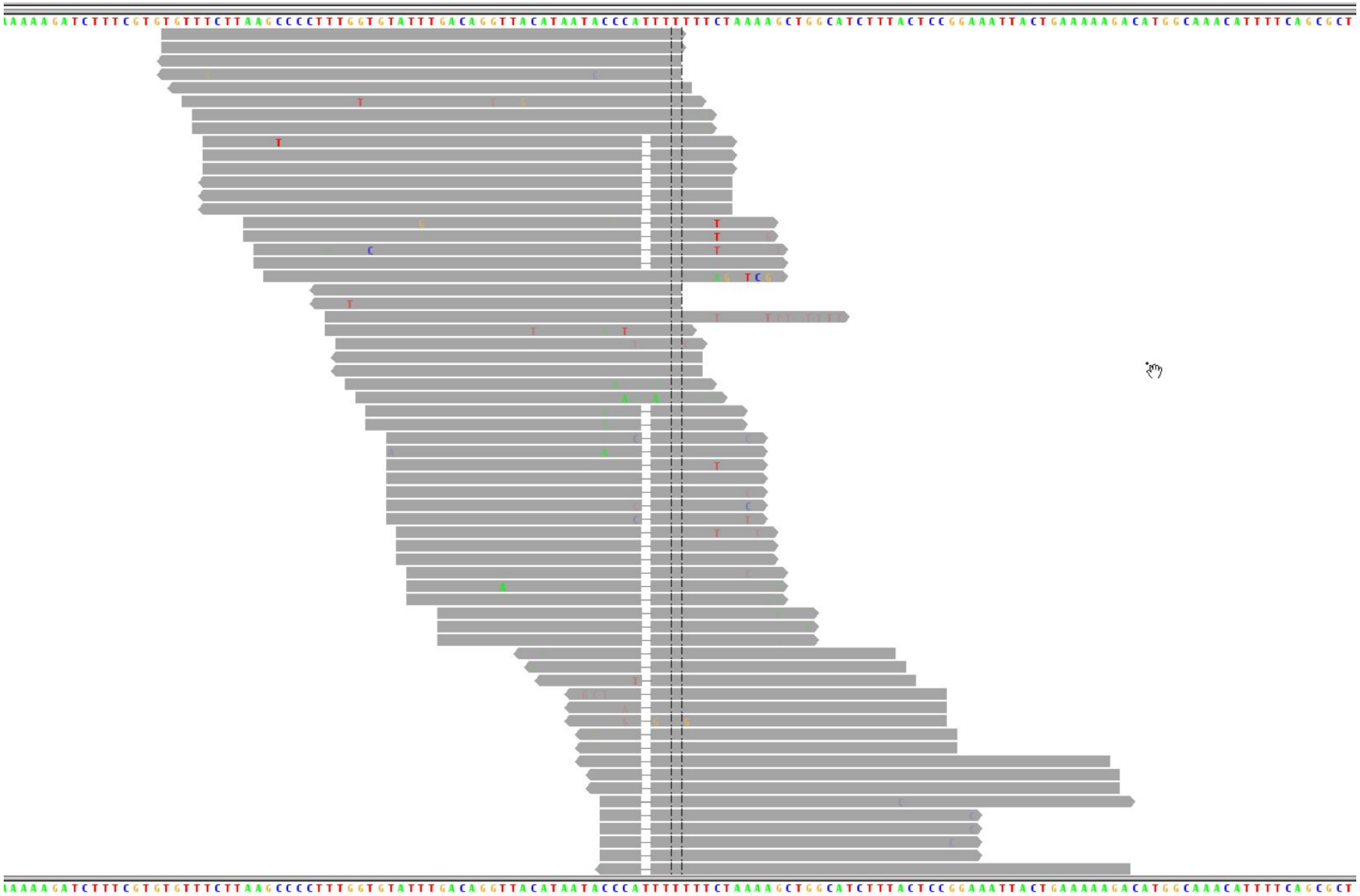
Several consecutive "SNVs" only found on reads ending on the left of the homopolymer

7bp "T" homopolymer run

Taken from GATK team

Variant Calling Pipeline

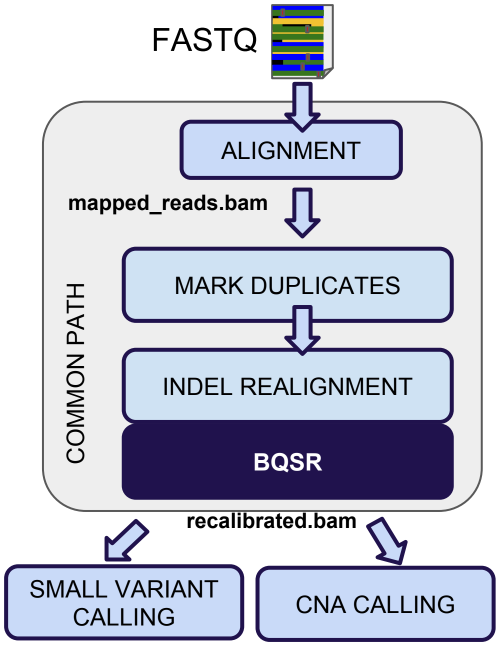
3. Indel realignment


AAAAGATCTTTCGTGTGTTTCTTAAGCCCCTTTGGTGTATTTGACAGGTTACATAATACCCATTTTTTTTCTAAAAGCTGGCATCTTTACTCCGGAAATTACTGAAAAAGACATGGCAAACATTTTCAGCGCT
AAAAGATCTTTCGTGTGTTTCTTAAGCCCCTTTGGTGTATTTGACAGGTTACATAATACCCATTTTTTTTCTAAAAGCTGGCATCTTTACTCCGGAAATTACTGAAAAAGACATGGCAAACATTTTCAGCGCT

