AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

# Bioinformatics and Genome Analyses

## Tools for variant analysis of next-generation genome sequencing data

## Lecture 1

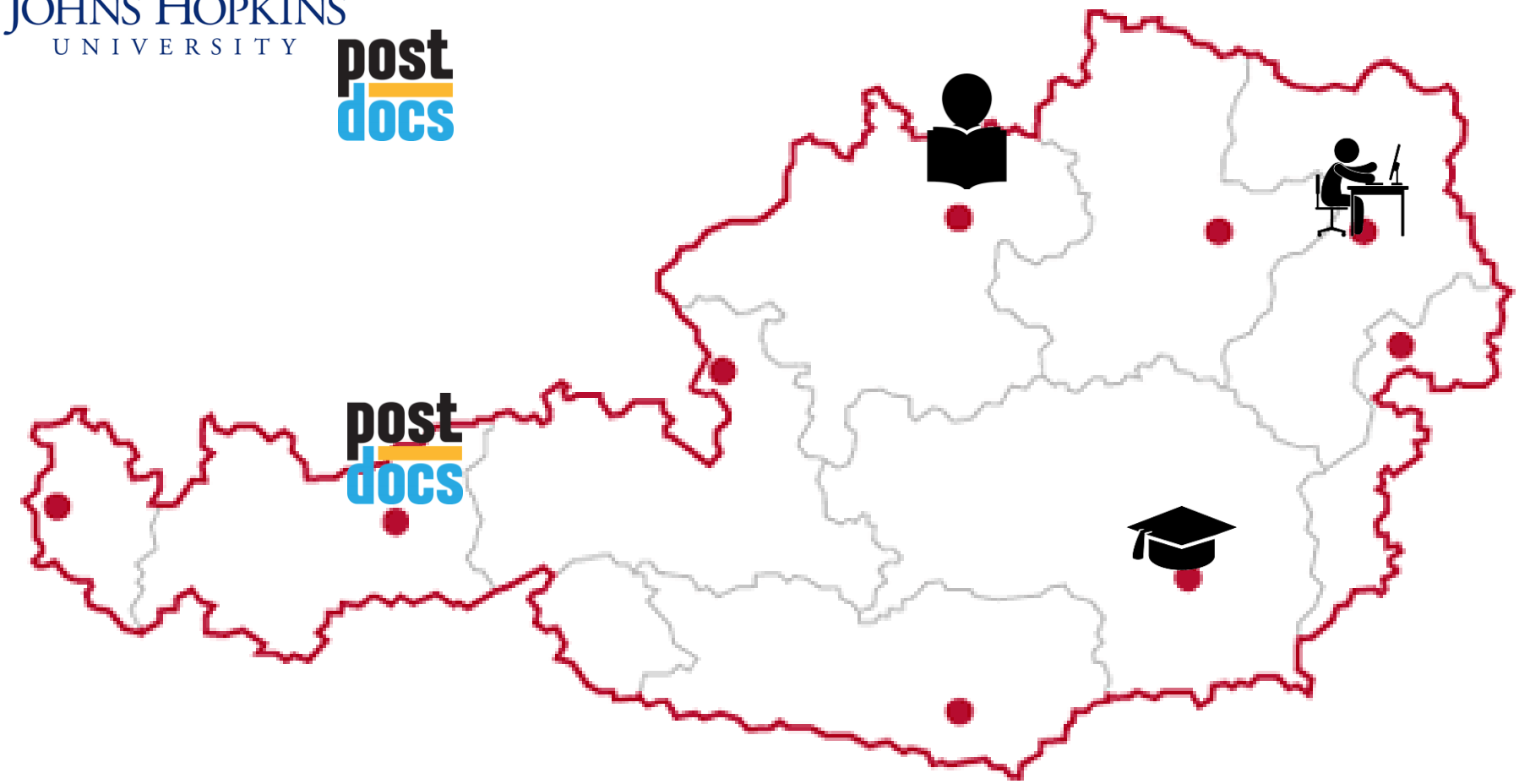
Stephan Pabinger
stephan.pabinger@ait.ac.at

# My background

JOHNS HOPKINS UNIVERSITY

post docs

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY



post docs

# My background

**Bioinformatics**

- Software development
- Web tools
- Pipeline design

**Working with sequencing data**

- DNASeq
- RNASeq
- MethylationSeq

**Data analysis**

- DNA data
- Protein data
- Peptide data
- Immunomics data

# What to expect

- Finish with an understanding of major concepts and tools
- Know how to perform variant calling
- Ready to make informed choices about what kind of variant calling tools you may need
- Focus is on *variant calling* and *functional annotation*
  - Alignment refinement
  - Variant calling
  - Variant annotation and filtering
  - Visualization

# Resources

Information about practicals can be found here:

https://github.com/spabinger/BioinformaticsAndGenomeAnalyses2019

Please make sure that you can access the resource

# GIT

# GIT

## Information

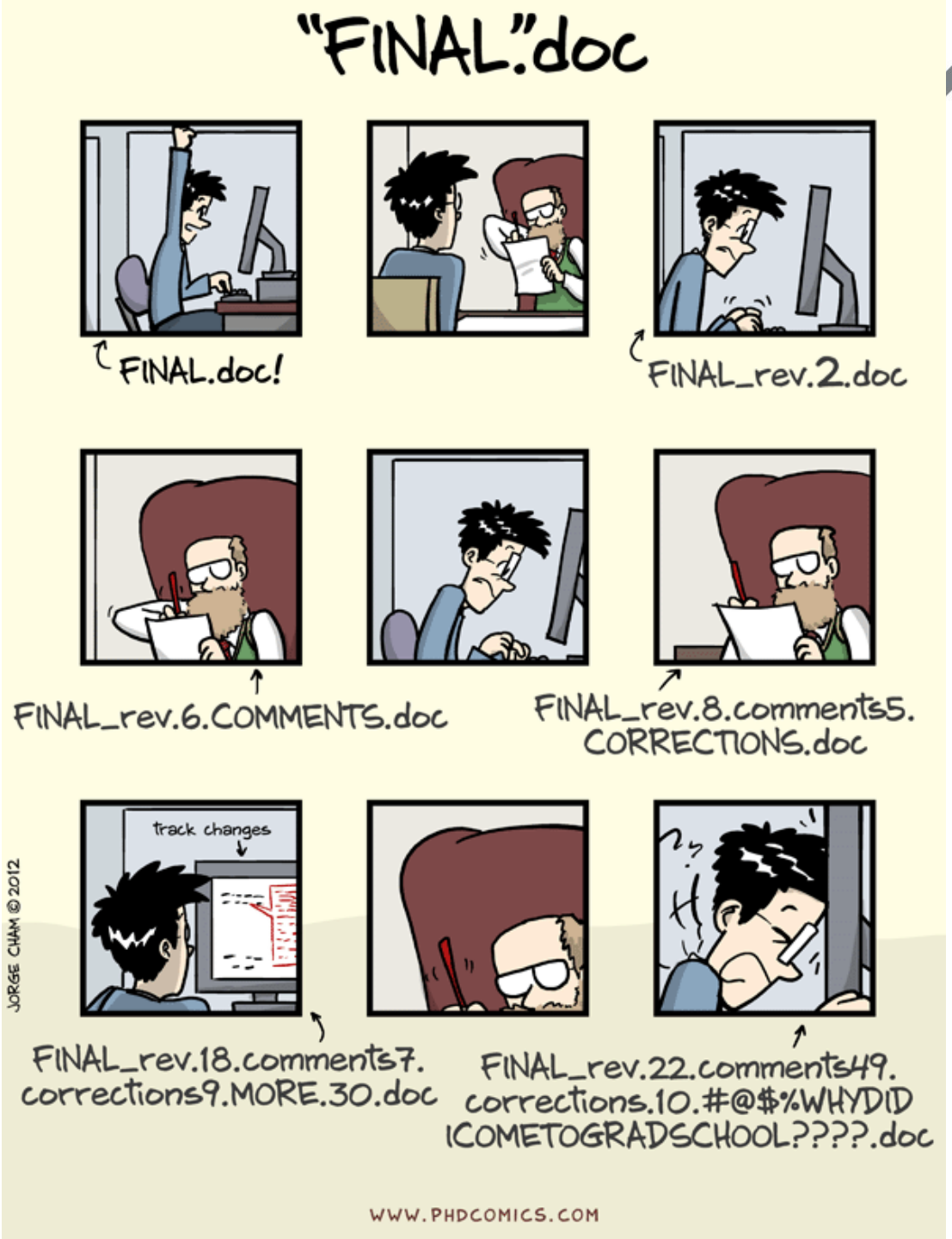
- Distributed revision control system
- Developed by Linus Torvalds (Linux developer)
- Local repositories & remote repositories

## Keep track of changes

- Code
- Manuscript
- Presentations
- Data analysis

## Master/PhD thesis

## Merging collaborators' changes


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

# Start with GIT

- Create a new directory
- Open it (cd into it)
- Perform
  `git init`

- Work ….

- Add files you want to store in the repository (locally)
  `git add XY.txt`
  `git add *` (to add everything)

- Commit files
  `git commit –m „Performed first analysis“`

# Start with GIT

```
stephan@shaq /tmp $ mkdir super_analysis
stephan@shaq /tmp $ cd super_analysis/
stephan@shaq /tmp/super_analysis $ git init
Initialized empty Git repository in /tmp/super_analysis/.git/
stephan@shaq /tmp/super_analysis $ touch XY.txt
stephan@shaq /tmp/super_analysis $ touch rawfile.txt
stephan@shaq /tmp/super_analysis $ git add XY.txt
stephan@shaq /tmp/super_analysis $ git add *
stephan@shaq /tmp/super_analysis $ git commit -m "Performed first analysis"
[master (root-commit) 915d159] Performed first analysis
 2 files changed, 0 insertions(+), 0 deletions(-)
 create mode 100644 XY.txt
 create mode 100644 rawfile.txt
stephan@shaq /tmp/super_analysis $
```

# Features

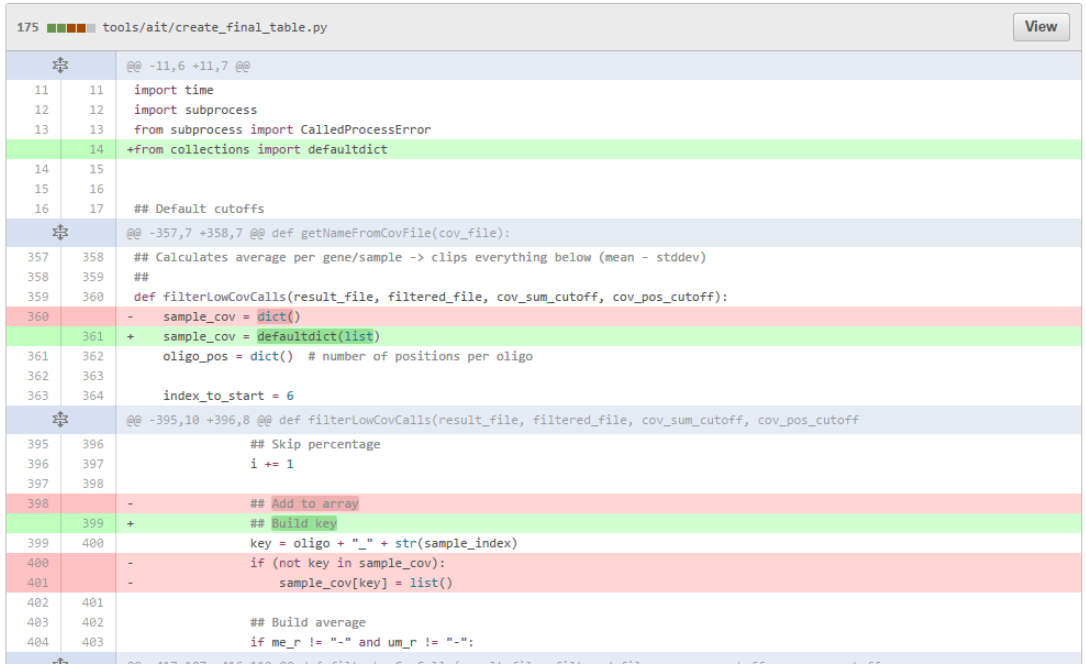
- Status of your project
  `git status`

```
stephan@shaq /tmp/super_analysis $ git status
On branch master
Untracked files:
  (use "git add <file>..." to include in what will be committed)

        nextanalysis.py

nothing added to commit but untracked files present (use "git add" to track)
stephan@shaq /tmp/super_analysis $
```

- History
- Go to a previous version
- Branching (implement a new feature without disrupting main code)
- Merging between different versions/branches



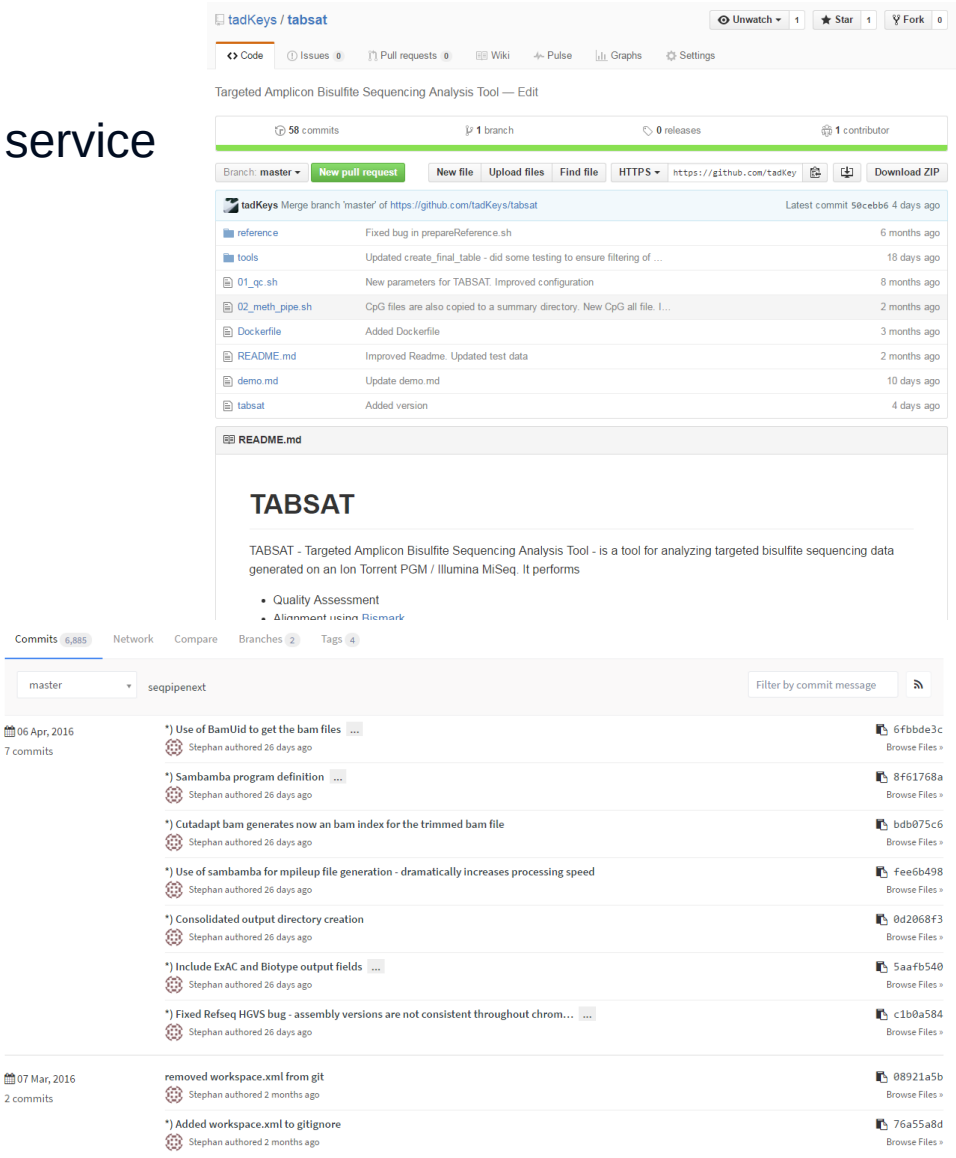
# Github - Gitlab

## Github

- Web-based Git repository hosting service
- Free to use
- Request for private repositories

## Gitlab

- Community version
- Open Source
- Share code, analyses, etc
- Easy to transfer to Github



# Remote repositories

- Clone repository
  ```
  git clone username@host:/path/to/repository
  ```

- Push files to remote repository
  ```
  git push
  ```

# GIT – Hints- Resources

- Make lots of commits
- Don‘t commit large files (they should be on the server)

**Information**

http://rogerdudler.github.io/git-guide/

http://kbroman.org/github_tutorial/

http://nyuccl.org/pages/GitTutorial/

https://swcarpentry.github.io/git-novice/

**Windows**

- https://git-for-windows.github.io/
  → Graphical user-interface (for init, add, commit, push, compare)

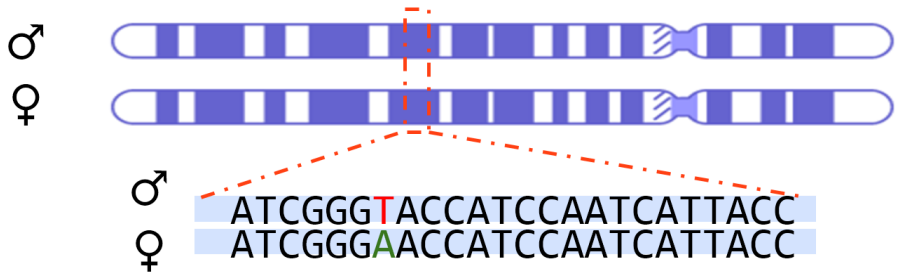
**Linux & Mac**

- Packages and GUIs available

# Terms

# Genetics Terms

**Humans are diploid:** Our genome is comprised of a paternal and a maternal "haplotype“
→ form our "genotype"




**Gene:** A hereditary unit consisting of a sequence of DNA that occupies a specific location on a chromosome and determines a particular characteristic in an organism

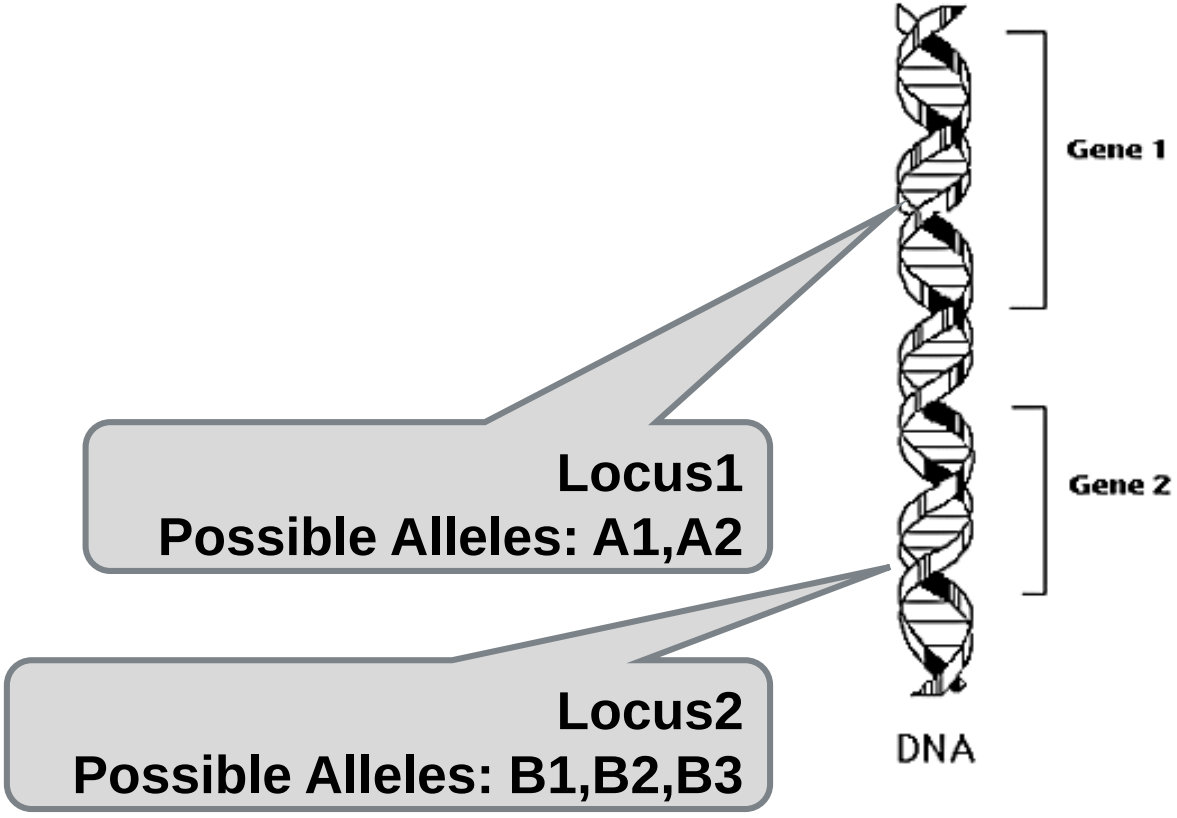
**Genotype:** Genetic makeup, distinguished from the physical appearance

**Phenotype:** The observable physical or biochemical characteristics as determined by both genetic makeup and environment

# Genetics terms

**Locus** location of a *gene/marker* on the chromosome

**Allele** one **variant form** of a gene/marker at a particular locus differ in their nucleotide sequence



# History

History of Molecular Diagnostics | Sequencing | and more

# The Molecular Biology Timeline

| | |
|---|---|
| 1865 | Gregor Mendel, Law of Heredity |
| 1866 | Johann Miescher, Purification of DNA |
| 1949 | Sickle Cell Anemia Mutation |