Taken from GATK team

4. Base Quality Score Recalibration

WORKFLOW:



FASTQ

ALIGNMENT

mapped_reads.bam

MARK DUPLICATES

INDEL REALIGNMENT

BQSR

COMMON PATH

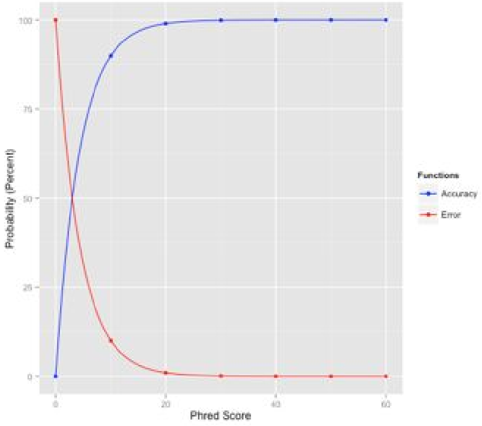
recalibrated.bam

SMALL VARIANT CALLING

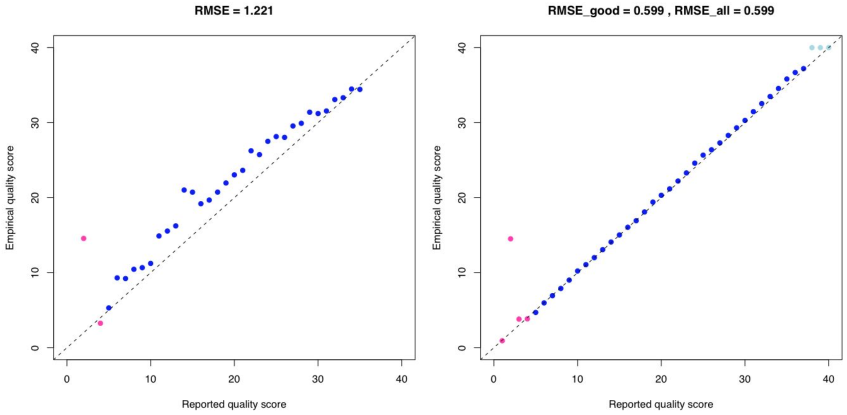
CNA CALLING

Phred Quality : each position of the sequence has its particular **base Quality score**.

The individual quality measures are NOT very important during the alignment step (mapping), but crucial during Variant calling.



Different NGS technologies have their particular bias in QS depending on the context. They could **correct empirically** these biases.



Original

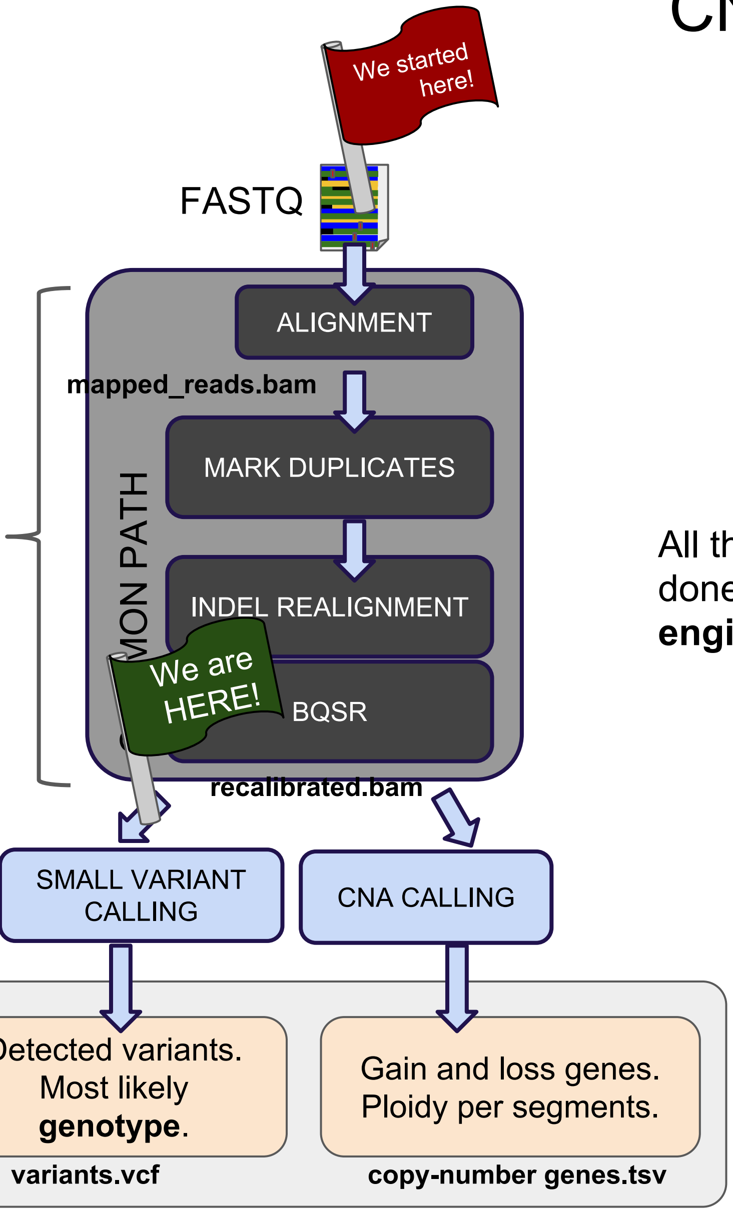
After BQSR recalibration

METHOD: by **GATK**
http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr

Point mutations and CNV Calling Process

The first two parts are **done**.