# Sickle Cell Anemia

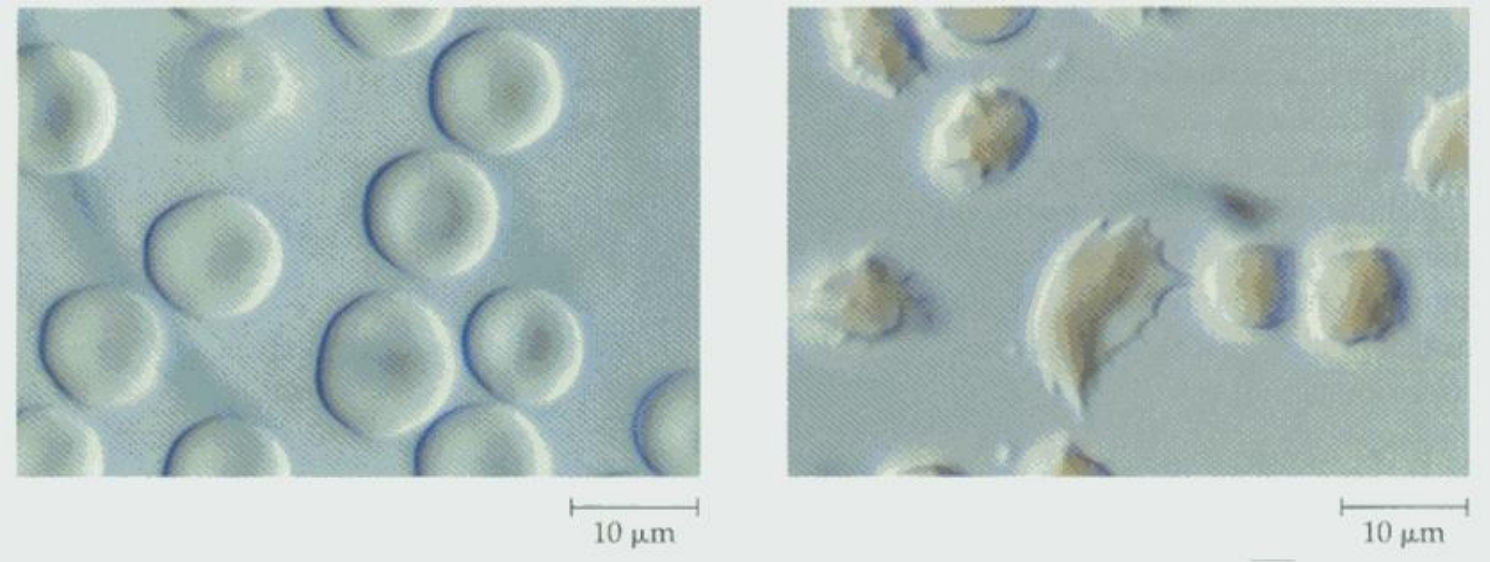
**Deformation of red blood cells**

**Epidemiology:**

1 out of 500 with African heritage; also common in Asia and Mediterranean region

**Genetics**

Single point mutation (SNP) → 6. Codon of the β-globin Gens



# The Molecular Biology Timeline

| | |
|---|---|
| 1865 | Gregor Mendel, Law of Heredity |
| 1866 | Johann Miescher, Purification of DNA |
| 1949 | Sickle Cell Anemia Mutation |
| 1953 | Watson and Crick, Structure of DNA |
| 1970 | Recombinant DNA Technology |
| 1977 | DNA sequencing |

# DNA Sequencing - mid 1970s

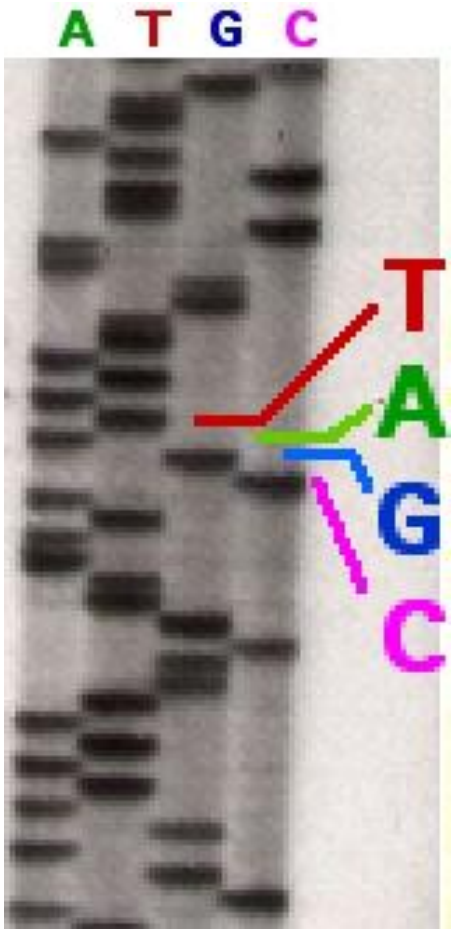
Mutations/Variations can be detected using DNA sequencing

**Sanger Sequencing**

- Dideoxy DNA sequencing paired with gel electrophoresis
- DNA is 5’ labeled with radioactivity
- Small amount of Dideoxy base added to 4 separate primer extension reactions
- Run on a gel to determine bases at each position by size
- Still considered the gold standard for validating sequencing data

Maxam-Gilbert Sequencing

- Chemical modification and cleavage paired with gel electrophoresis
- DNA is 5’ labeled with radioactivity (gamma-P ATP)
- Exposed to chemical agents that cause specific DNA breaks
- Run on a gel and the pattern reveals which base is at each site




Sanger

# The Molecular Biology Timeline

| | |
|---|---|
| 1865 | Gregor Mendel, Law of Heredity |
| 1866 | Johann Miescher, Purification of DNA |
| 1949 | Sickle Cell Anemia Mutation |
| 1953 | Watson and Crick, Structure of DNA |
| 1970 | Recombinant DNA Technology |
| 1977 | DNA sequencing |
| 1985 | *In Vitro* Amplification of DNA (PCR) |
| 2001 | The Human Genome Project |

# Human Genome Project

U.S. Government project coordinated by the Dept. of Energy and NIH

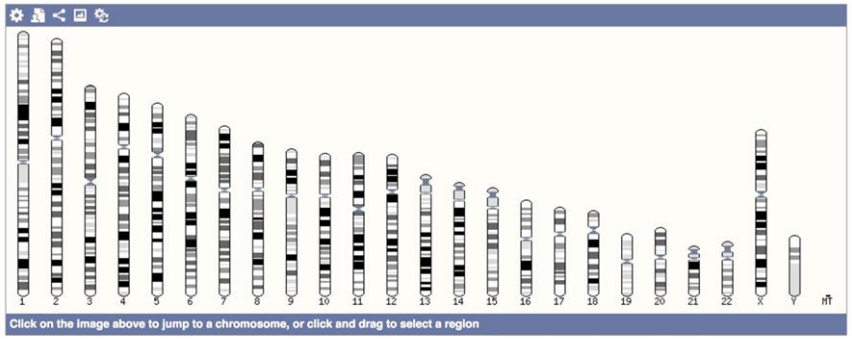
Total cost of over 3 billion $

**Goals of the Human Genome Project (1990–2006)**

- Identify all of the genes in human DNA
- Determine the sequences of the 3 billion bases that make up human DNA
- Create databases
- Develop tools for data analysis
- Address the ethical, legal, and social issues that arise from genome research

# The human genome - basic stats

- ~ 3 billion base pairs (haploid)
- ~ 20,000 protein coding genes
- ~ 200,000 coding transcripts (isoforms of a gene that each encode a distinct protein product)




| | Gencode[a] | Ensembl[b] | RefSeq[c] | CHESS[d] |
|---|---|---|---|---|
| Protein-coding genes | 19,901 | 20,376 | 20,345 | 21,306 |
| lncRNA genes | 15,779 | 14,720 | 17,712 | 18,484 |
| Antisense RNA | 5501 | | 28 | 2694 |
| Miscellaneous RNA | 2213 | 2222 | 13,899 | 4347 |
| Pseudogenes | 14,723 | 1740 | 15,952 | |
| Total transcripts | 203,835 | 203,903 | 154,484 | 323,827 |

**Summary**

| | |
|---|---|
| **Assembly** | GRCh38.p7 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.22, Dec 2013 |
| **Database version** | 87.38 |
| **Base Pairs** | 3,547,762,741 |
| **Golden Path Length** | 3,096,649,726 |
| **Genebuild by** | Ensembl |
| **Genebuild method** | Full genebuild |
| **Genebuild started** | Jan 2014 |
| **Genebuild released** | Jul 2014 |
| **Genebuild last updated/patched** | Jun 2016 |
| **Gencode version** | GENCODE 25 |

**Gene counts (Primary assembly)**

| | |
|---|---|
| **Coding genes** | 20,441 (incl 526 readthrough) |
| **Non coding genes** | 22,219 |
| **Small non coding genes** | 5,052 |
| **Long non coding genes** | 14,727 (incl 214 readthrough) |
| **Misc non coding genes** | 2,222 |
| **Pseudogenes** | 14,606 (incl 5 readthrough) |
| **Gene transcripts** | 198,002 |

Open questions: How many genes do we have? Salzberg, 2018, BMC Biology; http://uswest.ensembl.org/Homo_sapiens/Location/Genome

# Forms of genetic variations

**Single nucleotide substitution**
Replacement of one nucleotide with another

(Recap) Tandem repeat
ATTCG ATTCG ATTCG

**Microsatellites or mini-satellites**
These tandem repeats often present high levels of inter- and intra-specific polymorphism

**Deletions or insertions**
Loss or addition of one or more nucleotides

**Structural variations**
Changes in chromosome number, segmental rearrangements and deletions

# Sequencing

# Technological advances of Sanger sequencing

**1977**: Fred Sanger → 700 bases per day → 118,000 years to sequence the human genome

**1985**: ABI 370 (first automated sequencer) → 5000 bases per day →16,000 years

**1995**: ABI 377 (Bigger gels, better chemistry & optics, more sensitive dyes, faster computers) → 19,000 bases per day → 4,400 years

**1999**: ABI 3700 (96 capillaries, 96 well plates, fluid handling robots) → 400,000 bases per day → 205 years

*To sequence with a coverage of 10X*

# Sequencing duration - calculation

| Genome size | Coverage | Throughput | Days | Years |
|---|---|---|---|---|
| 3.000.000.000 | 10 | 400.000 | 75.000 | 205 |
| 3.000.000.000 | 10 | 19000 | 1.578.947 | 4.326 |
| 3.000.000.000 | 10 | 5000 | 6.000.000 | 16.438 |
| 3.000.000.000 | 10 | 700 | 42.857.143 | 117.417 |

# Second Generation Sequencing

- Developed to **increase throughput of Sanger sequencing**
- Can sequence **many molecules in parallel**
  - Does not require homogenous input
  - DNA sequenced as clusters or in nanowells
  - Single machine can sequence 3-10 Billion independent DNA fragments at once
  - Single Sanger Sequencer maxes out at 1152 reactions per machine

- Time from DNA to genome reduced from 10 years to 1 day!

# Disadvantages of 2nd Generation Tech

**Rely on amplification to create libraries and clusters**

- All polymerases have an inherent error rate ($10^{-6}$-$10^{-7}$)
- Errors introduced every 10 million to 100 million bases
- Secondary validation of variants is key

**GC bias**

- PCR bias against GC rich sequences
- Exome capture bias against GC rich sequences

**Short reads can miss large structural variations**

- Genome Translocations and inversions likely will be missed
- Require significant read depth at break points for these variations to be detected

**Trouble detecting small insertions and deletions**

- Short reads computationally hard to align and call

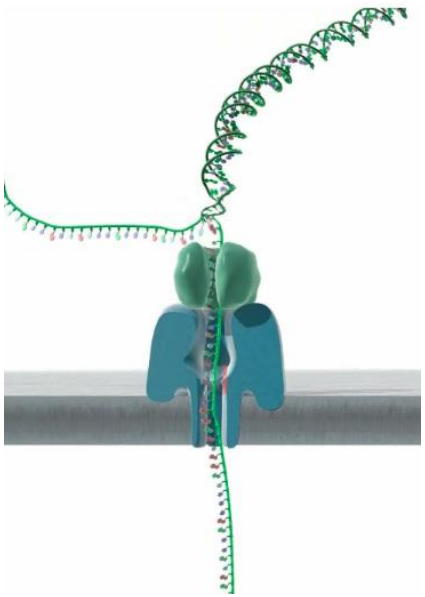
# Third Generation Sequencing

**Single molecule sequencing**

- Less complex sample prep
- Much longer read length (1-100kb)
- Many technical hurdles with very high error rates
- Expensive

**Sequencing by synthesis - Pacific Biosciences**

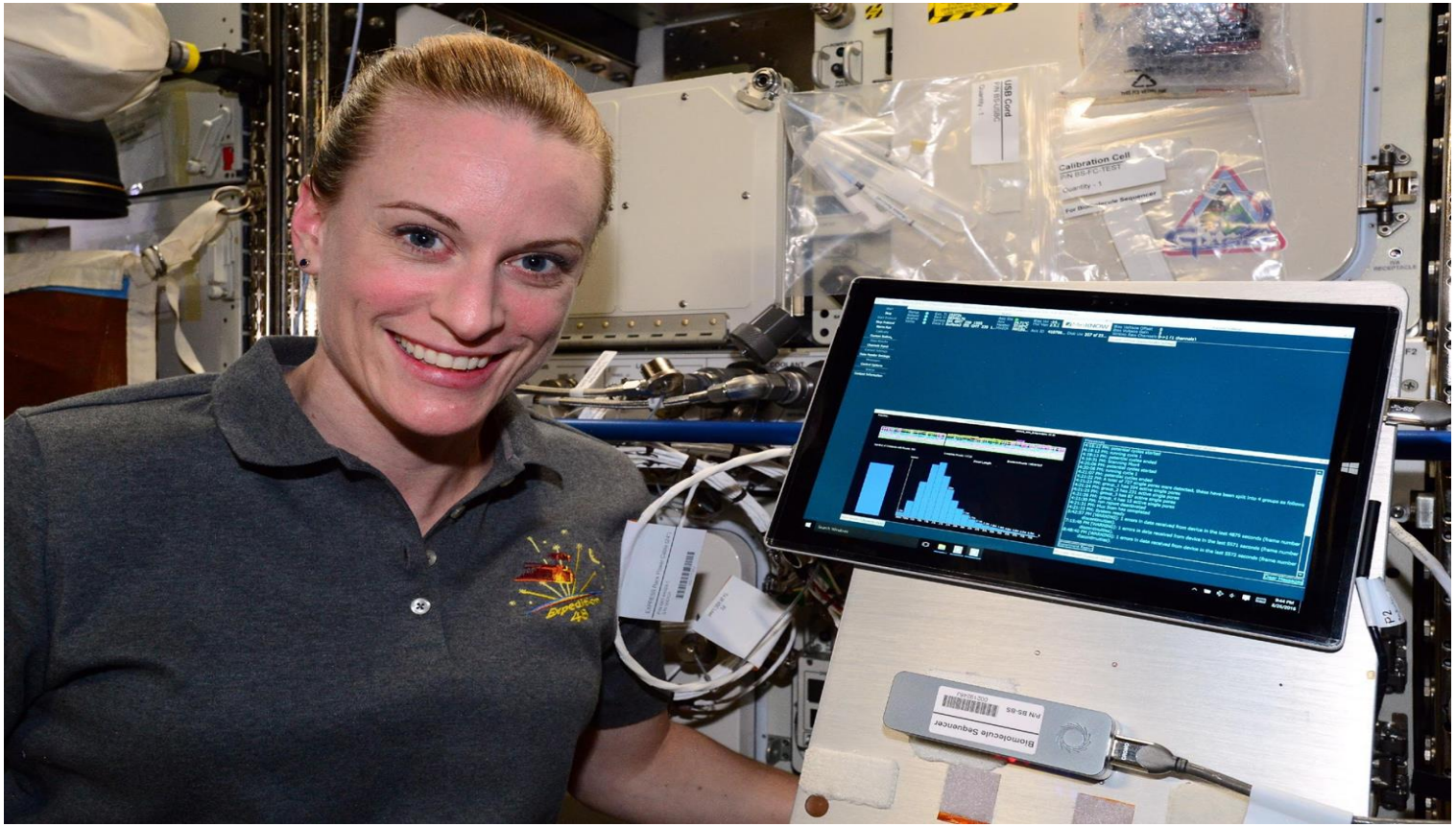
- 1 DNA molecule and 1 polymerase in each well (zero-mode waveguide)
- 4 different marker on the phosphate of the nucleotide
  → Polymerase interacts; marked freed
- No “theoretical” limit to DNA fragment length



**Direct sequencing by passing DNA through a nanopore**

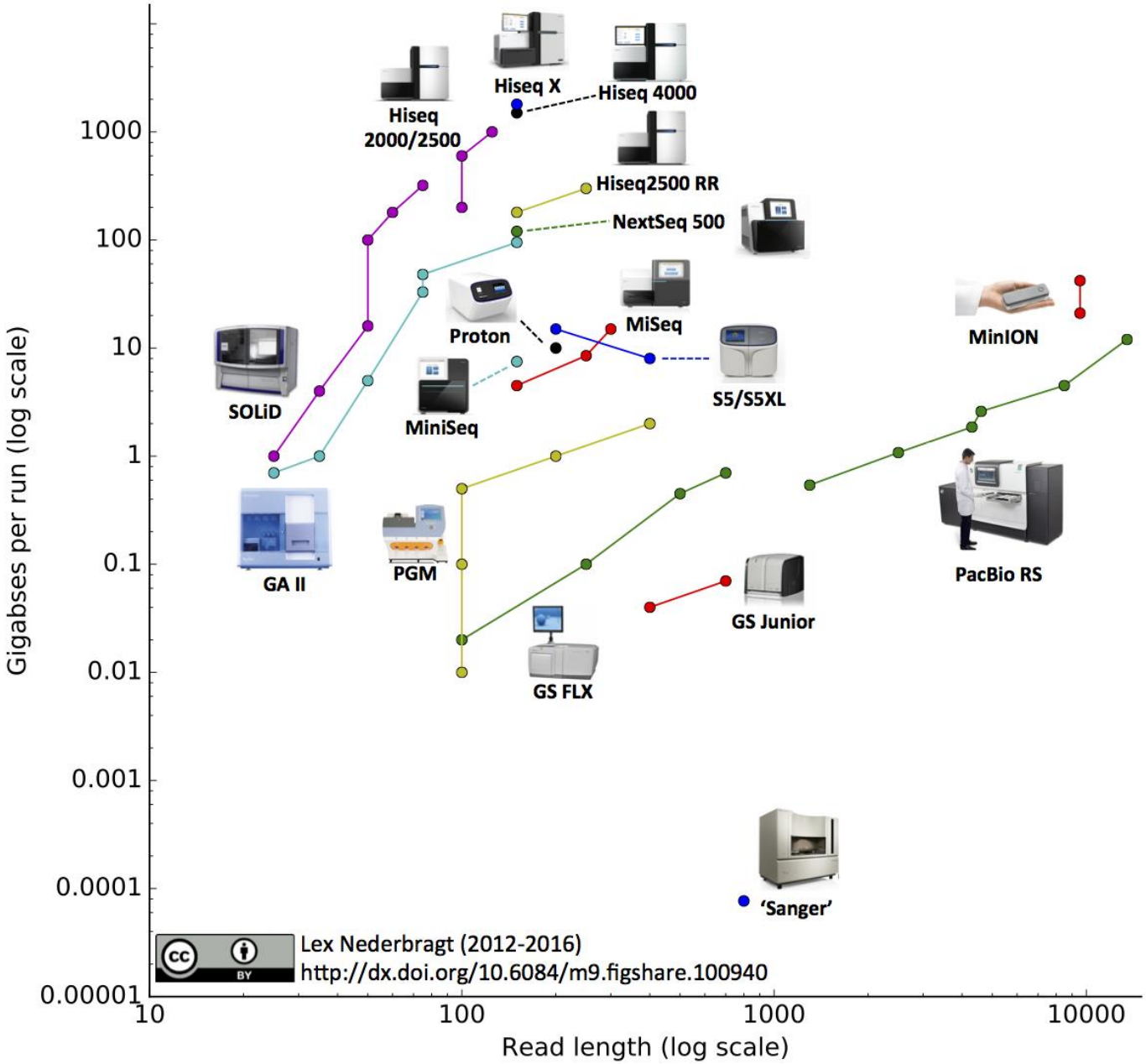
- Bases fed through a membrane bound nanopore
- Ionic difference between both sides of the membrane

# Nanopore – very portable




Kate Rubins sequencing DNA on the ISS

# High throughput sequencing



# Error Rates

| Instrument | Primary Errors | Single-pass Error Rate (%) | Final Error Rate (%) |
|---|---|---|---|
| Illumina | Substitutions | ~0.1 | ~0.1 |
| Ion Torrent | INDELs | ~1 | ~1 |
| Oxford Nanopore | Deletions | ≥4 | 4 |
| PacBio RS | INDELs | ~13 | ≤1 |

# Flavors of Sequencing

## Whole Genome Sequencing

- Obtain whole blood or tissue sample
- Create sequencing libraries of all DNA fragments

## Whole Exome Sequencing

- Utilizes a selection protocol
- Attach complimentary RNA or DNA strands to beads
- Fish out **only** coding DNA sequences
- Create sequencing libraries from enriched DNA
- Reduces cost and analysis time

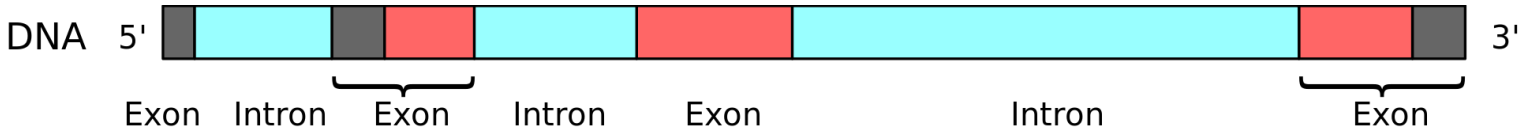
## Custom Capture

- Same protocol as Exome sequencing
- Only target desired DNA sequences

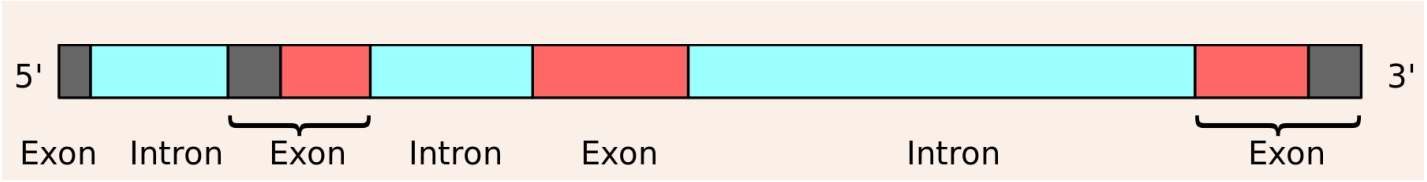
## Amplicon Sequencing

- Use PCR to amplify target DNA
- Sequence amplified DNA (Amplicon)

# Sequencing Techniques

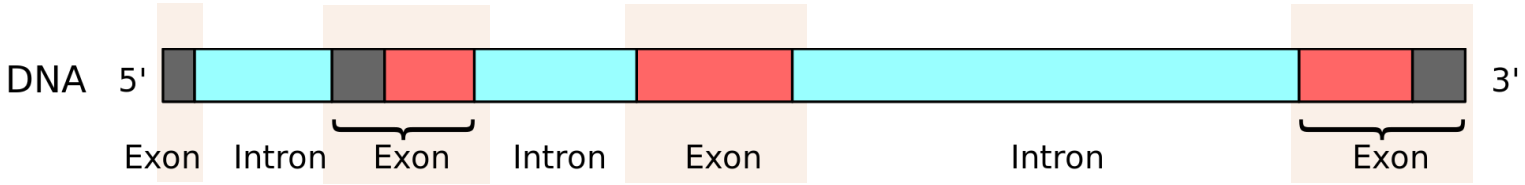


# Sequencing Techniques



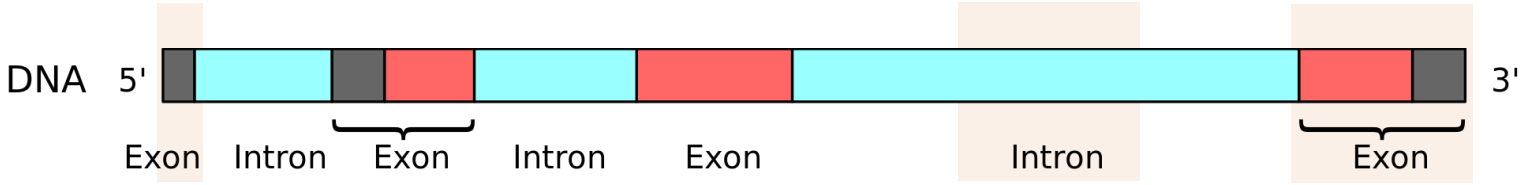

- Whole genome sequencing

# Sequencing Techniques



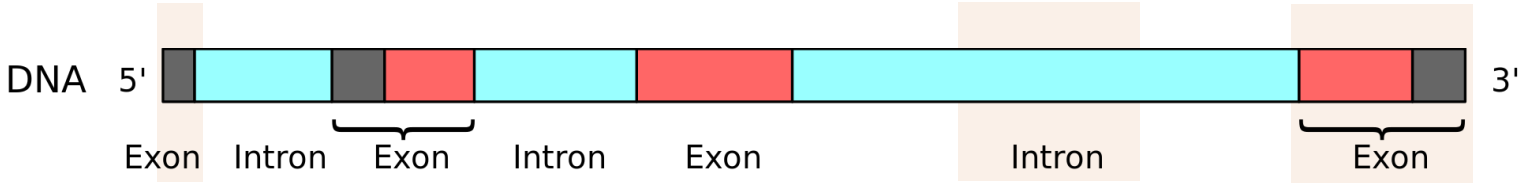

- Whole genome sequencing
- Whole exome sequencing
- Custom capture

# Sequencing Techniques




- Whole genome sequencing
- Whole exome sequencing
- Custom capture
- Amplicon sequencing

# Sequencing Techniques

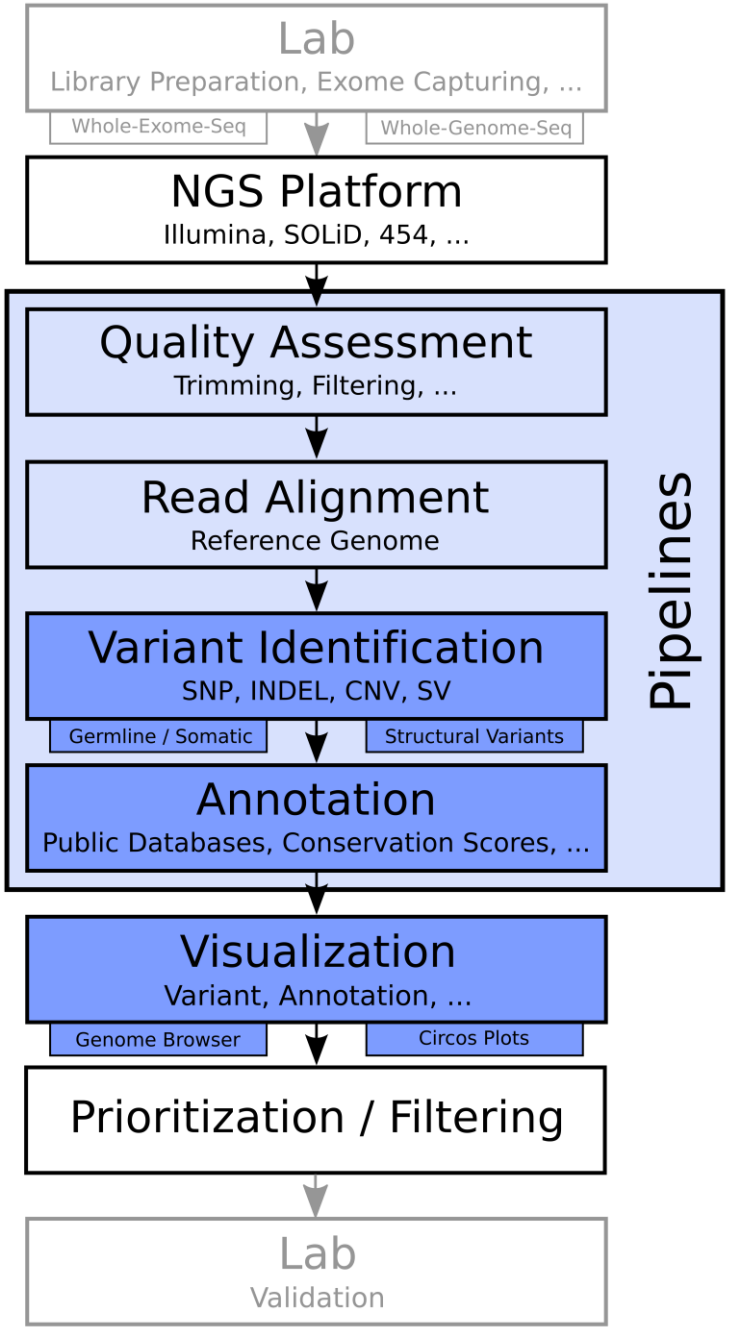



- Whole genome sequencing
- Whole exome sequencing
- Custom capture
- Amplicon sequencing

What is the best technology for my use-case?

- Research question?
- Number of samples?
- Cost?
- Future strategies?

# Workflow overview


Lab
Library Preparation, Exome Capturing, ...
Whole-Exome-Seq
Whole-Genome-Seq
NGS Platform
Illumina, SOLiD, 454, ...
Quality Assessment
Trimming, Filtering, ...
Read Alignment
Reference Genome
Variant Identification
SNP, INDEL, CNV, SV
Germline / Somatic
Structural Variants
Annotation
Public Databases, Conservation Scores, ...
Pipelines
Visualization
Variant, Annotation, ...
Genome Browser
Circos Plots
Prioritization / Filtering
Lab
Validation

# FASTQ

# FASTQ

AUSTRIAN INSTITUTE OF TECHNOLOGY

Storing and defining sequences from next-generation sequencing technologies

| | |
|---|---|
| Sequence ID | `@SEQ_ID` |
| Sequence | `GATTTGTTTGGCTCGATCGATACATAATCA` |
| Separator | `+` |
| Quality scores | `!*((())()+***+*()(!!!+5CCFCF555` |

## Old format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

| | |
|---|---|
| **HWUSI-EAS100R** | the unique instrument name |
| **6** | flowcell lane |
| **73** | tile number within the flowcell lane |
| **941** | 'x'-coordinate of the cluster within the tile |
| **1973** | 'y'-coordinate of the cluster within the tile |
| **#0** | index number for a multiplexed sample (0 for no indexing) |
| **/1** | the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)* |

## New format

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

| | |
|---|---|
| **EAS139** | the unique instrument name |
| **136** | the run id |
| **FC706VJ** | the flowcell id |
| **2** | flowcell lane |
| **2104** | tile number within the flowcell lane |
| **15343** | 'x'-coordinate of the cluster within the tile |
| **197393** | 'y'-coordinate of the cluster within the tile |
| **1** | the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)* |
| **Y** | Y if the read is filtered, N otherwise |
| **18** | 0 when none of the control bits are on, otherwise it is an even number |
| **ATCACG** | index sequence |

# Cleaning up FASTQ files

**Before mapping make sure**

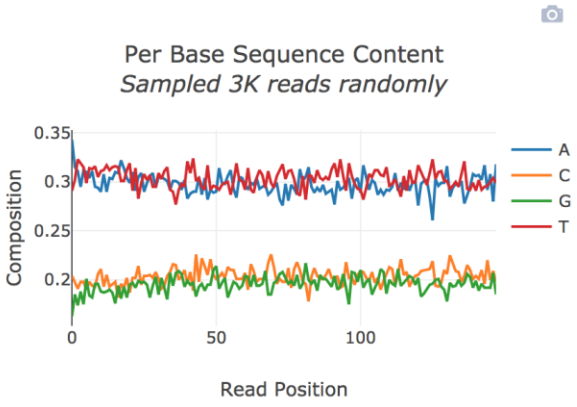
- Non genomic sequences are removed (barcodes, …)
- Adapter sequences are removed
- Clean contaminations (PRINSEQ, DeconSeq)
- Trim Ns
- Trim bad quality reads

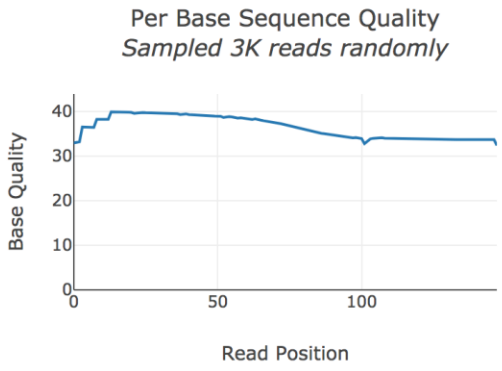
Skip cleaning → less read mapping; if not randomly distributed some areas → not enough coverage

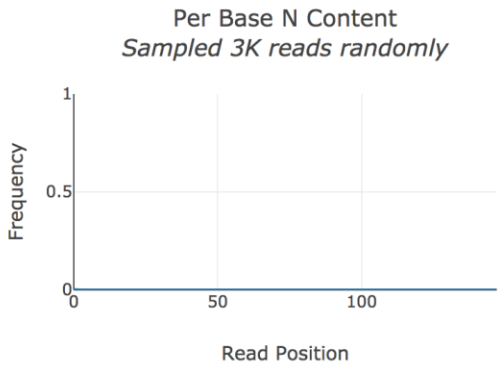
# Tools for FASTQ manipulation

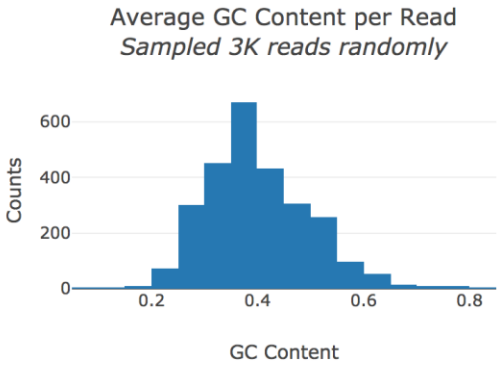
- FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  HTML output
- Fastx toolkit http://hannonlab.cshl.edu/fastx_toolkit/
  Lots of tools, charts, trimming, clipping, filtering
- Cutadapt https://code.google.com/p/cutadapt/
  Remove adapter sequences
- DeconSeq http://deconseq.sourceforge.net/
  User friendly interface, coverage plots, metagenomics datasets
- PRINSEQ: http://prinseq.sourceforge.net/
  HTML output, trimming, filtering, contaminations
- Trimmomatic http://www.usadellab.org/cms/?page=trimmomatic
  Paird-end trimming
- Atropos https://github.com/jdidion/atropos
  Multi-threading, paired-end, bisulfite-seq, …
- ...

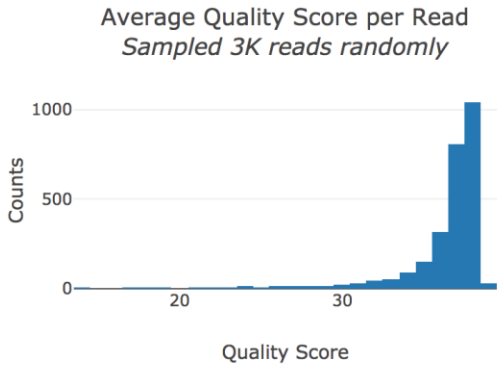
# www.fastq.bio

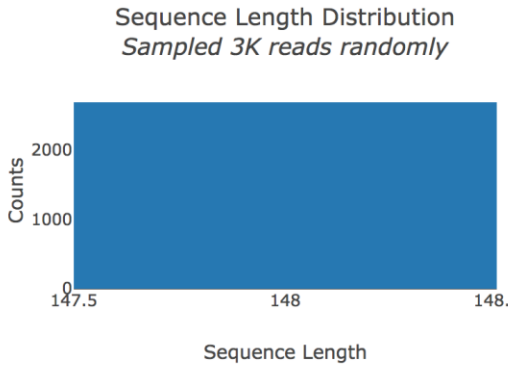


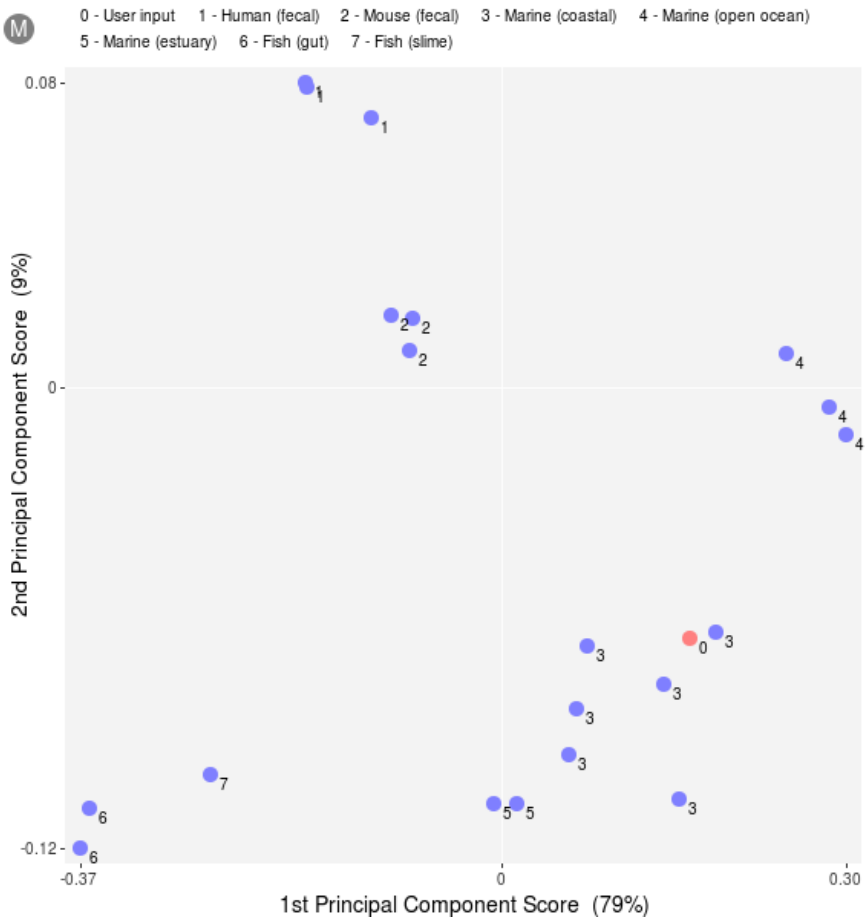



# Check for contamination

## PRINSEQ

- Dinucleotide (e.g., TA, GC, …) odds ratios
- Principal component analysis (PCA) to group metagenomes

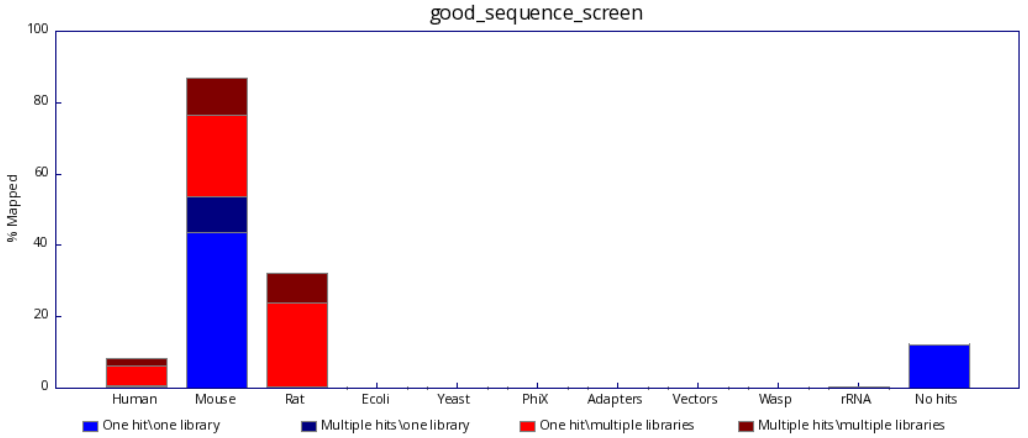


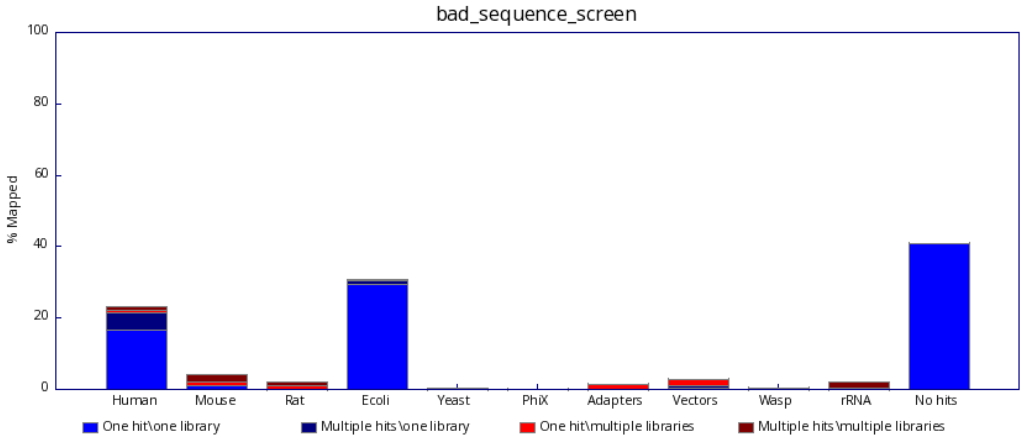
# Check for contamination

## FastQ Screen

Screen against a set of sequence databases:

- genomes of all of the organisms you work on
- PhiX
- Vectors
- …

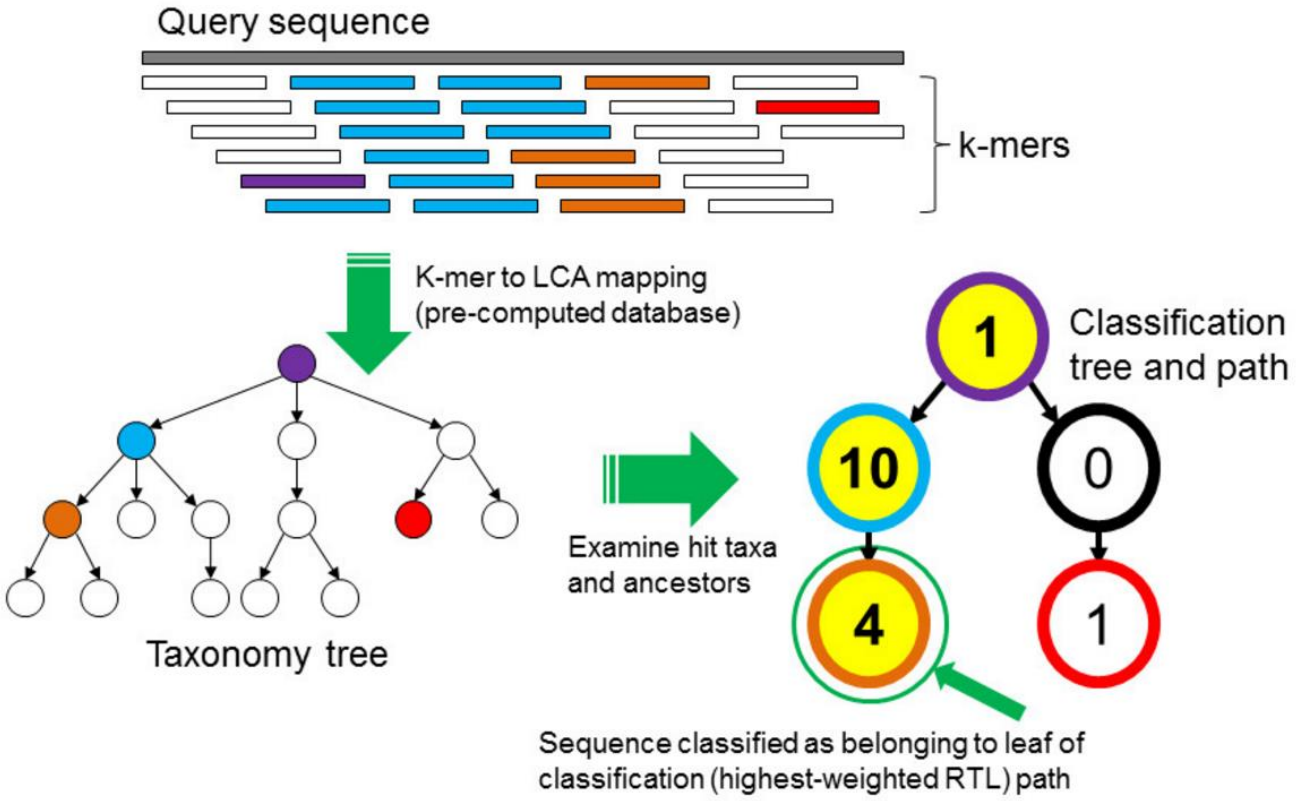


# Check for contamination

Kraken (https://ccb.jhu.edu/software/kraken/)

- Assigning taxonomic labels to short DNA sequences
- Detect metagenomics contaminations



# Phred quality score

Characterize the quality of DNA sequences

$$q = -10 \log_{10}(p)$$

p = error probability for the base

| Phred quality score | Probability | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

# Genome reference

# Genome reference

Reference genome is a “consensus” across all chromosomes of DNA pooled from multiple individuals

UCSC and Genome Reference Consortium (GRCh)

hg18, hg19, hg38 ←→ GRCh36, GRCh37, GRCh38

Newest version (hg38) release on Dez 24th 2013

Download (e.g.)