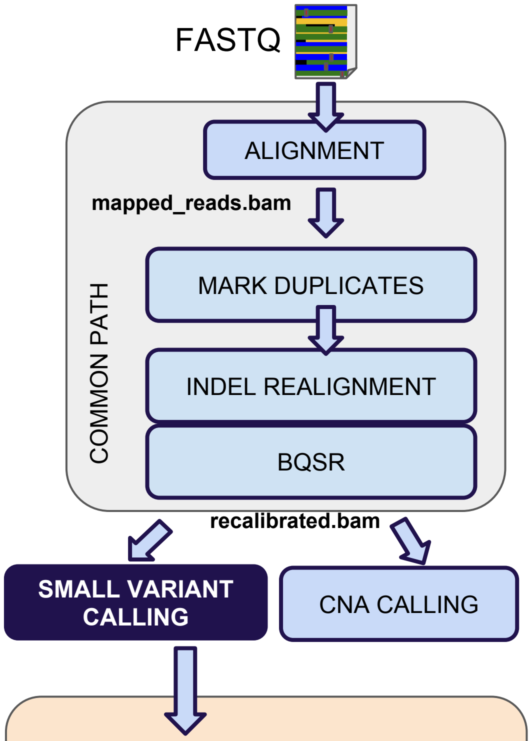
We are **ready** to discover variants on the sample.



All these steps are **automatically** done by the pipeline (**RUbioSeq's engine**).

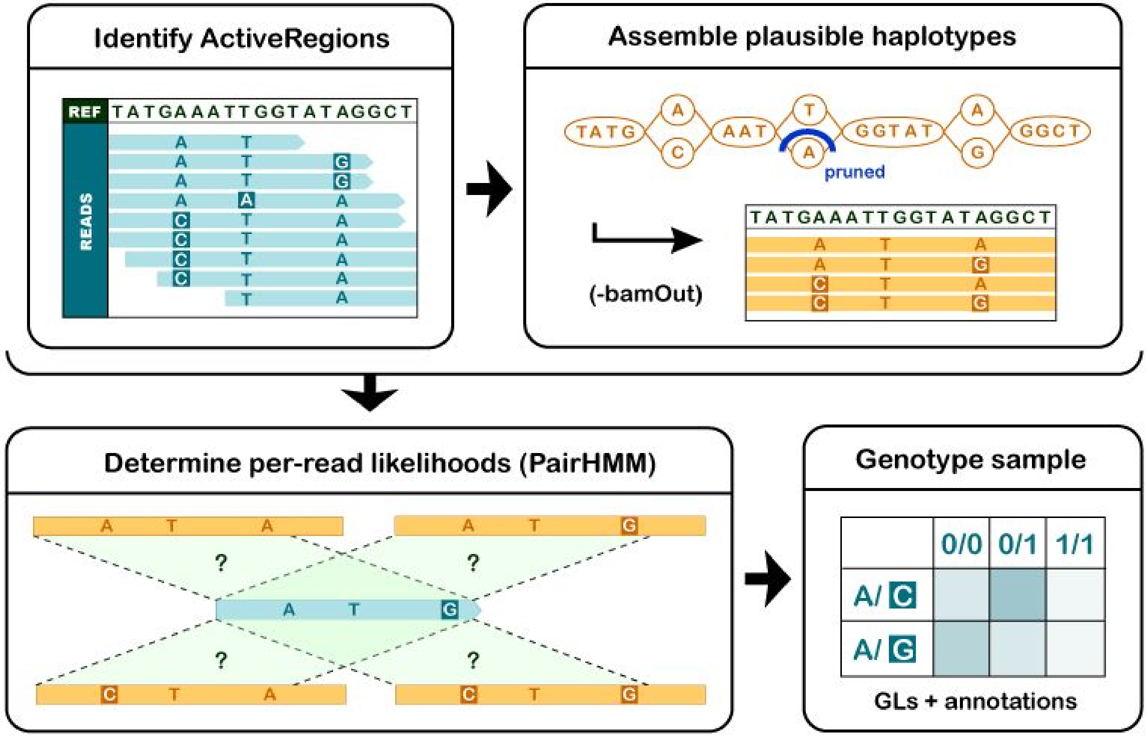
5. GATK Variant Calling Process : SNV & Indels

WORKFLOW:



- Detected variants:
  chr1:1234-1234:A>C
- Most likely **genotype**:
  AA or AC or CC