- http://hgdownload-test.cse.ucsc.edu/goldenPath/hg38/bigZips/

| | HG38 (UCSC) | GRCh38 |
|---|---|---|
| Prefix | Chr | - |
| Mitochondrial | chrM | MT |
| Order | chrM, chr1, chr2, …chrX, chrY | 1,2, …, X, Y, MT |

# Genome reference

**Indexing**

- Fai file (created by samtools faidx)
  contig, size, location, bases-per-line and for efficient random
- Dict file (created by Picard CreateSequenceDictionary)
  SAM style header describing the contents of the fasta file
- Different mapping programs

**Important**

- Choose one reference genome (well sorted, indexed) and stick to it
- Be sure that previous variant calls use same reference - otherwise convert coordinates (lift-over)

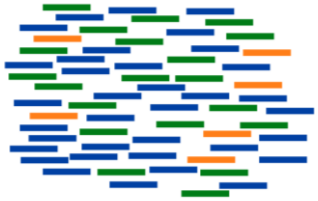
http://www.broadinstitute.org/gatk/guide/best-practices?bpm=DNAseq#data-processing-ovw

# Mapping and QC

SAM - Sequence Alignment/Map Format

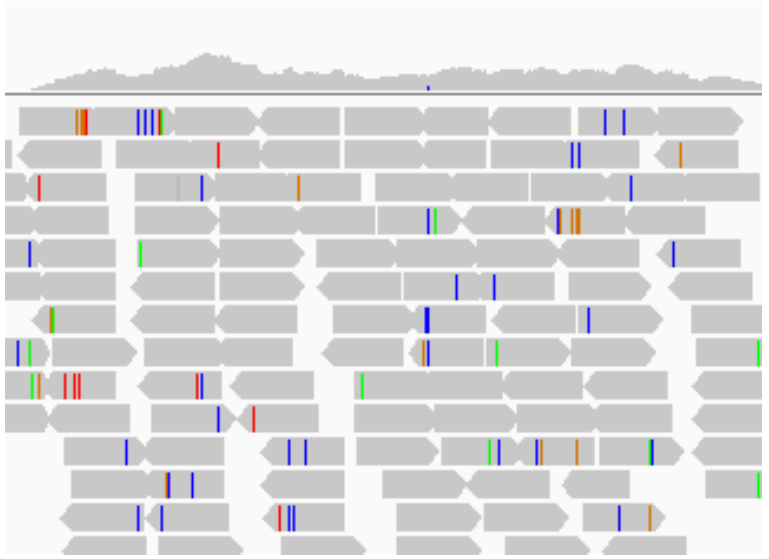
# Mapping - Principle


Short sequence reads

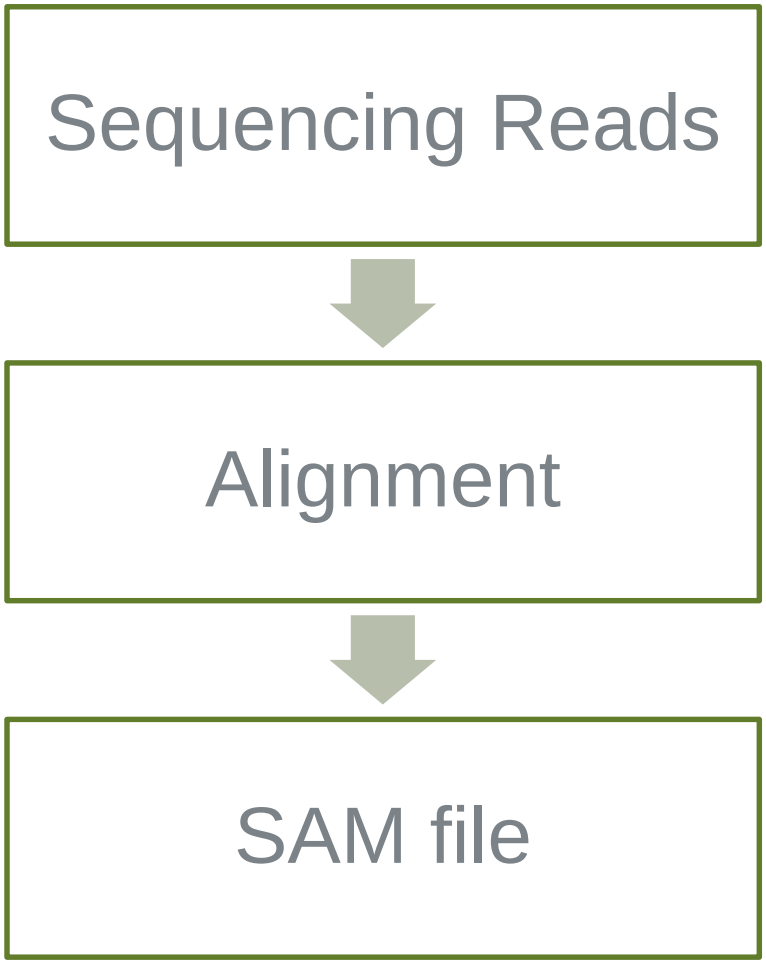


Alignment

Reference



Adapted from http://raetschlab.org//members/research/transcriptomics/images/RNA-Sequencing.png


Sequencing Reads
Alignment
SAM file

# Sequence **mapping** versus **alignment**

**Mapping:** (quickly) find the best possible loci to which a sequence could be aligned

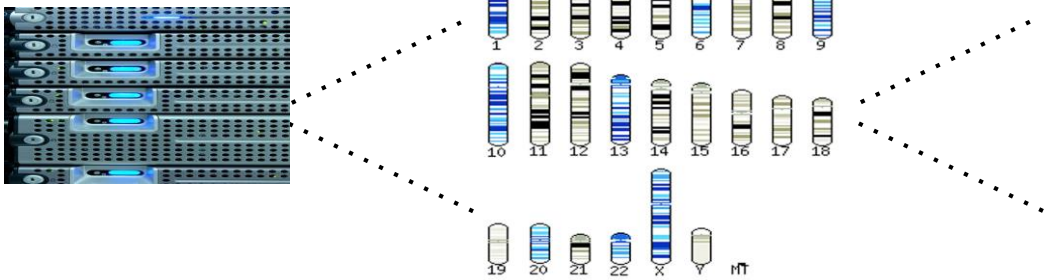
**Alignment:** for each locus to which a sequence can be mapped, determine the optimal base by base alignment of the query sequence to the reference sequence

# Paired-end sequencing

**1) A read-pair**

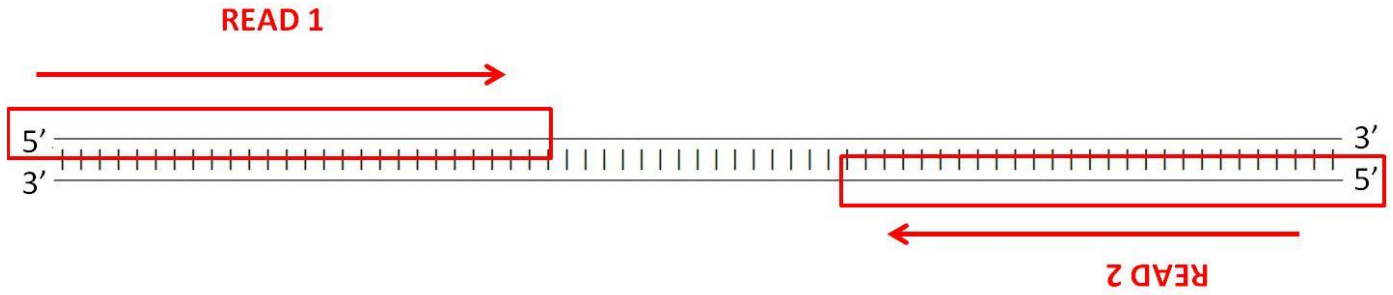
5’ GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC ----------//---------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGCAGC 5’

**2) A reference genome on a computer**



```
TATGGCAATTAAAATTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGCACAACAC
CGTGATGACATTGAAGTTGTAGGTATTAACGACTTAATCGACGTTGAATACATGGCTTATATGTTGAAAT
ATGATTCAACTCACGGTCGTTTCGACGGCACTGTTGAAGTGAAAGATGGTAACTTAGTGGTTAATGGTAA
AACTATCCGTGTAACTGCAGAACGTGATCCAGCAAACTTAAACTGGGGTGCAATCGGTGTTGATATCGCT
GTTGAAGCGACTGGTTTATTCTTAACTGATGAAACTGCTCGTAAACATATCACTGCAGGCGCAAAAAAAG
TTGTATTAACTGGCCCATCTAAAGATGCAACCCCTATGTTCGTTCGTGGTGTAAACTTCAACGCATACGC
AGGTCAAGATATCGTTTCTAACGCATCTTGTACAACAAACTGTTTAGCTCCTTTAGCACGTGTTGTTCAT
GAAACTTTCGGTATCAAAGATGGTTTAATGACCACTGTTCACGCAACGACTGCAACTCAAAAAACTGTGG
ATGGTCCATCAGCTAAAGACTGGCGCGGCGGCCGCGGTGCATCACAAAACATCATTCCATCTTCAACAGG
TATGGCAATTAAAATTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGCACAACAC
CGTGATGACATTGAAGTTGTAGGTATTAACGACTTAATCGACGTTGAATACATGGCTTATATGTTGAAAT
ATGATTCAACTCACGGTCGTTTCGACGGCACTGTTGAAGTGAAAGATGGTAACTTAGTGGTTAATGGTAA
AACTATCCGTGTAACTGCAGAACGTGATCCAGCAAACTTAAACTGGGGTGCAATCGGTGTTGATATCGCT
GTTGAAGCGACTGGTTTATTCTTAACTGATGAAACTGCTCGTAAACATATCACTGCAGGCGCAAAAAAAG
TTGTATTAACTGGCCCATCTAAAGATGCAACCCCTATGTTCGTTCGTGGTGTAAACTTCAACGCATACGC
AGGTCAAGATATCGTTTCTAACGCATCTTGTACAACAAACTGTTTAGCTCCTTTAGCACGTGTTGTTCAT
GAAACTTTCGGTATCAAAGATGGTTTAATGACCACTGTTCACGCAACGACTGCAACTCAAAAAACTGTGG
ATGGTCCATCAGCTAAAGACTGGCGCGGCGGCCGCGGTGCATCACAAAACATCATTCCATCTTCAACAGG
```

**3) Alignment of the read-pair to the reference genome gives coordinates describing where in the human genome the read-pair came from**



# SAM file format

The Sequence Alignment/Map (SAM) Format and SAMtools
Heng Li et al. Bioinformatics, 2009

Tab-delimited text file

Output of most alignment programs

- Header section
- Alignment section
- 11 Required columns
- Optional fields

# SAM Header

- Header lines start with @ symbol
- Always at top of file
- Contain lots of information about what was mapped, what it was mapped to, and how (metadata)
  - The version information for the SAM/BAM file
  - Whether or not and how the file is sorted
  - Information about the reference sequences
  - Any processing that was used to generate the various reads in the file
  - Software version

## 1.4 The alignment section: mandatory fields

In the SAM format, each alignment line typically represents the linear alignment of a segment. Each line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | `[!-?A-~]{1,255}` | Query template NAME |
| 2 | FLAG | Int | $[0, 2^{16}-1]$ | bitwise FLAG |
| 3 | RNAME | String | `\*|[!-()+-<>-~][!-~]*` | Reference sequence NAME |
| 4 | POS | Int | $[0, 2^{31}-1]$ | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | $[0, 2^{8}-1]$ | MAPping Quality |
| 6 | CIGAR | String | `\*|([0-9]+[MIDNSHPX=])+` | CIGAR string |
| 7 | RNEXT | String | `\*|=|[!-()+-<>-~][!-~]*` | Ref. name of the mate/next read |
| 8 | PNEXT | Int | $[0, 2^{31}-1]$ | Position of the mate/next read |
| 9 | TLEN | Int | $[-2^{31}+1, 2^{31}-1]$ | observed Template LENgth |
| 10 | SEQ | String | `\*|[A-Za-z=.]+` | segment SEQuence |
| 11 | QUAL | String | `[!-~]+` | ASCII of Phred-scaled base QUALity+33 |

# Example SAM file

AUSTRIAN INSTITUTE OF TECHNOLOGY

```
@PG     ID:bwa_sam      PN:bwa  PP:bwa_aln_fastq        VN:0.5.9-r16    CL:bwa sampe -a 1050 -r $rg_line -f $sam_file $reference_fasta $sai_file(s) $fastq_file(s)
@PG     ID:sam_to_fixed_bam     PN:samtools     PP:bwa_sam      VN:0.1.17 (r973:277)    CL:samtools view -bSu $sam_file | samtools sort -n -o - samtools_nsort_tmp | s
ort_tmp | samtools fillmd -u - $reference_fasta > $fixed_bam_file
@PG     ID:gatk_target_interval_creator PN:GenomeAnalysisTK     PP:sam_to_fixed_bam     VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysisTK.jar -T Realig
dels_file(s)
@PG     ID:bam_realignment_around_known_indels  PN:GenomeAnalysisTK     PP:gatk_target_interval_creator VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysis
bam_file -targetIntervals $intervals_file -known $known_indels_file(s) -LOD 0.4 -model KNOWNS_ONLY -compress 0 --disable_bam_indexing
@PG     ID:bam_count_covariates PN:GenomeAnalysisTK     PP:bam_realignment_around_known_indels  VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysisTK.jar -
recal_data.csv -knownSites $known_sites_file(s) -l INFO -L '1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;X;Y;MT' -cov ReadGroupCovariate -cov QualityScor
@PG     ID:bam_recalibrate_quality_scores       PN:GenomeAnalysisTK     PP:bam_count_covariates VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysisTK.jar -
.csv -I $bam_file -o $recalibrated_bam_file -l INFO -compress 0 --disable_bam_indexing
@PG     ID:bam_calculate_bq     PN:samtools     PP:bam_recalibrate_quality_scores       VN:0.1.17 (r973:277)    CL:samtools calmd -Erb $bam_file $reference_fasta > $
@PG     ID:bam_merge    PN:picard       PP:bam_calculate_bq     VN:1.53 CL:java $jvm_args -jar MergeSamFiles.jar INPUT=$bam_file(s) OUTPUT=$merged_bam VALIDATION_STR
@PG     ID:bam_mark_duplicates  PN:picard       PP:bam_merge    VN:1.53 CL:java $jvm_args -jar MarkDuplicates.jar INPUT=$bam_file OUTPUT=$markdup_bam_file ASSUME_SOR
@PG     ID:bam_merge.1  PN:picard       PP:bam_mark_duplicates  VN:1.53 CL:java $jvm_args -jar MergeSamFiles.jar INPUT=$bam_file(s) OUTPUT=$merged_bam VALIDATION_STR
@CO     $known_indels_file(s) = ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.indels_mills_devine_hg19_leftAligned_co
@CO     $known_indels_file(s) .= ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.low_coverage_vqsr.20101123.indels.site
@CO     $known_sites_file(s) = ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.dbsnp.build135.snps.sites.vcf.gz
SRR070823.24480225      163     11      60464   0       100M    =       60720   355     CCATGGTTAAACATAAATGCAAATATGTTAATGTTTACTGAATAACTTATCTGTGCCAAGTGGTGTATTAATGATTC
MNKJKMKNJNNNMGNDNNGNGGKKGNKNKKJNMILLLLCKLLFKKLBLHJGFHED    X0:i:3  X1:i:6  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@I@@@@@@@@@@@@
@@@@@@@@@@@@@GI
SRR070823.24480518      1187    11      60464   0       100M    =       60720   355     CCATGGTTAAACATAAATGCAAATATGTTAATGTTTACTGAATAACTTATCTGTGCCAAGTGGTGTATTAATGATTC
DJKBEIJBEJFIJCJGKEGKGCHIG@HIHDHNLKCJLIAKLLIKMIJLFBEE?EB    X0:i:3  X1:i:6  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@E@@@@@@@@@@@
@@@@@@@@@@@@@GG
SRR070823.24480225      83      11      60720   0       100M    =       60464   -355    TATGGAGTTTGATGTTATGTCAGGGTAATTACATGATTATATAATTAACAGGTTTCTTTTTAAATCAGCTATATCAA
GHHKKJJJJGJJJJJGHHGGJJJJGGGGGJGGGJIHGGHGGGGGHIEJIIEDCF6    X0:i:9  X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.24480518      1107    11      60720   0       100M    =       60464   -355    TATGGAGTTTGATGTTATGTCAGGGTAATTACATGATTATATAATTAACAGGTTTCTTTTTAAATCAGCTATATCAA
=EGJJF>?@DBADDD>BBB<AD:@2B@BABBC>B;::::E@DBDG;6E5=A??@6    X0:i:9  X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070531.23281260      99      11      61942   0       100M    =       62035   192     TCCATGCCGCTTTGATGACAGGATAATATGTAAGCTTTTCTATATTTCAGAAACTATATGACATGACGAAAAGTAAA
JJMLKHLIJMJHJKNHFGKNHMGGMJHJFGIGIINOLNIKMNENNIFILKGGHEA    X0:i:8  X1:i:1  MD:Z:100        RG:Z:SRR070531  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070531.23281260      147     11      62035   0       100M    =       61942   -192    GGAGGTATCCTGAATTGACTGAGAAATAAGGAGGTATTCCACAGAGAATATAAATAAACATATACTTAGTGTTCAAG
MKKJKJKKCJKKJNKJJJJMLJMLLLLIMJMMLJJIJIIJIIFIKKKKHGJECG6    X0:i:8  X1:i:2  MD:Z:100        RG:Z:SRR070531  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:CB@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.17243685      99      11      62388   0       100M    =       62452   163     CCTTGCCAATTGTGTTCTCCTTTATTTCCTGCTGGATATGACCACTGTCCTTCCATTGCATTGTATGTGTTTTTTAA
MGHMMMKIMLKHHJEKKFFDHJEKDEIDCCDGF?IEGJHIDEHJIGE?GAFDCCC    X0:i:10 X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@E@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.17243685      147     11      62452   0       100M    =       62388   -163    ATGTGTTTTTTAATAGACTTTAATGGTTCTCAAGTGATGCATTATTTAGTTTGGTTCTTTGAAACTTATATAAATGA
MJNMMMLJMLJNLLIMKKJFLJGGJIJJJLLJJGLIJKKKJKHLKIKHJGJHDA6    X0:i:10 X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.1968642       163     11      62725   0       100M    =       62918   292     TTTTTAACCATTACAAACAATGCTGCTATGAACATTCTTTTGTAAATCACCTGGTTCATATGTGCAAGATATCCTCT
LKMLMNNNNHKMMKJHNHNNMMNHHKKNNKNLNNNMGHIFIEIIIGJDHHGCCBD    X0:i:9  X1:i:1  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@EDDB@@@@@@@@@
@@@@@@@@@@@@AAF
```

# 2 FLAG

- Bitwise
- Picard (online) – explain flags

**Examples**

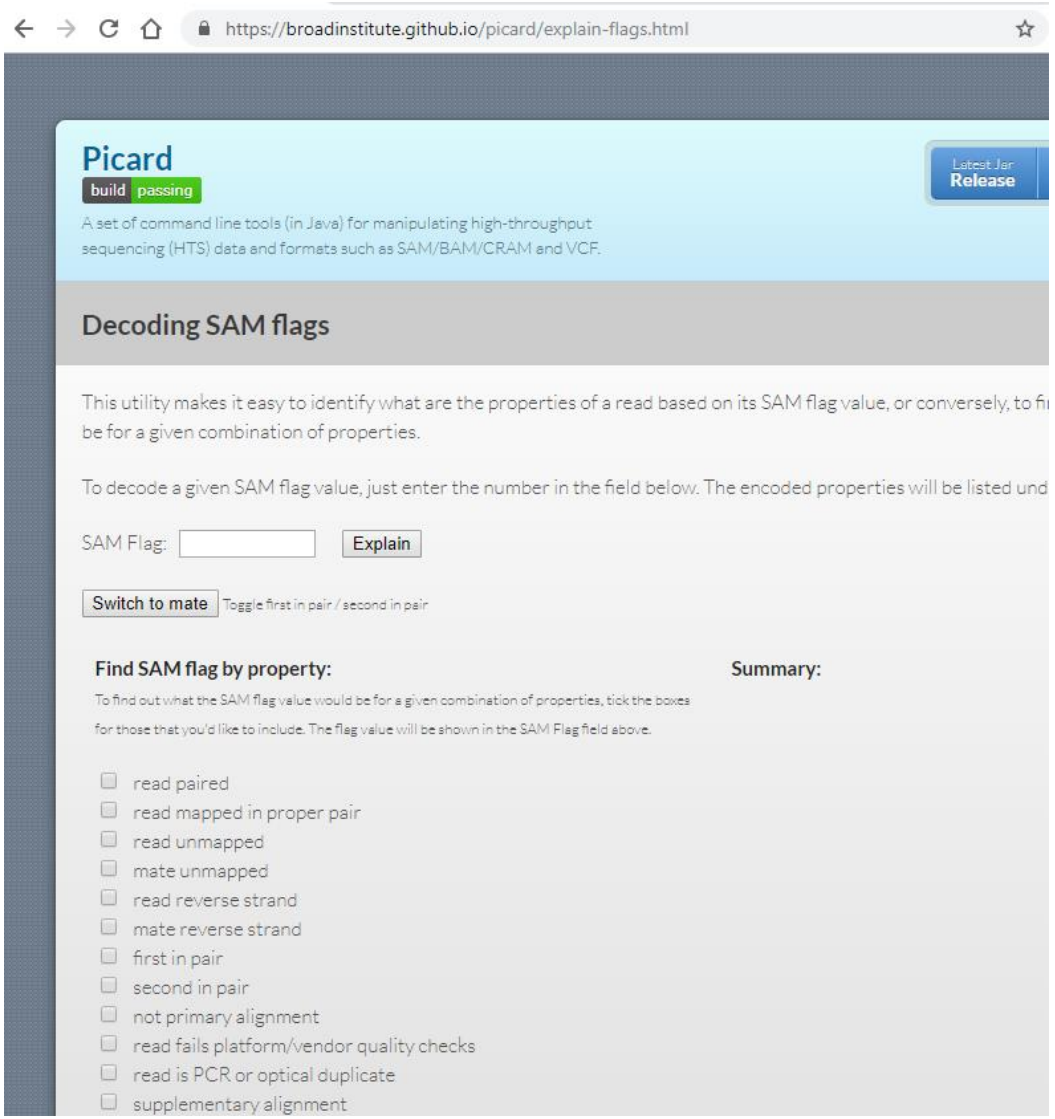
1 = 0001 -> PE read
4 = 0100 -> Unmappable
5 = 0101 -> Unmapped PE

# 2 FLAG

- Bitwise
- Picard (online) – explain flags

**Examples**

1 = 0001 -> PE read
4 = 0100 -> Unmappable
5 = 0101 -> Unmapped PE



# 3 RNAME & 4 POS & 5 MAPQ

AUSTRIAN INSTITUTE OF TECHNOLOGY

**RNAME**

- Reference name
  FASTA sequence name (e.g.: chr4)

**POS**

- Mapping position
  leftmost position in reference (e.g.: 142345)
  ! Reverse strand

**MAPQ**

- Mapping quality
- Phred score
- Depends on mapping program

# 6 CIGAR

Used as a compact way to represent sequence alignment

Ref `ACTCAGTG--GT`
Read `ACGC-TGCAGTTATATAGG`
Cigar 4M 1D 3M 2I 2M 7S

?