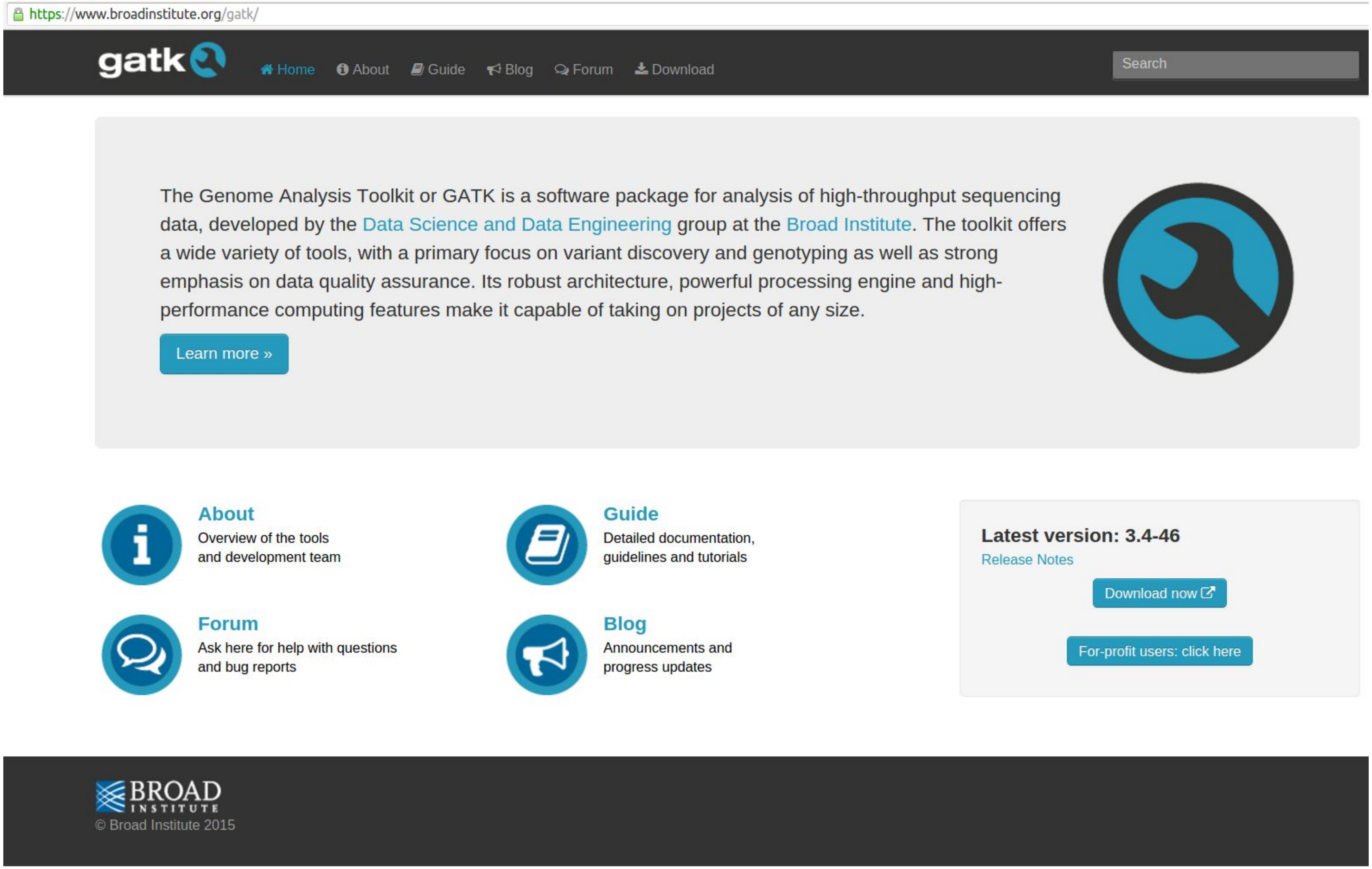
Haplotype Caller (new, included in **RUbioSeq v3.8.1)** : Variant calling based on the calculation of genotype likelihoods:



Assumptions: Diploid genome (2n).
Limitation: Allele freq > 20%.

Further reading:
http://gatkforums.broadinstitute.org/discussion/4148/hc-overview-how-the-haplotypecaller-works
HC steps 1-4: https://software.broadinstitute.org/gatk/documentation/topic?name=methods

GATK is in active development



The Genome Analysis Toolkit or GATK is a software package for analysis of high-throughput sequencing data, developed by the Data Science and Data Engineering group at the Broad Institute. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.

About
Overview of the tools and development team

Guide
Detailed documentation, guidelines and tutorials

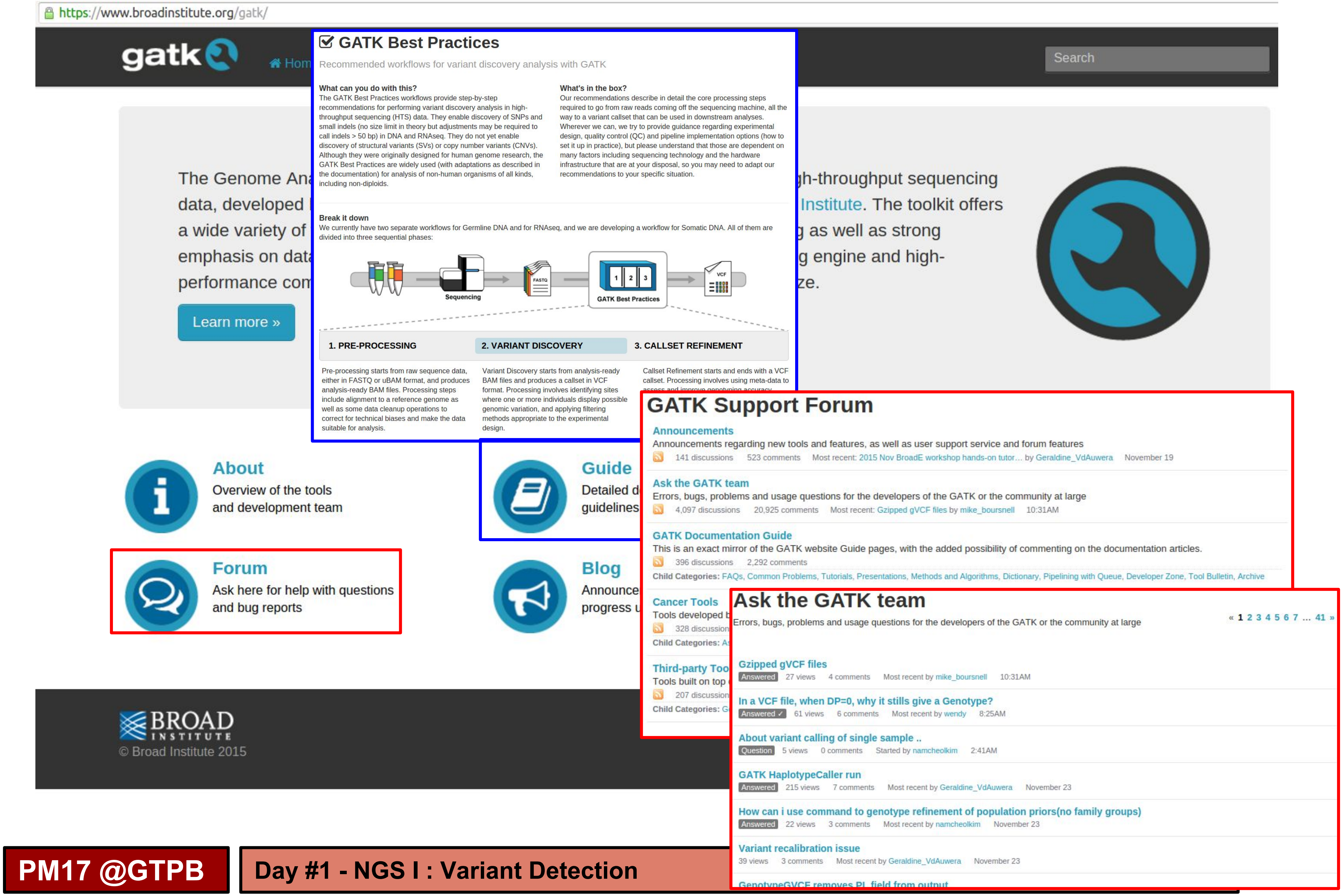
Forum
Ask here for help with questions and bug reports