← spaces just for visualization

| Op | BAM | Description |
|---|---|---|
| M | 0 | alignment match (can be a sequence match or mismatch) |
| I | 1 | insertion to the reference |
| D | 2 | deletion from the reference |
| N | 3 | skipped region from the reference |
| S | 4 | soft clipping (clipped sequences present in SEQ) |
| H | 5 | hard clipping (clipped sequences NOT present in SEQ) |
| P | 6 | padding (silent deletion from padded reference) |
| = | 7 | sequence match |
| X | 8 | sequence mismatch |

# CIGAR

Ref `TCGATCTTAGC---TACGATC`

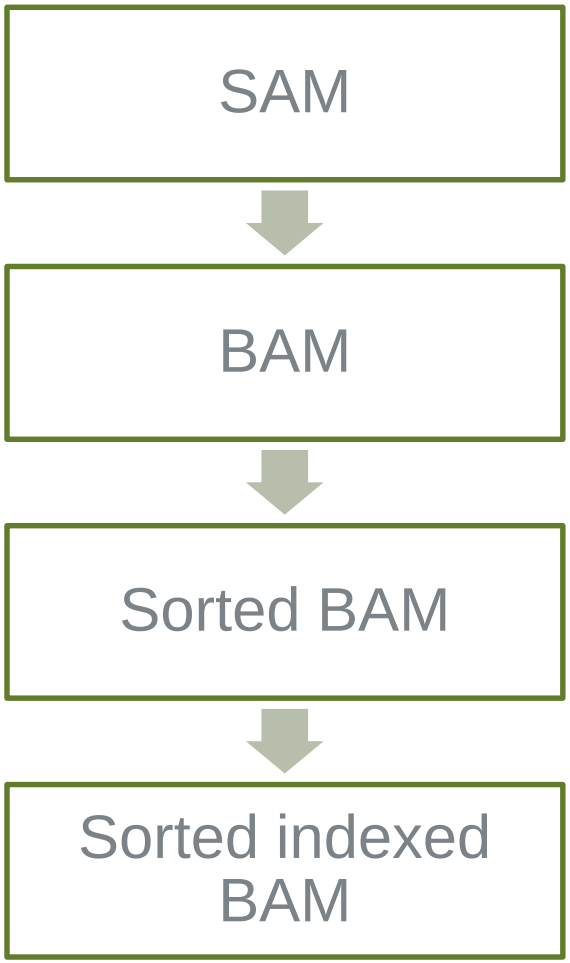
Read `TCAG---TAACTACTACGATC`

**?**

4M 3D 4M 3I 7M

# BAM – Binary SAM

# SAM / BAM




Convert (view)

sort

index

# BAM

**SAM**

- Information on the alignment of each read
- Optimized for readability and sequential access

**BAM (Binary SAM)**

- Compressed -> saves disk space; with BGZF (Blocked GNU Zip Format) - a variant of GZIP
- Can be sorted & indexed - quick viewing/searching (bigger than GZIP files)
- Cannot be read without a tool (samtools)

**uBAM**

- unmapped BAM → compress FASTQ files

**CRAM**

- Better lossless compression than BAM
- Cramtools for conversion from/to BAM
- http://www.ebi.ac.uk/ena/about/cram_toolkit

# Coordinate systems

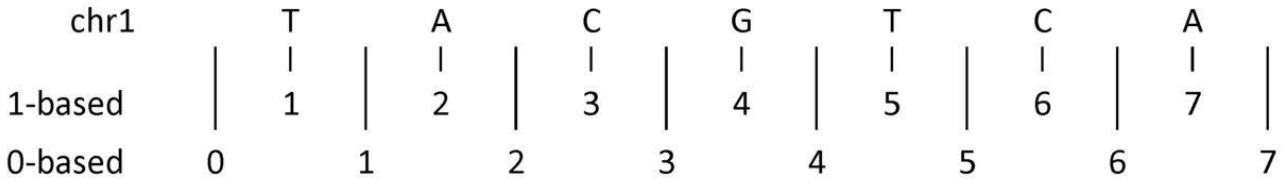
# Coordinate system

- 0 based → 0, 1, 2, … 9 | 1 based → 1, 2, 3, … 10

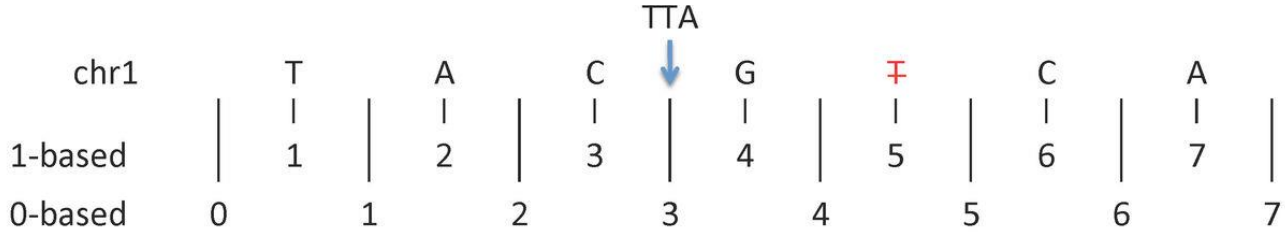
- BED – 0 based
- GFF – 1 based
- Ensembl uses a one-based coordinate system - UCSC use a zero-based coordinate system

| | 1 based | 0 based |
|---|---|---|
| Third element | 3 | 2 |
| First ten | 1, 10 | 0, 10 |
| Second ten | 11, 20 | 10, 20 |
| One base long at 10 | 10,10 | 9,10 |
| Interval | end – start + 1 | end – start |
| Five elements at 100 | 100, 104 | 99, 104 |

# Coordinate system




| | 1-based | 0-based |
|---|---|---|
| Indicate a single nucleotide | chr1:4-4 G | chr1:3-4 G |
| Indicate a range of nucleotides | chr1:2-4 ACG | chr1:1-4 ACG |
| Indicate a single nucleotide variant | chr1:5-5 T/A | chr1:4-5 T/A |




| | 1-based | 0-based |
|---|---|---|
| Indicate a deletion | chr1:5-5 T/- | chr1:4-5 T/- |
| Indicate an insertion | chr1:3-4 -/TTA | chr1:3-3 -/TTA |

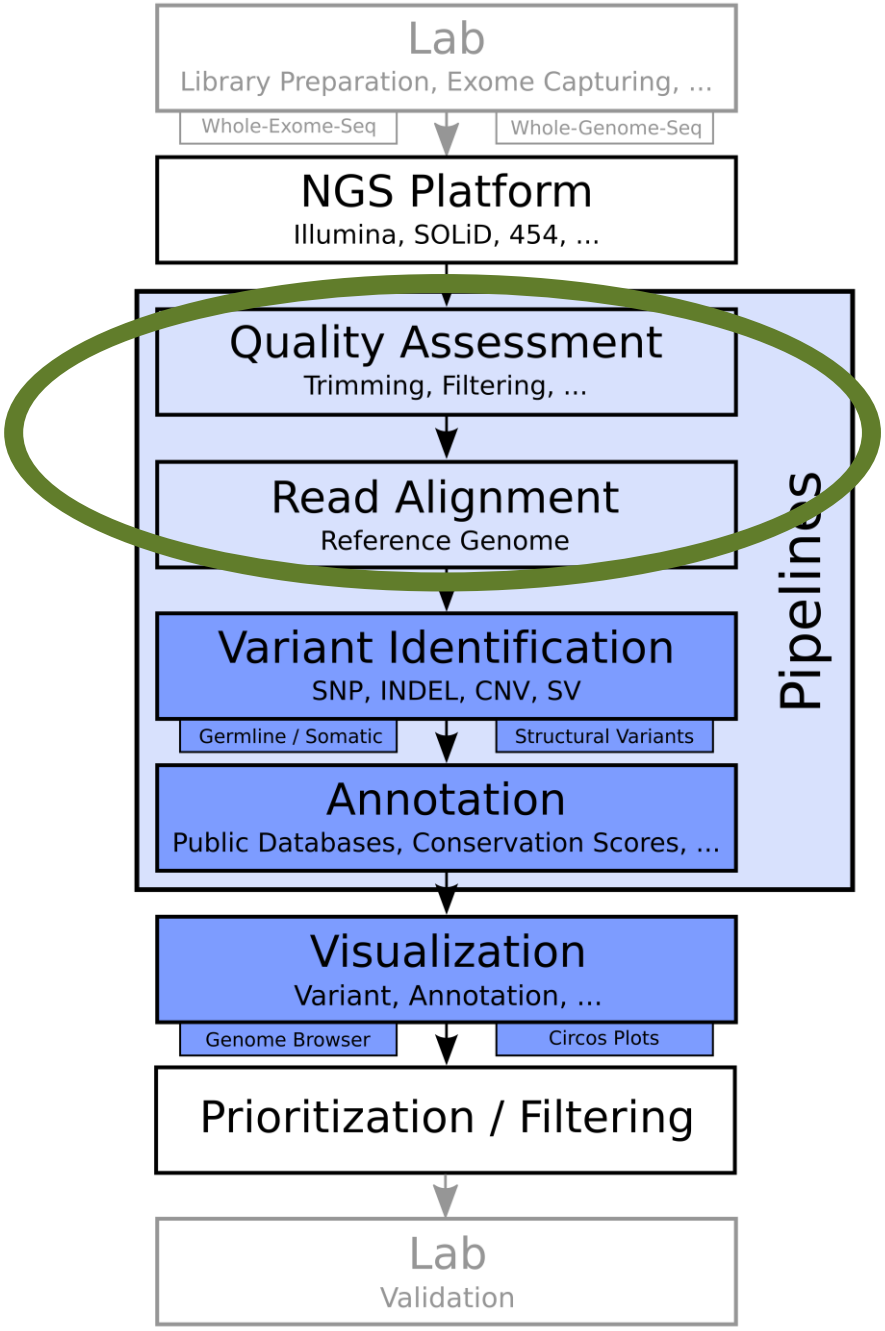
https://www.biostars.org/p/84686/ https://www.biostars.org/p/6373/

# Conversion tool

**convert_zero_one_based**

- Python CLI
- convert between zero and one based coordinate systems
- Use: `convert_zero_one_based --help`

https://github.com/griffithlab/convert_zero_one_based


Lab
Library Preparation, Exome Capturing, ...
Whole-Exome-Seq
Whole-Genome-Seq
NGS Platform
Illumina, SOLiD, 454, ...
Quality Assessment
Trimming, Filtering, ...
Read Alignment
Reference Genome
Variant Identification
SNP, INDEL, CNV, SV
Germline / Somatic
Structural Variants
Annotation
Public Databases, Conservation Scores, ...
Pipelines
Visualization
Variant, Annotation, ...
Genome Browser
Circos Plots
Prioritization / Filtering
Lab
Validation

# Quality check of alignment

Based on SAM/BAM files

Detect biases in the sequencing and/or mapping

**Metrics**

- Coverage / nucleotide distribution
- Reads mapped outside of a target (e.g., Exome sequencing)
- Number of mapped reads (wrong reference genome?)
- Insert size statistics
- Mapping quality - rule of thumb: Anything less than Q20 is not useful data

**Tools**

- Qualimap 2
  http://qualimap.bioinfo.cipf.es/
- bamstats
  http://bamstats.sourceforge.net/

# Read Group Tag (RG)

**RG - Meta information**

- ID: unique e.g., SRA number (Sequence read archive)
- PL: Sequencing platform
- PU: Platform unit (run name / flowcell-barcode-lane)
- LB: Library name
- PI: Insert size (Predicted mean insert size)
- SM: Sample (Individual)
- CN: Sequencing center

**PICARD**

Use Picard to add RG tags to your SAM files

# Read Group Tag – why important?

- It refers to a set of reads that were generated from a single run of a sequencing instrument.
- When multiplexing is involved, then each subset of reads originating from a separate library run on that lane will constitute a separate read group.
- To see the read group information for a BAM file, use the following command:

```
samtools view -H sample.bam | grep '@RG’
```

- Check that your FASTQ files have appropriate RQ when performing demultiplexing
- Important to have a correct RG tag → required by bioinformatics analysis tools (GATK, …)

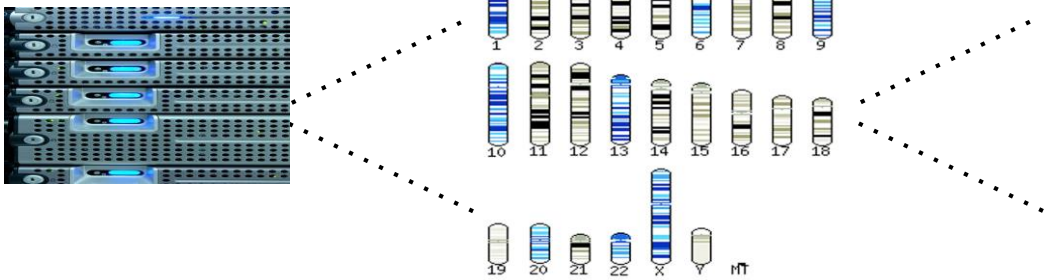
# Insert Size

# RECAP - Paired-end sequencing

## 1) A read-pair

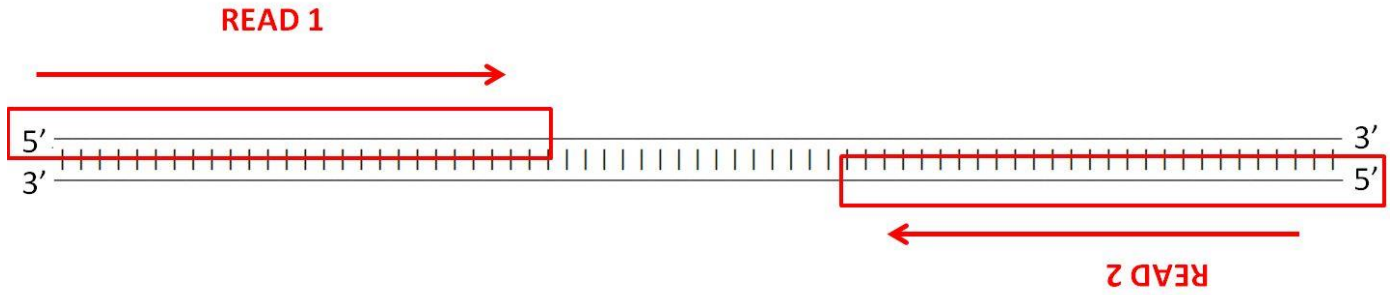
5’ GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC ----------//---------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGCAGC 5’

## 2) A reference genome on a computer

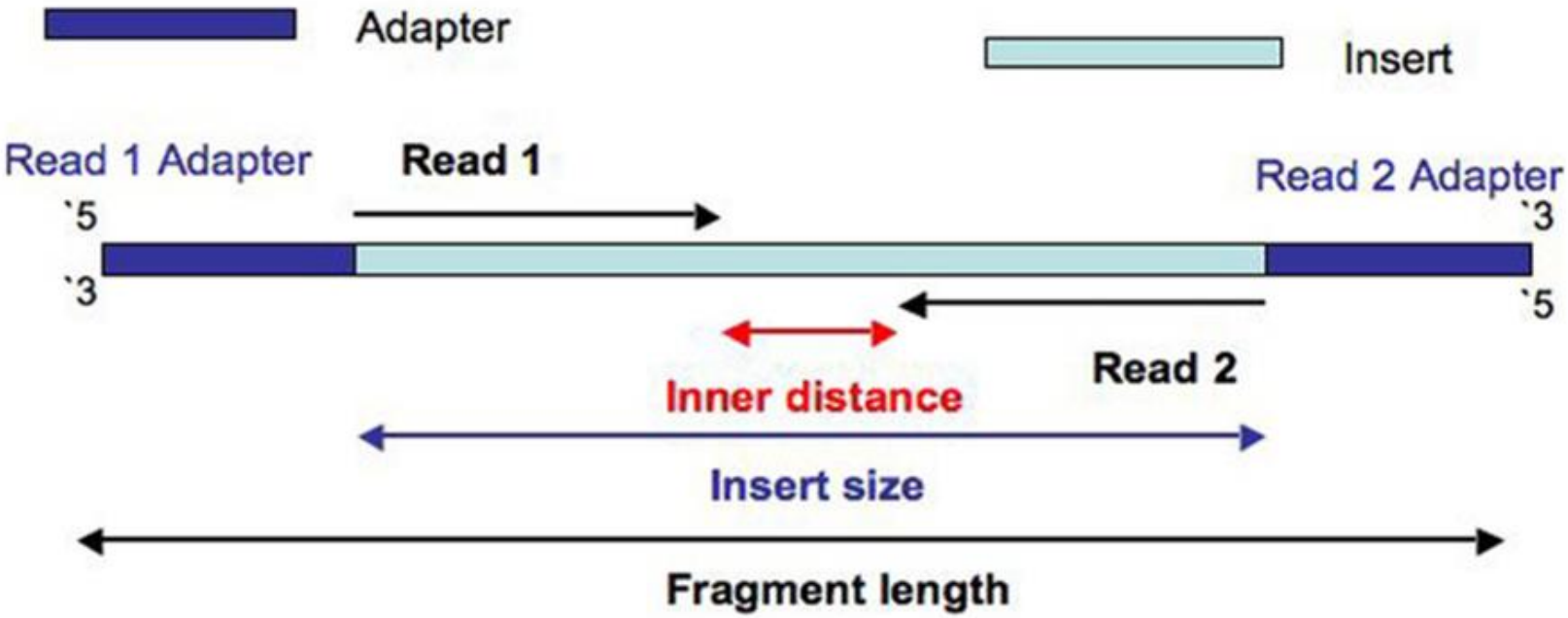


```
TATGGCAATTAAAATTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGCACAACAC
CGTGATGACATTGAAGTTGTAGGTATTAACGACTTAATCGACGTTGAATACATGGCTTATATGTTGAAAT
ATGATTCAACTCACGGTCGTTTCGACGGCACTGTTGAAGTGAAAGATGGTAACTTAGTGGTTAATGGTAA
AACTATCCGTGTAACTGCAGAACGTGATCCAGCAAACTTAAACTGGGGTGCAATCGGTGTTGATATCGCT
GTTGAAGCGACTGGTTTATTCTTAACTGATGAAACTGCTCGTAAACATATCACTGCAGGCGCAAAAAAAG
TTGTATTAACTGGCCCATCTAAAGATGCAACCCCTATGTTCGTTCGTGGTGTAAACTTCAACGCATACGC
AGGTCAAGATATCGTTTCTAACGCATCTTGTACAACAAACTGTTTAGCTCCTTTAGCACGTGTTGTTCAT
GAAACTTTCGGTATCAAAGATGGTTTAATGACCACTGTTCACGCAACGACTGCAACTCAAAAAACTGTGG
ATGGTCCATCAGCTAAAGACTGGCGCGGCGGCCGCGGTGCATCACAAAACATCATTCCATCTTCAACAGG
TATGGCAATTAAAATTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGCACAACAC
CGTGATGACATTGAAGTTGTAGGTATTAACGACTTAATCGACGTTGAATACATGGCTTATATGTTGAAAT
ATGATTCAACTCACGGTCGTTTCGACGGCACTGTTGAAGTGAAAGATGGTAACTTAGTGGTTAATGGTAA
AACTATCCGTGTAACTGCAGAACGTGATCCAGCAAACTTAAACTGGGGTGCAATCGGTGTTGATATCGCT
GTTGAAGCGACTGGTTTATTCTTAACTGATGAAACTGCTCGTAAACATATCACTGCAGGCGCAAAAAAAG
TTGTATTAACTGGCCCATCTAAAGATGCAACCCCTATGTTCGTTCGTGGTGTAAACTTCAACGCATACGC
AGGTCAAGATATCGTTTCTAACGCATCTTGTACAACAAACTGTTTAGCTCCTTTAGCACGTGTTGTTCAT
GAAACTTTCGGTATCAAAGATGGTTTAATGACCACTGTTCACGCAACGACTGCAACTCAAAAAACTGTGG
ATGGTCCATCAGCTAAAGACTGGCGCGGCGGCCGCGGTGCATCACAAAACATCATTCCATCTTCAACAGG
```