Blog
Announcements and progress updates

Latest version: 3.4-46

© Broad Institute 2015

GATK is in active development



https://www.broadinstitute.org/gatk/

The Genome Ana … data, developed … a wide variety of … emphasis on data … performance com … gh-throughput sequencing … Institute. The toolkit offers … g as well as strong … g engine and high- … ze.

Learn more »

About
Overview of the tools and development team

Guide
Detailed d… guidelines

Forum
Ask here for help with questions and bug reports

Blog
Announce… progress u…

BROAD INSTITUTE
© Broad Institute 2015

☑ GATK Best Practices

Recommended workflows for variant discovery analysis with GATK

What can you do with this?
The GATK Best Practices workflows provide step-by-step recommendations for performing variant discovery analysis in high-throughput sequencing (HTS) data. They enable discovery of SNPs and small indels (no size limit in theory but adjustments may be required to call indels > 50 bp) in DNA and RNAseq. They do not yet enable discovery of structural variants (SVs) or copy number variants (CNVs). Although they were originally designed for human genome research, the GATK Best Practices are widely used (with adaptations as described in the documentation) for analysis of non-human organisms of all kinds, including non-diploids.

What's in the box?
Our recommendations describe in detail the core processing steps required to go from raw reads coming off the sequencing machine, all the way to a variant callset that can be used in downstream analyses. Wherever we can, we try to provide guidance regarding experimental design, quality control (QC) and pipeline implementation options (how to set it up in practice), but please understand that those are dependent on many factors including sequencing technology and the hardware infrastructure that are at your disposal, so you may need to adapt our recommendations to your specific situation.

Break it down
We currently have two separate workflows for Germline DNA and for RNAseq, and we are developing a workflow for Somatic DNA. All of them are divided into three sequential phases:

1. PRE-PROCESSING — **2. VARIANT DISCOVERY** — **3. CALLSET REFINEMENT**

Pre-processing starts from raw sequence data, either in FASTQ or uBAM format, and produces analysis-ready BAM files. Processing steps include alignment to a reference genome as well as some data cleanup operations to correct for technical biases and make the data suitable for analysis.

Variant Discovery starts from analysis-ready BAM files and produces a callset in VCF format. Processing involves identifying sites where one or more individuals display possible genomic variation, and applying filtering methods appropriate to the experimental design.

Callset Refinement starts and ends with a VCF callset. Processing involves using meta-data to assess and improve genotyping accuracy

GATK Support Forum

Announcements
Announcements regarding new tools and features, as well as user support service and forum features
141 discussions 523 comments Most recent: 2015 Nov BroadE workshop hands-on tutor… by Geraldine_VdAuwera November 19

Ask the GATK team
Errors, bugs, problems and usage questions for the developers of the GATK or the community at large
4,097 discussions 20,925 comments Most recent: Gzipped gVCF files by mike_boursnell 10:31AM

GATK Documentation Guide
This is an exact mirror of the GATK website Guide pages, with the added possibility of commenting on the documentation articles.
396 discussions 2,292 comments
Child Categories: FAQs, Common Problems, Tutorials, Presentations, Methods and Algorithms, Dictionary, Pipelining with Queue, Developer Zone, Tool Bulletin, Archive

Cancer Tools
Tools developed b…
328 discussion…
Child Categories: A…

Third-party Too…
Tools built on top …
207 discussion…
Child Categories: G…

Ask the GATK team

Errors, bugs, problems and usage questions for the developers of the GATK or the community at large

« 1 2 3 4 5 6 7 … 41 »

Gzipped gVCF files
Answered 27 views 4 comments Most recent by mike_boursnell 10:31AM

In a VCF file, when DP=0, why it stills give a Genotype?
Answered ✓ 61 views 6 comments Most recent by wendy 8:25AM

About variant calling of single sample ..
Question 5 views 0 comments Started by namcheolkim 2:41AM

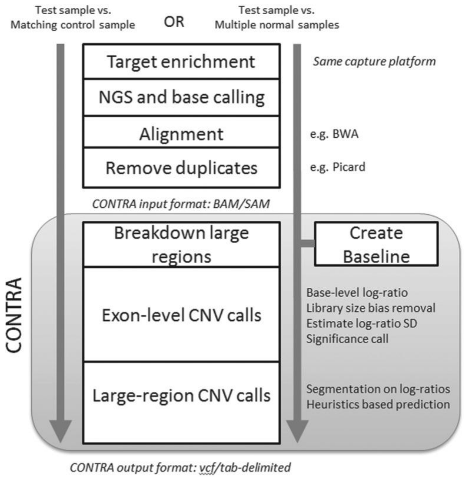
GATK HaplotypeCaller run
Answered 215 views 7 comments Most recent by Geraldine_VdAuwera November 23

How can i use command to genotype refinement of population priors(no family groups)
Answered 22 views 3 comments Most recent by namcheolkim November 23

Variant recalibration issue
39 views 3 comments Most recent by Geraldine_VdAuwera November 23

GenotypeGVCF removes PL field from output

6. CNA Variant Calling

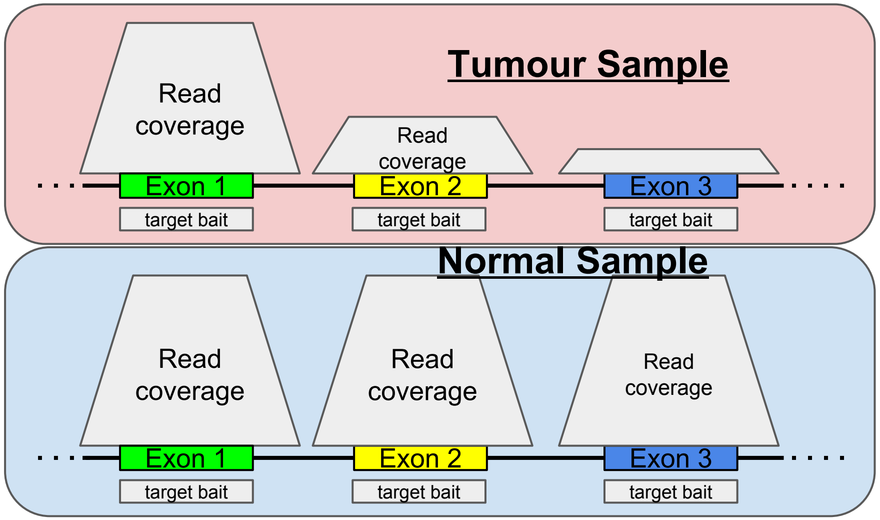


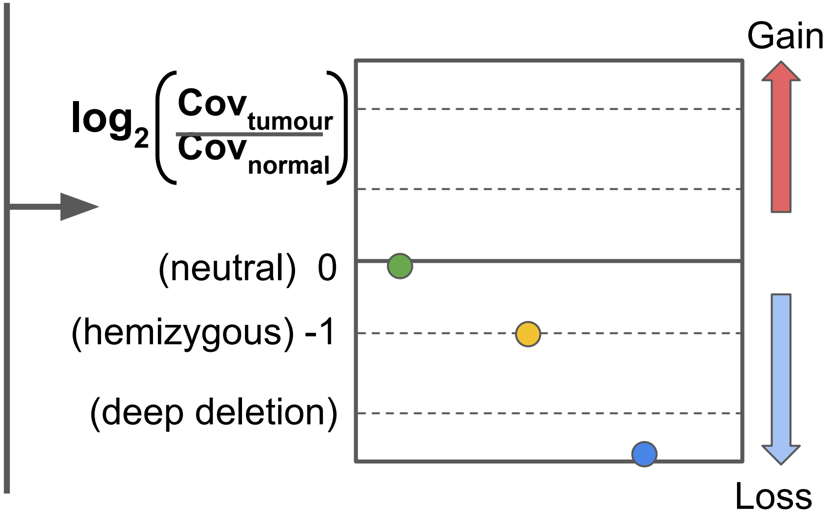
- **Normalization**: Split large regions. GC-content bias, unbalanced library size effect on log-ratios.

- Read-depth coverage & log2 CN ratio are corrected.

- Significance:
  Assumption: log2-transformed coverage fits a normal distribution:

  $\mathrm{RLR} \sim \mathrm{N}(\mu_d, \sigma_d)$; Two-tailed P-value. multiple testing correction (FDR).





Li J et al. CONTRA: copy number analysis for targeted resequencing. (2012) Bioinformatics

- # What have we learnt?

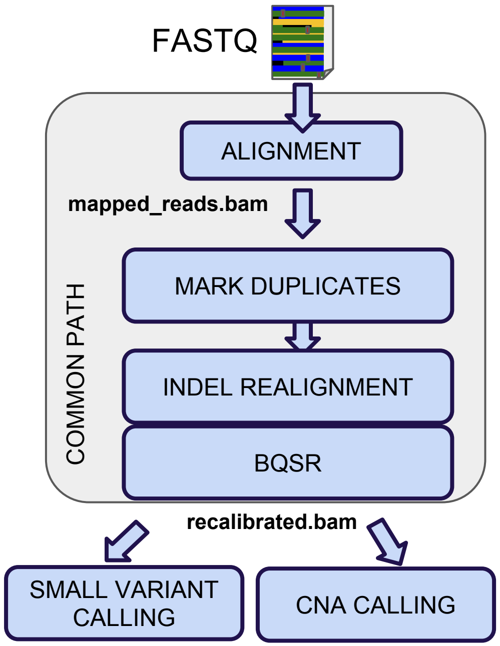
Main concepts:

- Variants.
- How to detect them.
- Differences between platforms.