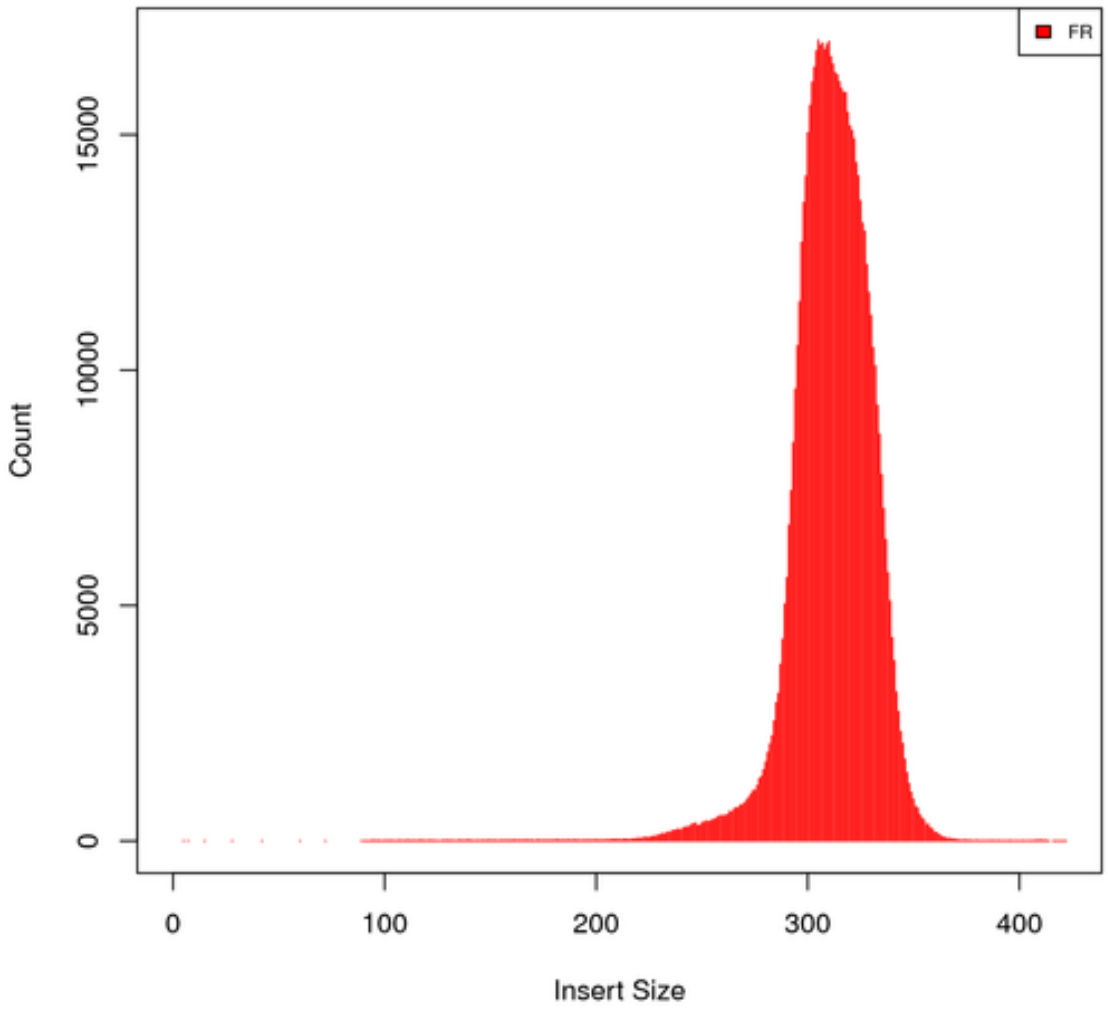
## 3) Alignment of the read-pair to the reference genome gives coordinates describing where in the human genome the read-pair came from



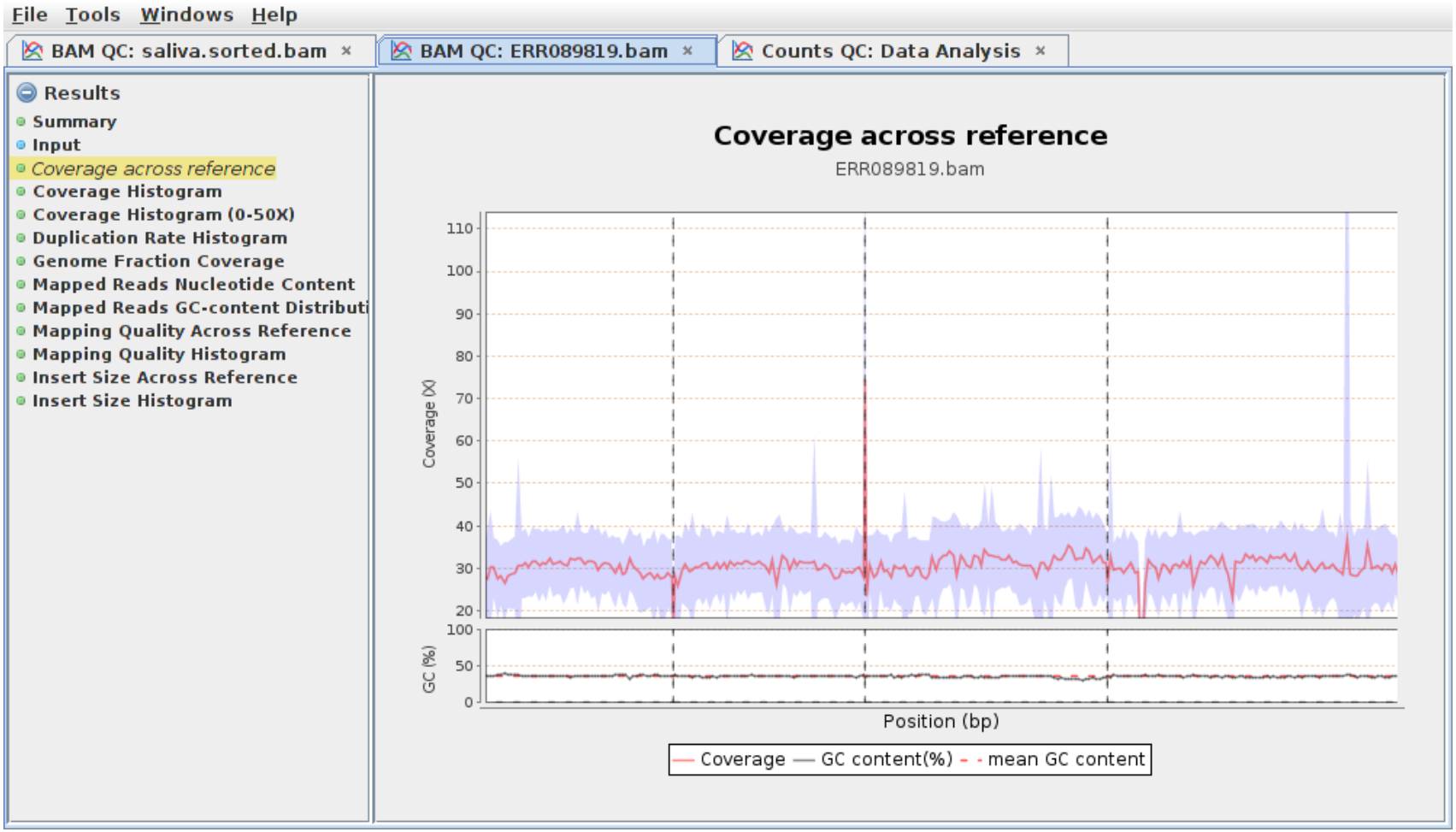
# Insert size


Adapter
Insert
Read 1 Adapter
Read 1
Read 2 Adapter
`5
`3
`3
`5
Read 2
Inner distance
Insert size
Fragment length

# Insert size histogram



# Qualimap 2


File Tools Windows Help
BAM QC: saliva.sorted.bam
BAM QC: ERR089819.bam
Counts QC: Data Analysis
Results
Summary
Input
Coverage across reference
Coverage Histogram
Coverage Histogram (0-50X)
Duplication Rate Histogram
Genome Fraction Coverage
Mapped Reads Nucleotide Content
Mapped Reads GC-content Distributi
Mapping Quality Across Reference
Mapping Quality Histogram
Insert Size Across Reference
Insert Size Histogram
Coverage across reference
ERR089819.bam
Coverage (X)
GC (%)
Position (bp)
Coverage
GC content(%)
mean GC content

# Read duplicates

AUSTRIAN INSTITUTE OF TECHNOLOGY

**Origin**

- PCR amplification step in library preparation
  1. Get DNA pieces (shatter / enrich DNA)
  2. Ligate adapters to both ends of the fragments
  3. PCR amplify the fragments with adapters
  4. Put fragments on beads or across flowcells
  5. Amplify fragments
  6. Sequence

**Identification**

- Have the same starting position

http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965708/

# Read duplicates

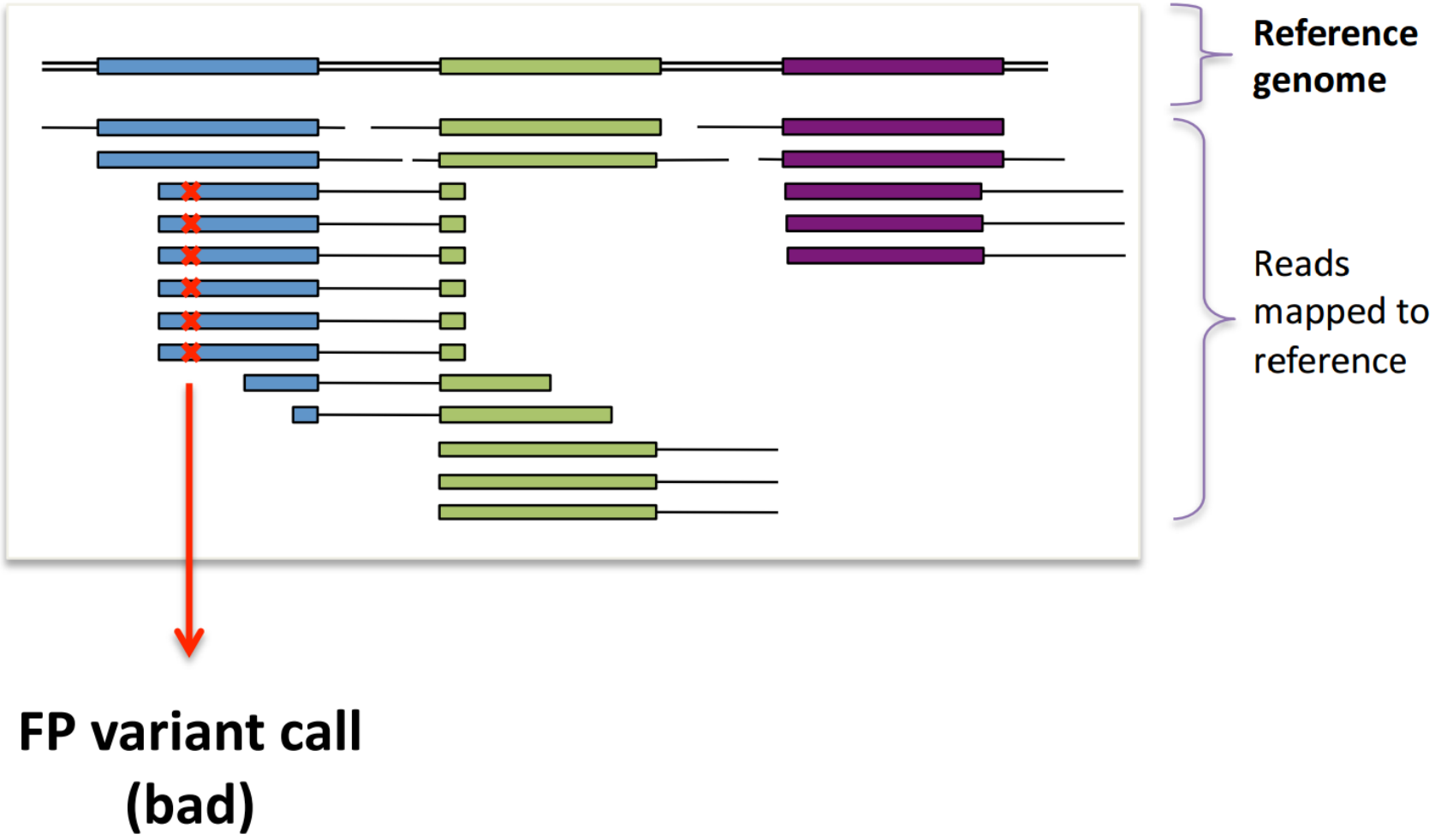
**Problems/causes**

- More steps during PCR amplification with little input material → more duplicates
  - This may result in duplicate DNA fragments in the final library

- Higher rates (~30%) arise when too little starting material is used → more amplification of the library is needed

- May result in false SNP calls (statistical model gets mixed up)

# Read duplicates

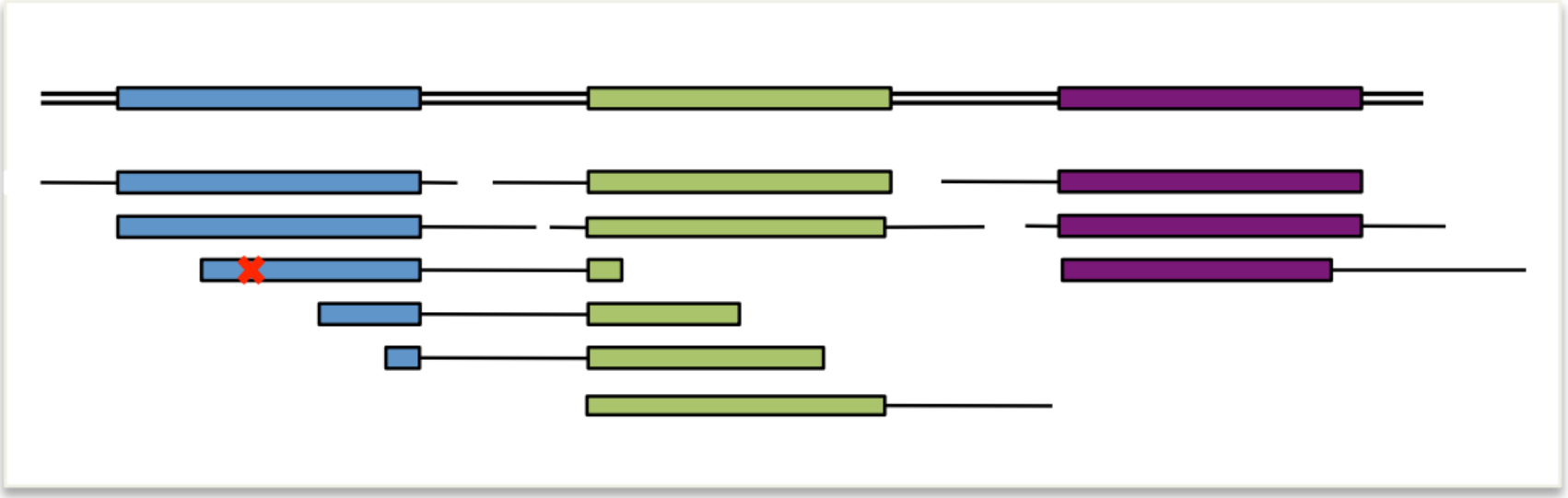



http://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-1-Map_and_Dedup.pdf

# Read duplicates - removal

- → Identify reads that map to the same location
- → Remove all but one

After marking duplicates



# Read duplicates - removal

**Attention!**

**Do not remove for**

- Haloplex enrichment (nonrandom fragmentation method)
- PCR based enrıchment

**! You would remove results**

# Base Quality Score Recalibration

- Various sources of systematic error
  → over / under estimated base quality scores

- Quality score assigned to single base in isolation
  (assigned by the sequencing machines)

- Variant calling algorithms rely heavily on base quality scores

**Solution** → Correct base quality scores

# Base Quality Score Recalibration

Apply machine learning to model these errors empirically and adjust the quality scores accordingly

- First the program builds a model of covariation based on the data
  - Reported quality score
  - Position in the read (cycle)
  - Preceding and current base – sequence context (homopolymer, ...)
- and a set of known variants (1000g, dbSNP, large private cohort)
  - discount most of the real genetic variation

- First pass: calculate new QS based on the model
  Second pass: adjust the base quality scores

→ Visual inspection with before/after plots

Good explanation: http://zenfractal.com/2014/01/25/bqsr/

# Base Quality Score Recalibration

**WES**

- For WES restrict to capture targets
  off-target sites are likely to have higher error rates

**Organism with no known variants?**

- Call variants -> apply stringent filter -> use these for recalibration
- Repeat previous steps

https://www.broadinstitute.org/gatk/events/slides/1506/GATKwr8-B-3-Base_recalibration.pdf

# Informative NGS QC page

# Working with SAM/BAM files

# SAMTOOLS

http://samtools.sourceforge.net

**View**

```
samtools view –h <file.bam>
samtools view <file.bam> chr2:20,100,000-20,200,000
samtools view –f 0x02 <file.bam> > <only_proper_paired.sam>
```

**Sort**

```
samtools sort -o <sorted.bam> <aln.bam>
```

**Index**

```
samtools index <sorted.bam>
(only for BAM)
```

# SAMTOOLS

**Simple stats (reads mapped, reads paired, ...)**

```
samtools flagstat <file.bam>
```

**Stats for each chr/contig - reads mapped and unmapped**

```
samtools idxstats <sorted.bam> (and indexed)
```

**Convert SAM to BAM**

```
samtools view -S -h -b <aln.sam> > <aln.bam>
```

**Converting SAM to a sorted BAM file (without intermediate file)**

```
samtools view -Shb <file.sam> | samtools sort -o
file_sorted.bam -
```

http://davetang.org/wiki/tiki-index.php?page=SAMTools

# Tools for SAM/BAM files

# Other tools

**SAMBAMBA** - „Multithreaded" SAMtools - https://github.com/lomereiter/sambamba

- view
- sort
- index
- merge
- flagstat
- markdup

**elprep -** https://github.com/exascience/elprep

- High-performance tool for preparing .sam / .bam / .cram files
- In-memory and multi-threaded application
- Requires lots of memory (WGS ~256GB)
- Replacement for SAMTOOLS & Picard

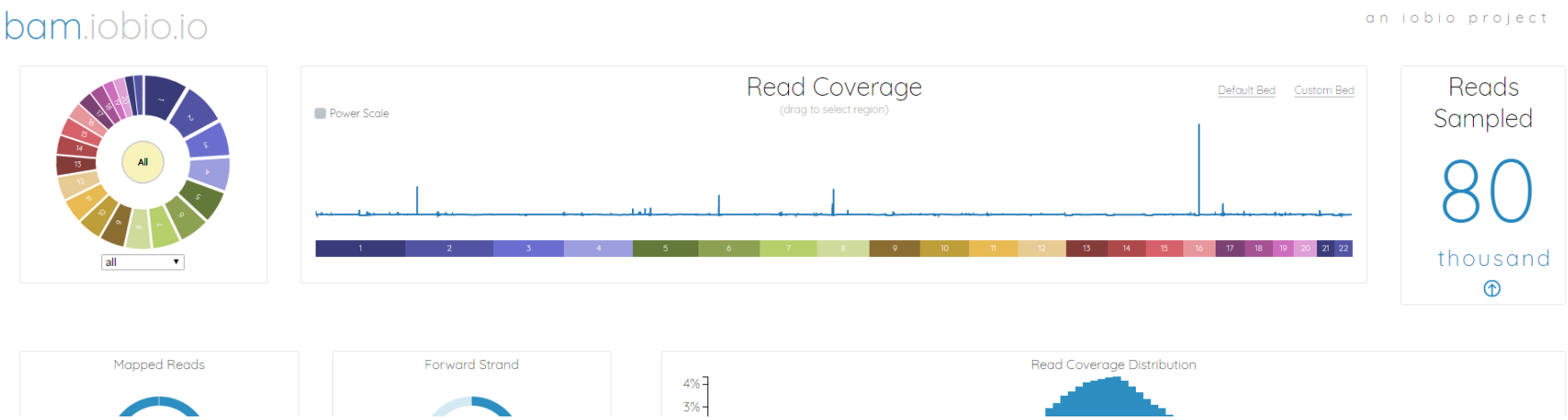
# PICARD & iobio

**PICARD**

- JAVA based tool (http://picard.sourceforge.net/)
- BuildBamIndex, FastqToSam, MergeSamFiles, …

**bam.iobio.io**

- Web-based (http://bam.iobio.io)
- Coverage overview
- Mapping overview



# Practicals – Day 1

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

# Bioinformatics and Genome Analyses

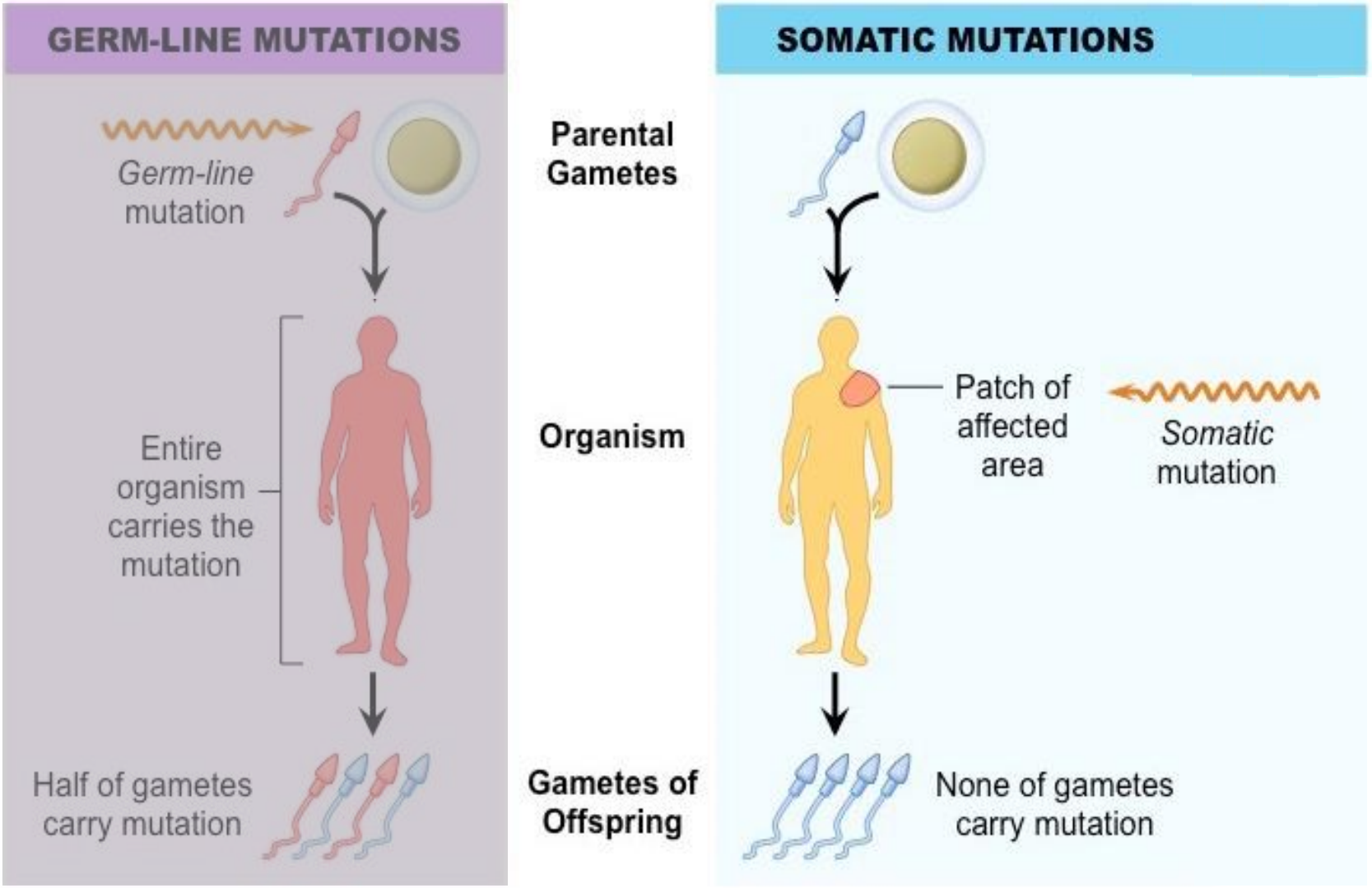
## Tools for variant analysis of next-generation genome sequencing data

## Lecture 2

Stephan Pabinger
stephan.pabinger@ait.ac.at

# Genetic variations

# Somatic mutations



# Types of genetic variations

**SNV / SNP**

- A single nucleotide — A, T, C or G — in the genome differs between members of a population or chromosome pairs
- Originally defined as occurring at least in one individual of the population (these definitions may shift in time)
- SNV (single nucleotide variant) if observed very rarely
- SNP, SNV → may fall within
  - coding sequences of genes
  - non-coding regions of genes
  - intergenic regions

# Types of genetic variations

## INDEL

- Insertion / deletion of bases
- Coding regions of the genome - produce a frameshift mutation (unless multiple of 3)
- There are approximately 190-280 frameshifting INDELs in each person.
  A map of human genome variation from population-scale sequencing". Nature 467 (7319)

# Types of genetic variations

**Structural variations (SV)**

- Variation in structure of an organism's chromosome

- Insertions
  Deletions
  CNV
  Inversions
  Translocations

# Human genome

- Typical genome differs from the reference human genome at **4.1 million to 5.0 million sites**.