Requirements:

- Files.
- Methods

step-by-step WORKFLOW

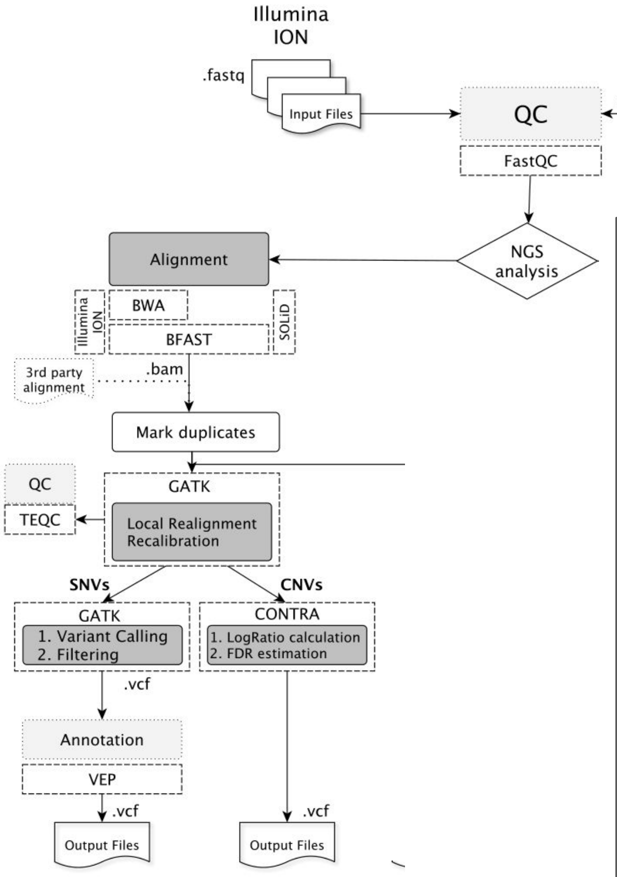


- # Questions?

Thanks to Miriam Rubio-Camarillo & Gonzalo Gómez for the support and development of RUbioSeq.

First hands-on :: Configure RUbioSeq+

Configuration files



It is already installed.

Let’s configure it!

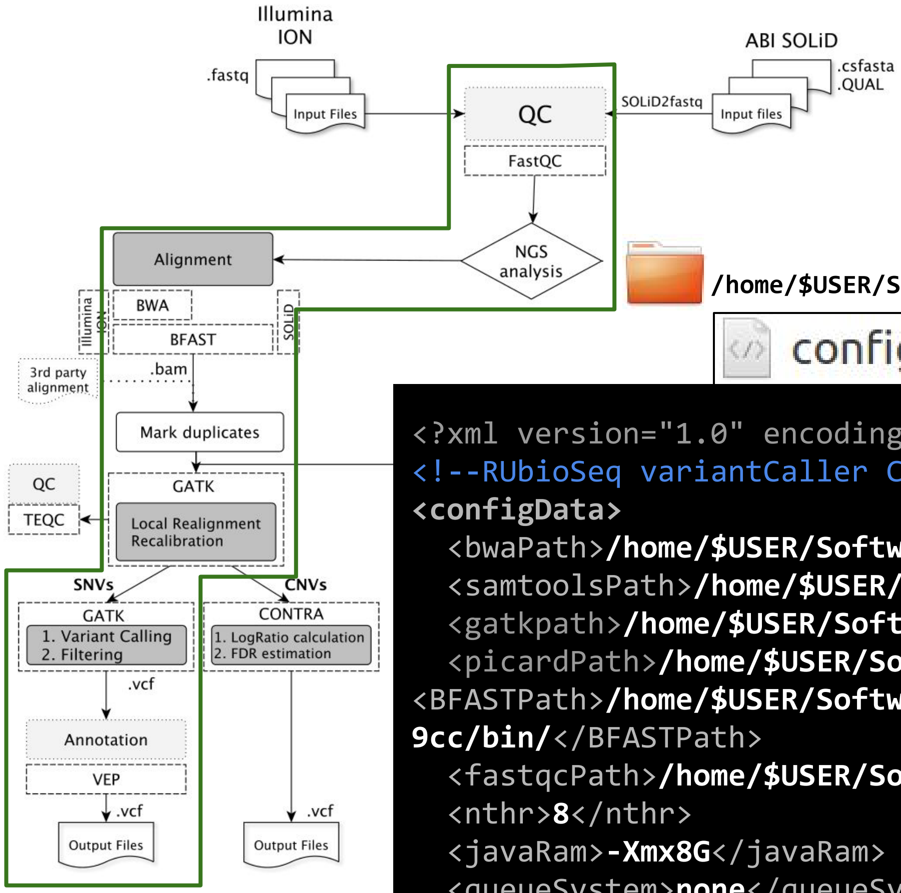
- What will we do?

Configure the pipeline for the detection of:

- Small variants (SNV + indels)
- Copy-Number Variants

First hands-on :: Configure RUbioSeq+

Configuration for Small variant analysis



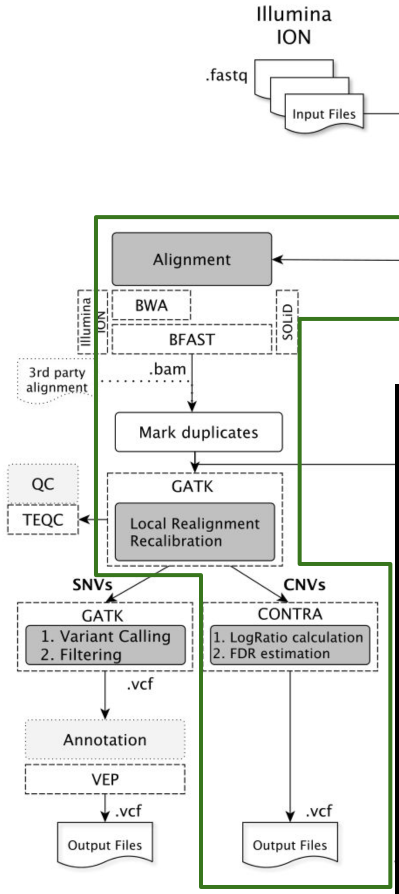
/home/$USER/Software/RUbioSeq3.8.1/variantCalling/config/snv