- **~99.9%** of variants consist of **SNVs and INDELs**

- Structural variants affect more bases
  - 2,100 to 2,500 structural variants (~1,000 large deletions, ~160 copy-number variants, ~1100 insertions)
  - Affecting ~20 million bases of sequence

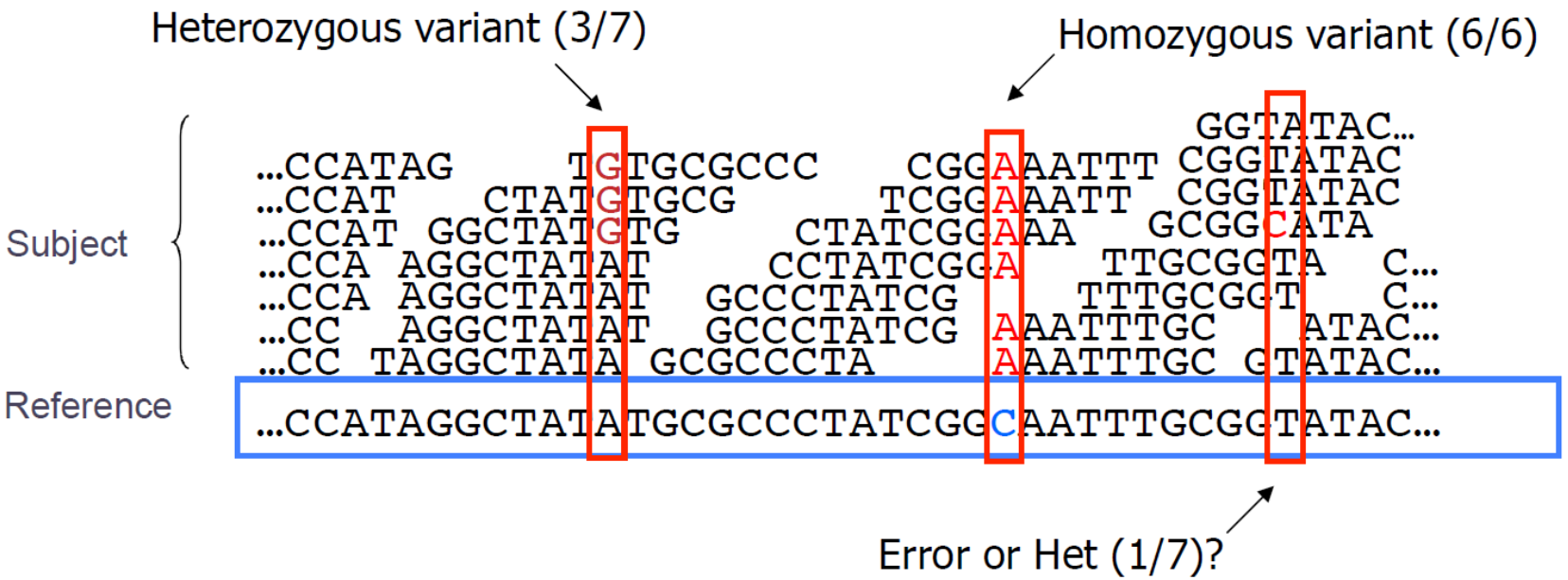
- African ancestry populations harbor the greatest numbers of variant sites (as predicted by the out-of-Africa model of human origins)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/

The 1000 genome project; A global reference for human genetic variation; Nature 2015

# Variant Calling

# Genotyping theory




- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!
- Sequencing instruments make mistakes
  – Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

# (Other) reasons for a mismatch

- True SNP

OR

- Error generated in library preparation
- Base calling error
  - May be reduced by improved base calling methods, but cannot be eliminated
- Misalignment (mapping error)
  - Local realignment to improve mapping
- Error in reference genome sequence

# More difficulties

- High depth - low quality regions → likely due to copy number or other larger structural events
- Repetitive regions → artefactual variants
- Regions of low complexity (~ 2% of genome)
  polypurine (AG), AT-rich regions, simple tandem repeats

Linderman et al. **Analytical validation of whole exome and whole genome sequencing for clinical applications.** *BMC Medical Genomics* 2014, **7**:20

Heng Li. **Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples.** http://arxiv.org/pdf/1404.0929v1.pdf

AUSTRIAN INSTITUTE OF TECHNOLOGY

# Genotype variant calling

# Variant callers

Bayesian genotype model - evaluates probability of genotype given read data

**Basic model - Bayes Theorem**

P(genotype|data) $\propto$ P(data|genotype) P(genotype)

P(genotype): prior probability for variant (Genome wide SNP rate)

P(data|genotype): likelihood for observed (called) allele type

Likelihood P(data|genotype) - what s' known to affect base calling

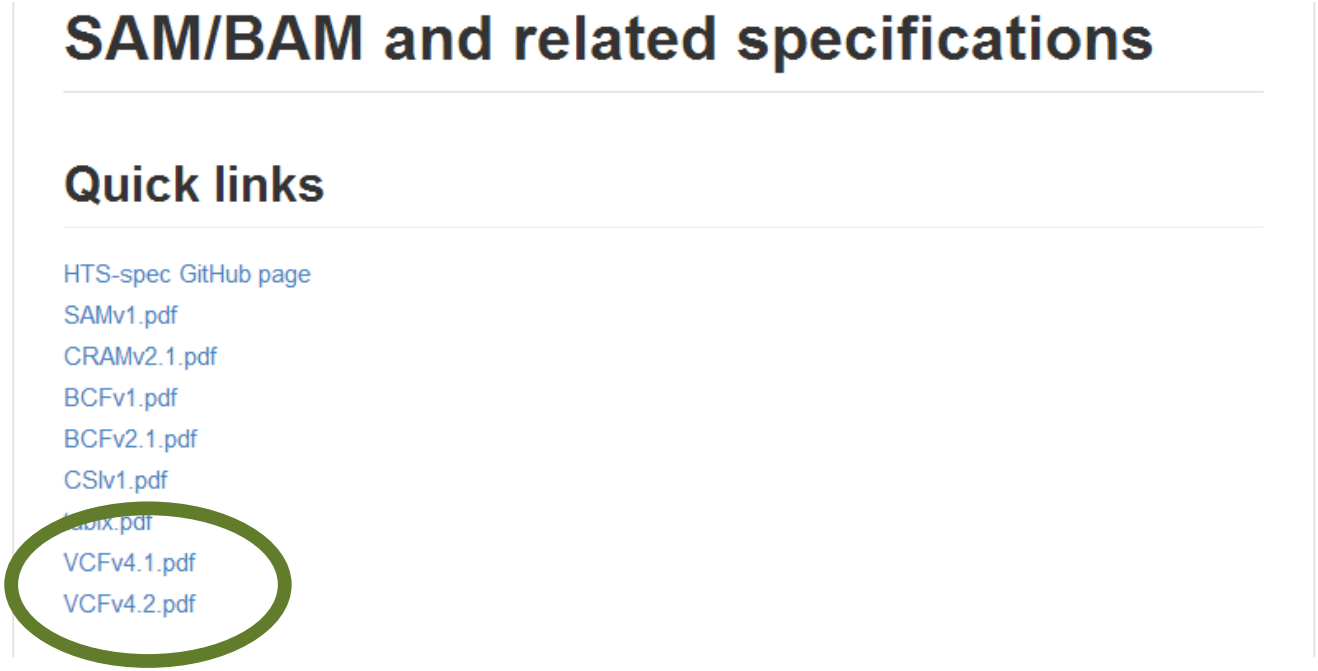
- Error rate increases as cycle numbers increase
- Error rate depends on substitution type (Ti/Tv)
- Error rate depends on local sequence environment
- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Proximity to INDEL

# Variant Call Format VCF

File format to store variant information

# Specification

https://github.com/samtools/hts-specs




**More information**

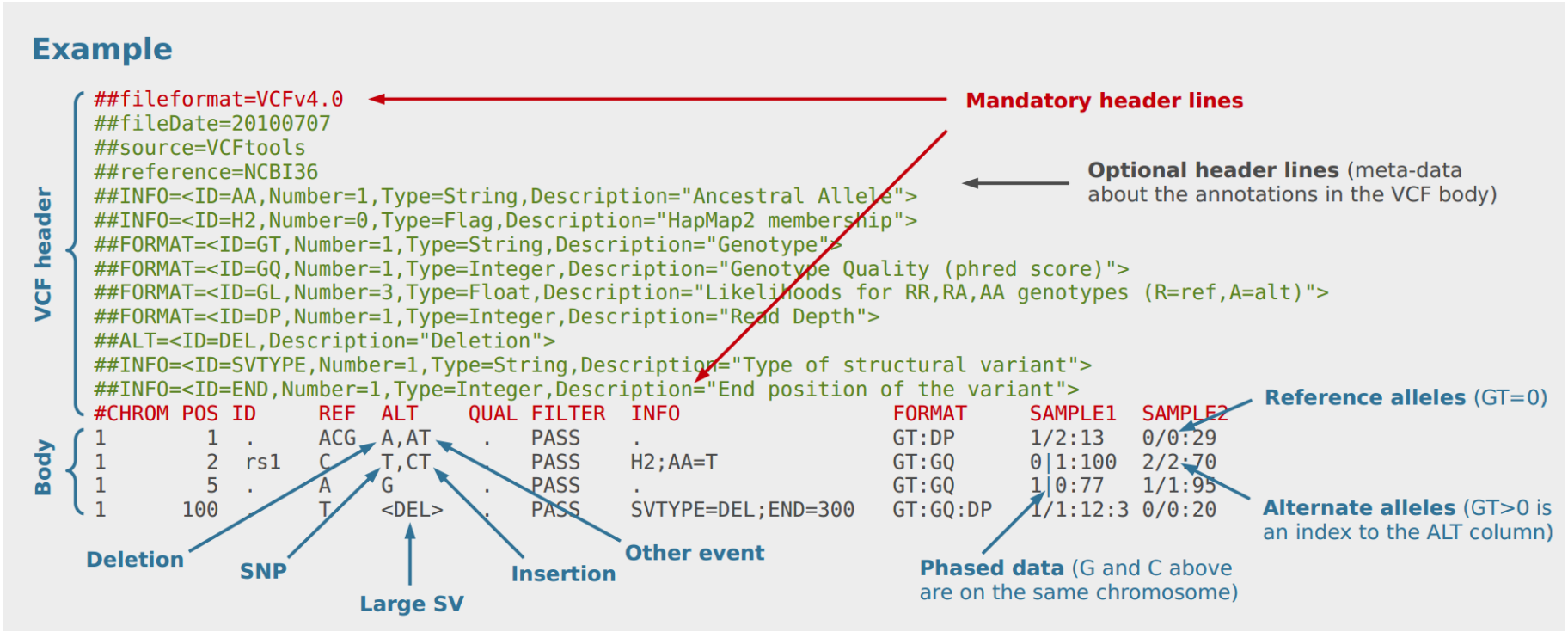
- http://vcftools.sourceforge.net/VCF-poster.pdf
- https://www.biostars.org/p/12964/

# VCF file format

AUSTRIAN INSTITUTE OF TECHNOLOGY

| | |
|---|---|
| CHROM | chromosome / contig |
| POS | the reference position with the 1st base having pos 1<br>for INDELs this is actually the base preceding the event |
| ID | id, if dbSNP variant - rs number |
| REF | reference base for INDELs, the reference string must include the base before the event |
| ALT | comma separated list of alternate non-reference alleles called on at least one of the samples |
| QUAL | phred-scaled quality score of the assertion |
| FILTER | PASS if the position has passed all filter criteria, otherwise list why filter was not passed |
| INFO | additional information |

# VCF - Example




http://vcftools.sourceforge.net/VCF-poster.pdf

# Format fields

Specifies type of data present for each genotype

- e.g.: GT:DP:GQ:MQ
- fields defined in metadata header

| | |
|---|---|
| GT | Genotype |
| DP | Read depth at position for sample |
| DS | Downsampled because of too much coverage |
| GQ | Genotype quality encoded as a phred quality |
| MQ | Mapping quality |
| QD | Variant quality score over depth |

...

# Genotype field

- GT: genotype, encoded as alleles separated by either | or /
  - 0 for the ref, 1 for the 1st allele listed in ALT, 2 for the second, etc
  - REF=A and ALT=T
- genotype 0/0 means homozygous reference A/A
- genotype 0/1 means heterozygous A/T
- genotype 1/1 means homozygous alternate T/T
  - /: genotype unphased and | genotype phased
    (Phased data are ordered along one chromosome https://www.biostars.org/p/7846/)
- ...

```
chr1    873762  .       T   G   [CLIPPED] GT:AD:DP:GQ:PL    0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047   A   G   [CLIPPED] GT:AD:DP:GQ:PL    1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431  C   T   [CLIPPED] GT:AD:DP:GQ:PL    0/1:1,3:4:25.92:103,0,26
```

http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk

# VCF – Example

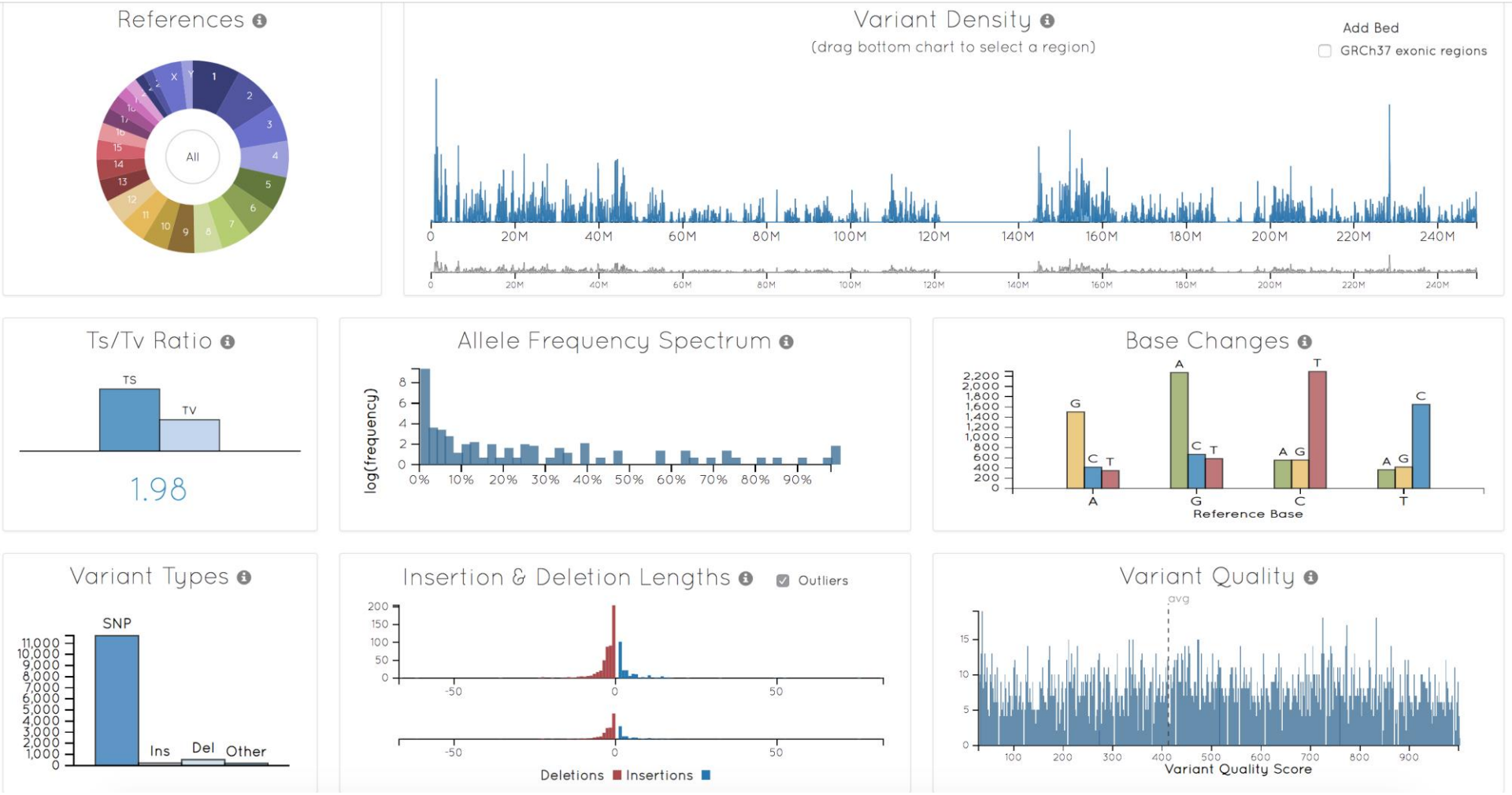
(taken from Thomas Keane)

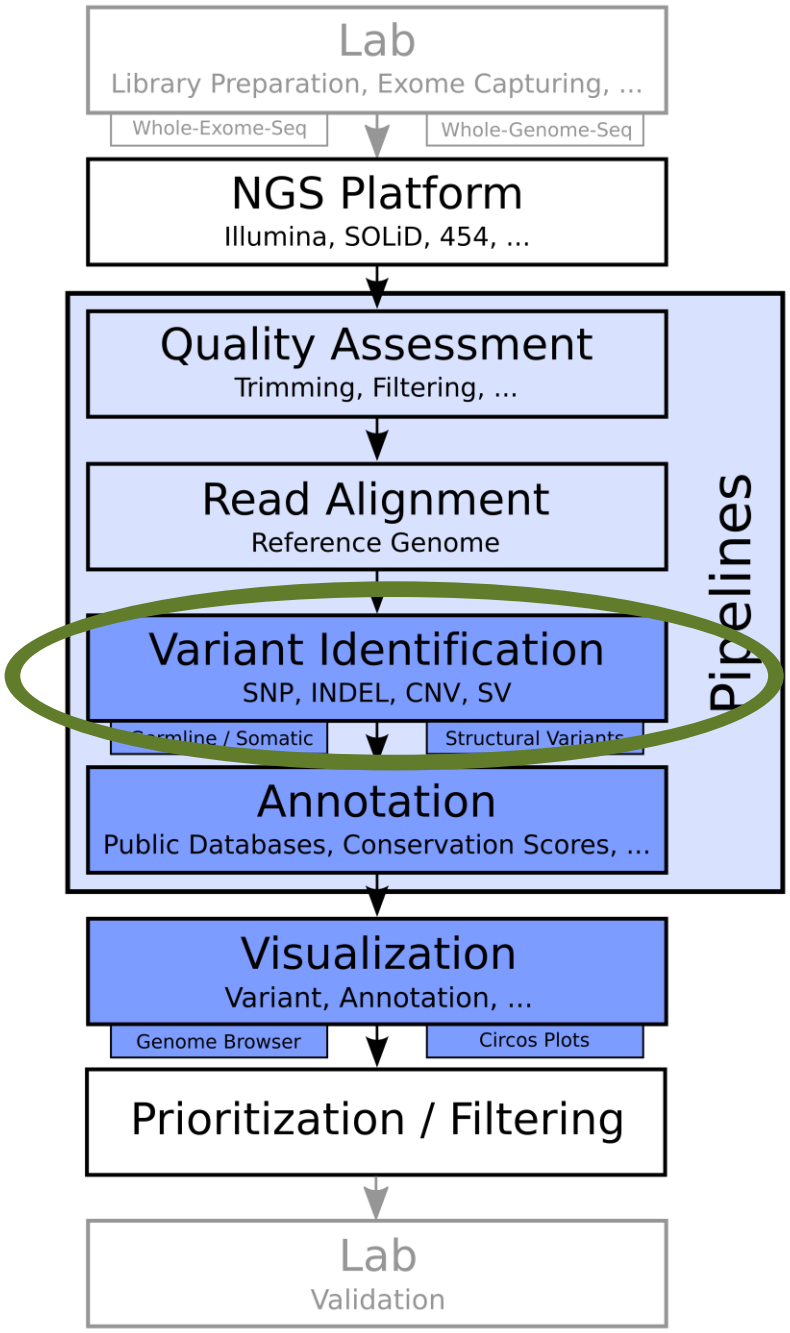
AUSTRIAN INSTITUTE OF TECHNOLOGY

```
##fileformat=VCFv4.2
##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
```

- What version of the human reference genome was used?
- What does the DB INFO tag stand for?
- What does the ALT column contain?
- At position 17330, what is the total depth? What is the depth for sample NA00002?
- At position 17330, what is the genotype of NA00002?
- Which position is a tri-allelic SNP site?
- What sort of variant is at position 1234567?