configProgramPaths.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!--RUbioSeq variantCaller CONFIG FILE-->
<configData>
  <bwaPath>/home/$USER/Software/bwa-0.7.10/</bwaPath>
  <samtoolsPath>/home/$USER/Software/samtools-0.1.19/</samtoolsPath>
  <gatkpath>/home/$USER/Software/GATK-3.4-0/GenomeAnalysisTK-3.4-0/</gatkpath>
  <picardPath>/home/$USER/Software/picard-tools-1.107/</picardPath>
<BFASTPath>/home/$USER/Software/bfast-bwa-ed42c18ea7f48af862935be52f1c072b1d560
9cc/bin/</BFASTPath>
  <fastqcPath>/home/$USER/Software/FastQC/</fastqcPath>
  <nthr>8</nthr>
  <javaRam>-Xmx8G</javaRam>
  <queueSystem>none</queueSystem>
  <queueName>none</queueName>
  <!--<multicoreName>multicore</multicoreName>
  <multicoreNumber>4</multicoreNumber>-->
</configData>
```

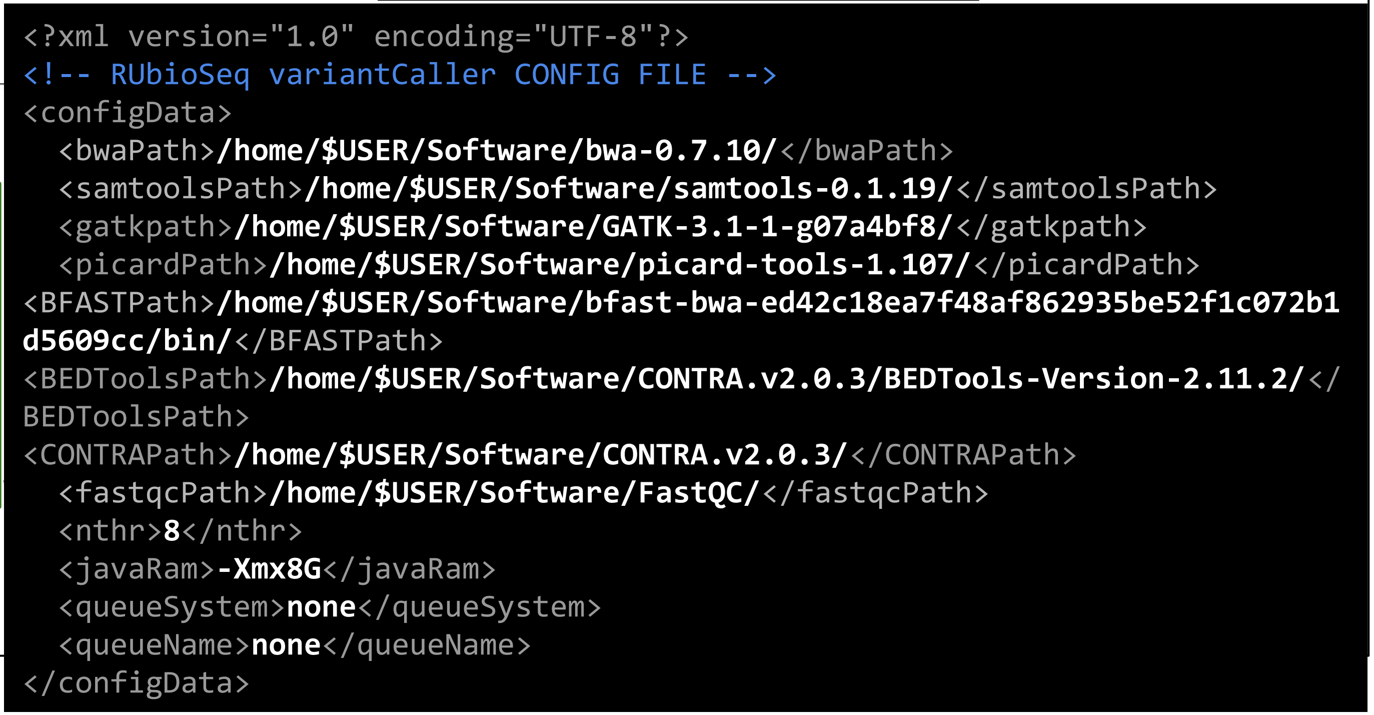
First hands-on :: Configure RUbioSeq+

Configuration for CNV analysis



/home/$USER/Software/RUbioSeq3.8.1/variantCalling/config/cnv

configProgramPaths.xml



```
<?xml version="1.0" encoding="UTF-8"?>
<!-- RUbioSeq variantCaller CONFIG FILE -->
<configData>
  <bwaPath>/home/$USER/Software/bwa-0.7.10/</bwaPath>
  <samtoolsPath>/home/$USER/Software/samtools-0.1.19/</samtoolsPath>
  <gatkpath>/home/$USER/Software/GATK-3.1-1-g07a4bf8/</gatkpath>
  <picardPath>/home/$USER/Software/picard-tools-1.107/</picardPath>
<BFASTPath>/home/$USER/Software/bfast-bwa-ed42c18ea7f48af862935be52f1c072b1
d5609cc/bin/</BFASTPath>
<BEDToolsPath>/home/$USER/Software/CONTRA.v2.0.3/BEDTools-Version-2.11.2/</
BEDToolsPath>
<CONTRAPath>/home/$USER/Software/CONTRA.v2.0.3/</CONTRAPath>
  <fastqcPath>/home/$USER/Software/FastQC/</fastqcPath>
  <nthr>8</nthr>
  <javaRam>-Xmx8G</javaRam>
  <queueSystem>none</queueSystem>
  <queueName>none</queueName>
</configData>
```