# vcf.iobio.io


References
All
Variant Density
(drag bottom chart to select a region)
Add Bed
GRCh37 exonic regions
Ts/Tv Ratio
TS
TV
1.98
Allele Frequency Spectrum
log(frequency)
Base Changes
Reference Base
Variant Types
SNP
Ins
Del
Other
Insertion & Deletion Lengths
Outliers
Deletions
Insertions
Variant Quality
avg
Variant Quality Score




Variant callers

# GATK - Genome Analysis Toolkit

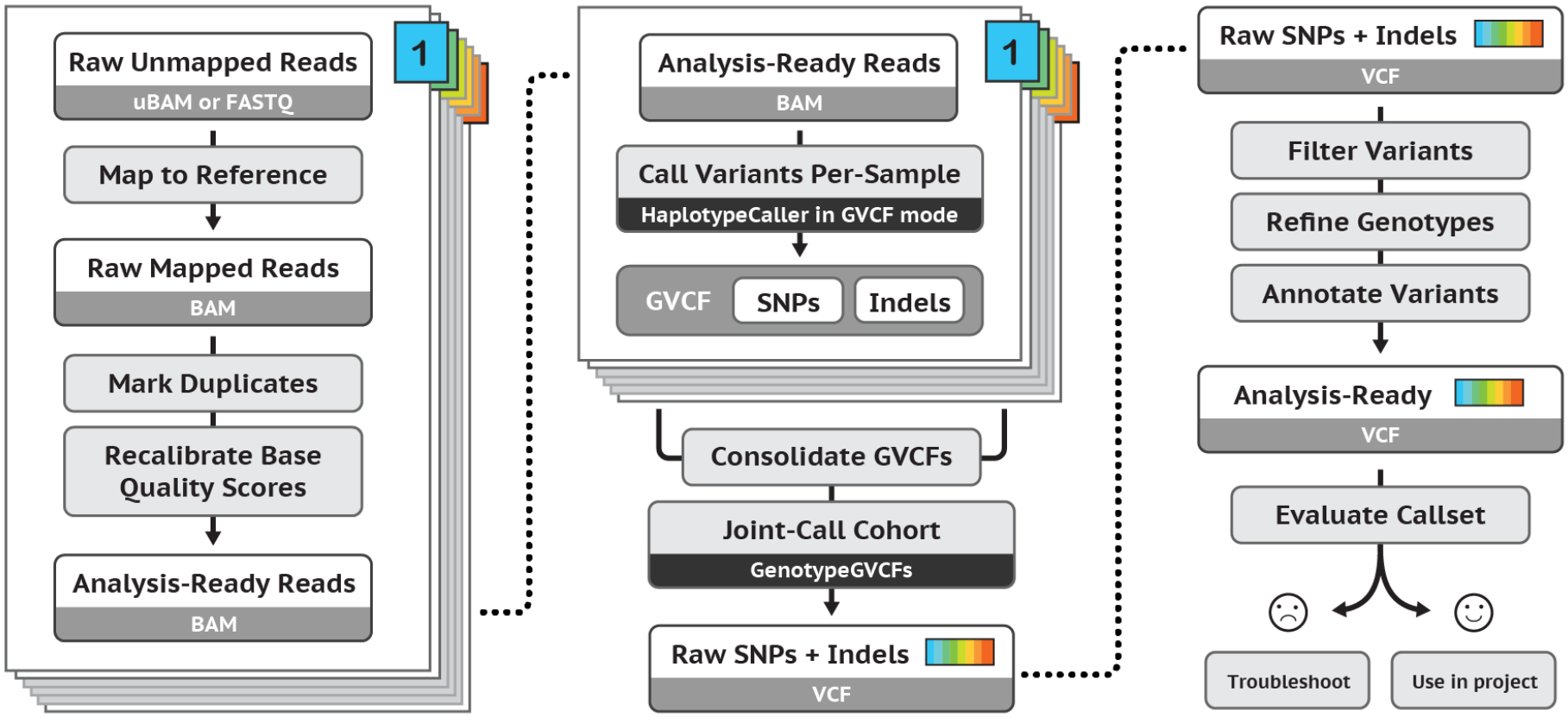
Variant calling pipeline

# GATK

- JAVA, command line software
- Linux (Mac)
- GATK4 is open-source under a BSD 3-clause

| Permissions | Limitations | Conditions |
| --- | --- | --- |
| ✔ Commercial use | ✖ Liability | ℹ License and copyright notice |
| ✔ Modification | ✖ Warranty | |
| ✔ Distribution | | |
| ✔ Private use | | |

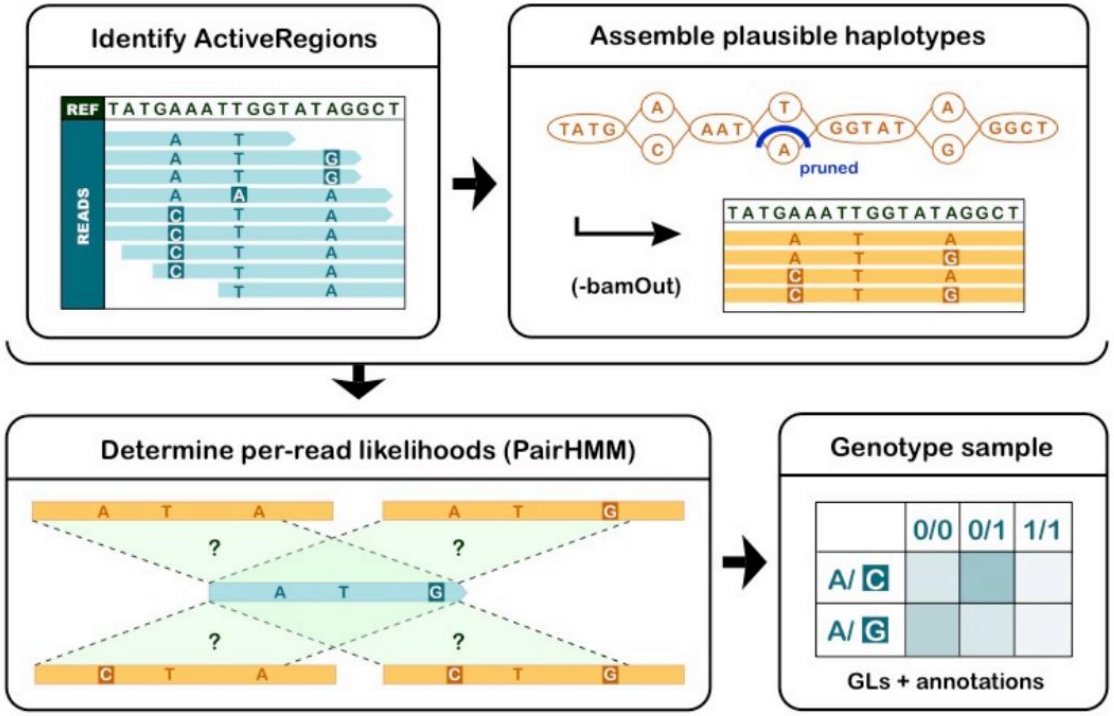
# GATK


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1
Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project

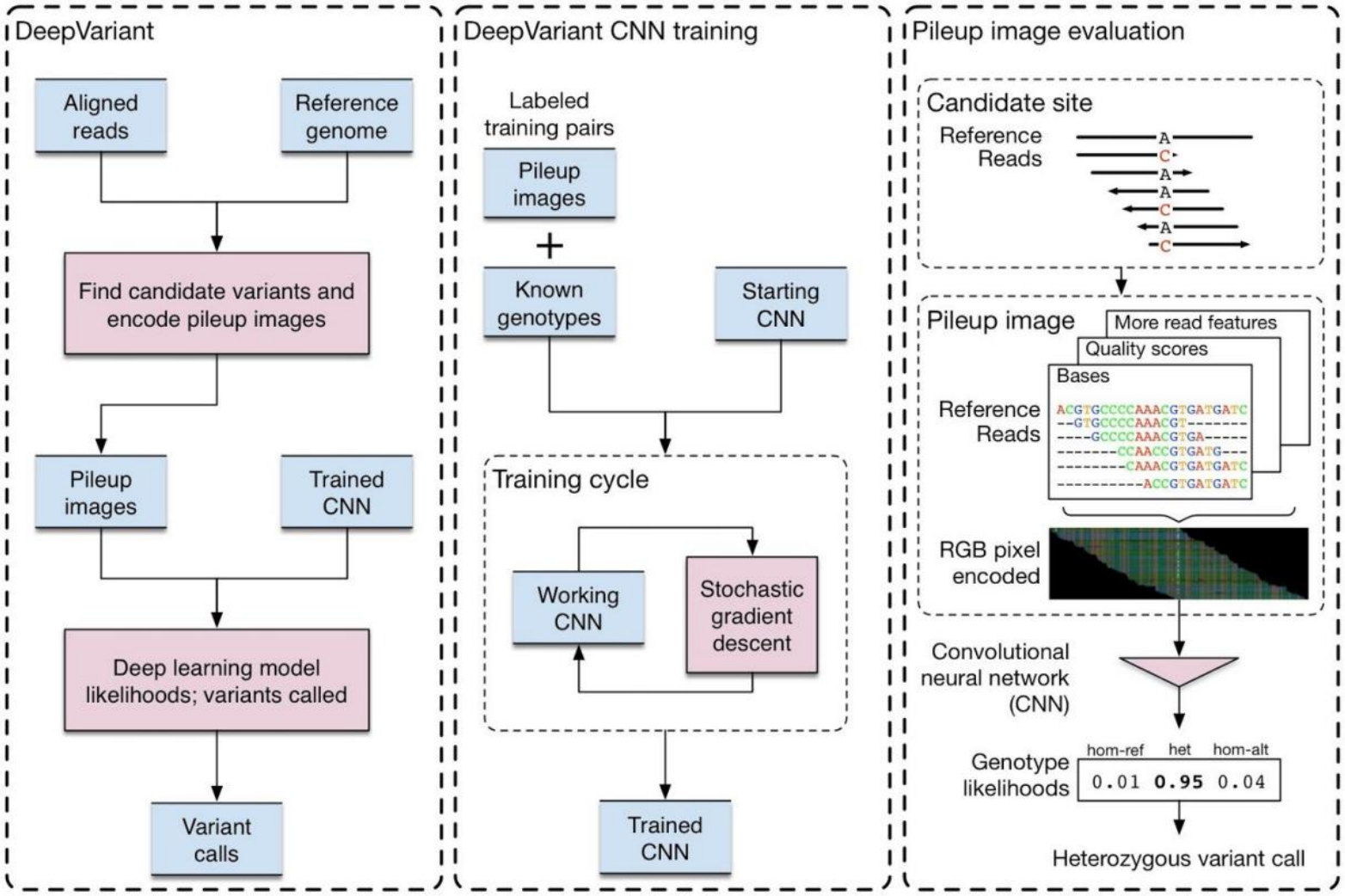
# Haplotype Caller

Calls SNVs and INDELs

- **Identify**: sliding Window, count mismatches, INDELs
- **Assemble**: local re-assembly; collect most likely haplotypes; align with SW
- **Score**: use HMM model to score haplotypes
- **Genotype**: use Bayesian model to determine most likely haplotypes



# Deep Variant




**A universal SNP and small-indel variant caller using deep neural networks**
Poplin et al. (2018) https://www.nature.com/articles/nbt.4235

# Popular germline variant callers

**Freebayes**

- SNVs and INDELs
- Determines the most-likely combination of genotypes for the population at each position in the reference

**VarDict**

- Ultra sensitive variant caller
- Amplicon bias aware variant calling (targeted sequencing)

**VarScan**

- Robust heuristic/statistic approach to call variants
- Thresholds for read depth, base quality, variant allele frequency, and statistical significance.

# Variant callers

https://github.com/deaconjs/ThousandVariantCallersRepo

README.md

## ThousandVariantCallersRepo

The Thousand Variant Callers Project is a comprehensive survey of software available (~250 pack
mutations in human DNA. Included are links to published studies and software repositories, as w
each on benchmarking, algorithmic details, application notes, and so on.

See the wiki page or browse the markdown files for the surveys.

Tictac parallelizes variant calling, and contains scripts for a variety of callers from this survey.

This is a free, community-driven resource so if you see something incorrect or incomplete or hav
a hand! To add content please clone the repo, edit the wiki markdown files, and make a pull requ

SNP Variant Callers

| caller | pubyear | from | study | source | algorithm |
|---|---|---|---|---|---|
| graphtyper | 2017 | deCODE genetics | study | source | Population-scale genotyping using pangenome graphs |
| muse | 2016 | MD Anderson Cancer Center | study | source | F81 Markov Substitution Model |
| sinvict | 2016 | Simon Frasier University, Canada | study | source | |
| multigems | 2016 | University of California, Riverside | study | source | Multinomial Bayesian, base and alignment quality priors |
| somaticseq | 2015 | Roche Bina | study | source | meta-caller, decision tree |
| discosnp | 2015 | Genscale France | study | source | reference-free, de bruijn graph |
| 2kplus2 | 2015 | Norwich Research Park, UK, Sainsbury lab | study | source | reference-free, de bruijn graph |
| exscalibur | 2015 | University of Chicago | study | source | |
| multisnv | 2015 | Cambridge Tavare | study | source | joint paired, timepoint pooling |
| rarevator | 2015 | University of Florence | study | source | Fisher's exact test, conserved loci only |
| snv-ppilp | 2015 | University of Helsinki, Finland | study | source | perfect phylogeny/integer linear programming |
| platypus | 2014 | U Oxford | study | source | Haplotype, bayesian, multi-sample, local realignment |
| baysic | 2014 | Baylor/Genformatic LLC | study | source | Meta-caller, Bayesian, unsupervised |
| haprnuc | 2014 | Kyoto University, Japan | study | source | Haplotype, Bayesian HMM |
| snpest | 2014 | U Copenhagen | study | source | reference-free, generative probabilistic |
| variantmaster | 2014 | Geneva Medical School, Switzerland | study | source | reference-free, pedigree inference |
| mutect | 2013 | Broad Getz | study | source | Beta-binomial, Variable Allele Fraction, filter population SNPs |
| niks | 2013 | Max Planck Institute for Plant Breeding Research, Germany | study | source | |
| ebcall | 2013 | Vanderbilt Zhao | study | source | Heuristic, multiple feature |
| shearwater | 2013 | U Cambridge/Welcome Trust | study | source | Beta-binomial, DeepSNV with aggregate control counts |
| shimmer | 2013 | NHGRI Larsen | study | source | Fisher's exact test, variant read count > N |
| bubbleparse | 2013 | Norwich Research Park Sainsbury Lab, UK | study | source | Reference-free, de Bruijn graph |
| cake | 2013 | Welcome Trust Adams | study | source | Meta-caller, simple 2x consensus, post-filter |
| denovogear | 2013 | WashU St Louis Conrad | study | source | Beta-binomial, pedigree |
| qsnp | 2013 | U Queensland | study | source | Heuristic, min 3 reads, post-filter |
| rvd | 2013 | Stanford University School of Medicine | study | source | Beta-binomial |
| seurat | 2013 | Translational Genomics Research Institute | study | source | Joint-paired, beta-binomial |
| snptools | 2013 | Baylor College of Medicine | study | source | Haplotype, Bayesian HMM |
| vcmm | 2013 | RIKEN Japan | study | source | Multinomial Bayesian, priors corrected Illumina q-score |
| vip | 2013 | Case Western, Li lab | study | source | Overlapping Pools |
| virmid | 2013 | UCSD Bafna | study | source | Joint-paired, Beta-binomial, purity estimation |
| varscan2 | 2012 | WashU St Louis Wilson | study | source | Heuristic, min 3 reads, filter |
| jointsnvmix | 2012 | U British Columbia Vancouver | study | source | Joint-paired, Beta-binomial |

# To consider ...

- Correctly formatted reference genome

**Important note about human genome reference versions**

If you are using human data, your reads must be aligned to one of the official b3x (e.g. b36, b37) or hg1x (e.g. hg18, hg19) references. The contig ordering in the reference you used must exactly match that of one of the official references canonical orderings. These are defined by historical karotyping of largest to smallest chromosomes, followed by the X, Y, and MT for the b3x references; the order is thus 1, 2, 3, ..., 10, 11, 12, ... 20, 21, 22, X, Y, MT. The hg1x references differ in that the chromosome names are prefixed with "chr" and chrM appears first instead of last. The GATK will detect misordered contigs (for example, lexicographically sorted) and throw an error. This draconian approach, though unnecessary technically, ensures that all supplementary data provided with the GATK works correctly. You can use ReorderSam to fix a BAM file aligned to a missorted reference sequence.

http://www.broadinstitute.org/gatk/guide/article?id=1213

- BAM file
  - sorted
  - indexed
  - with RG

# Ensemble variant calling

**Method**

- Combine multiple VCF caller outputs into one callset
- Specify how many callers need to identify a variant (heuristic step)
- Use included and excluded variants to train a support vector machine
  → use this classifier to identify trusted variants

**Validation**

- Used a pair of replicates
- Compared to variants from a single calling method, the ensemble method produced **more concordant variants** when comparing the replicates, with **fewer discordants**

https://github.com/chapmanb/bcbio.variation.recall

# Other types of variant calling

**Structural variant calling**

- Identify large deletions, insertions, translocations, inversions

**Copy Number Variation (CNV) calling**

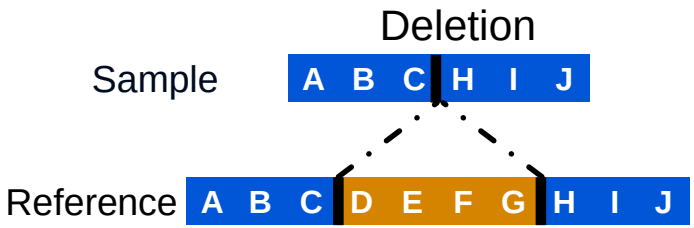
- Which parts of the genome are amplified or deleted?

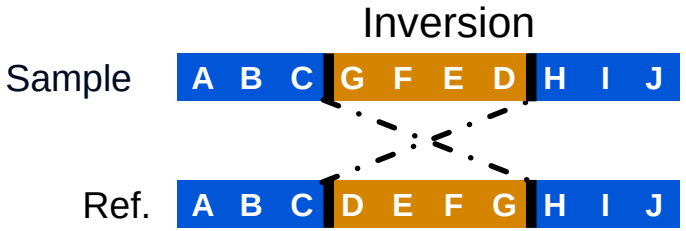
**Somatic variant calling**

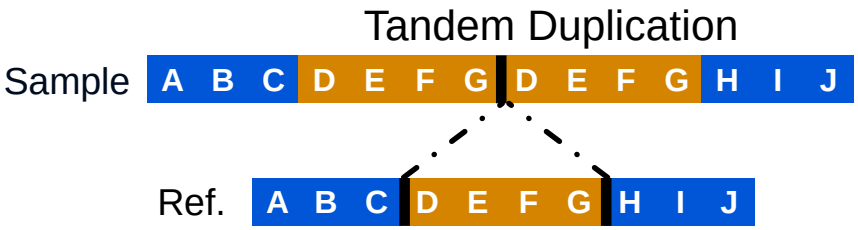
- Find acquired mutations

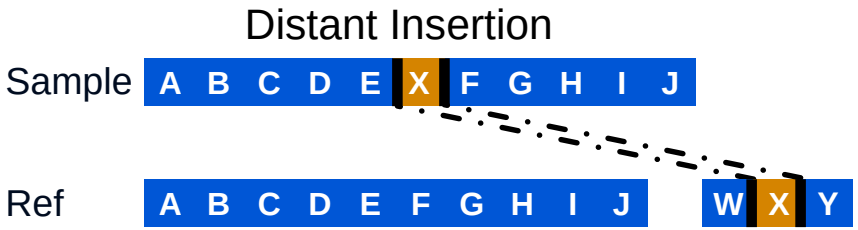
# Structural variation calling

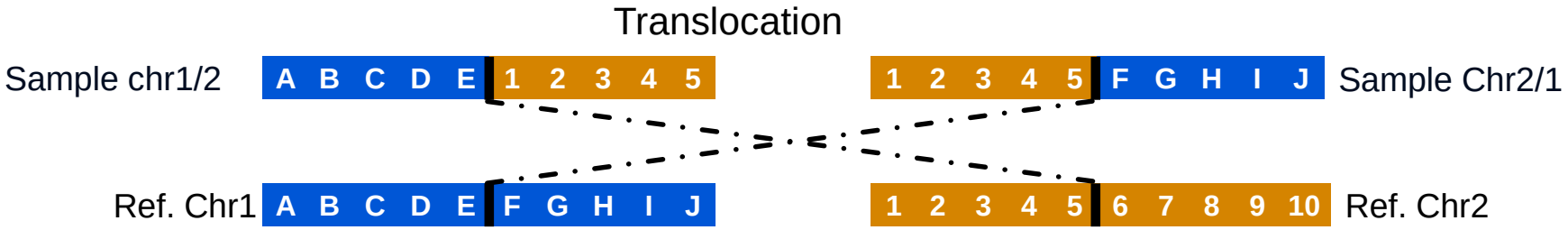
# Structural variations



# Why is structural variation relevant / important?

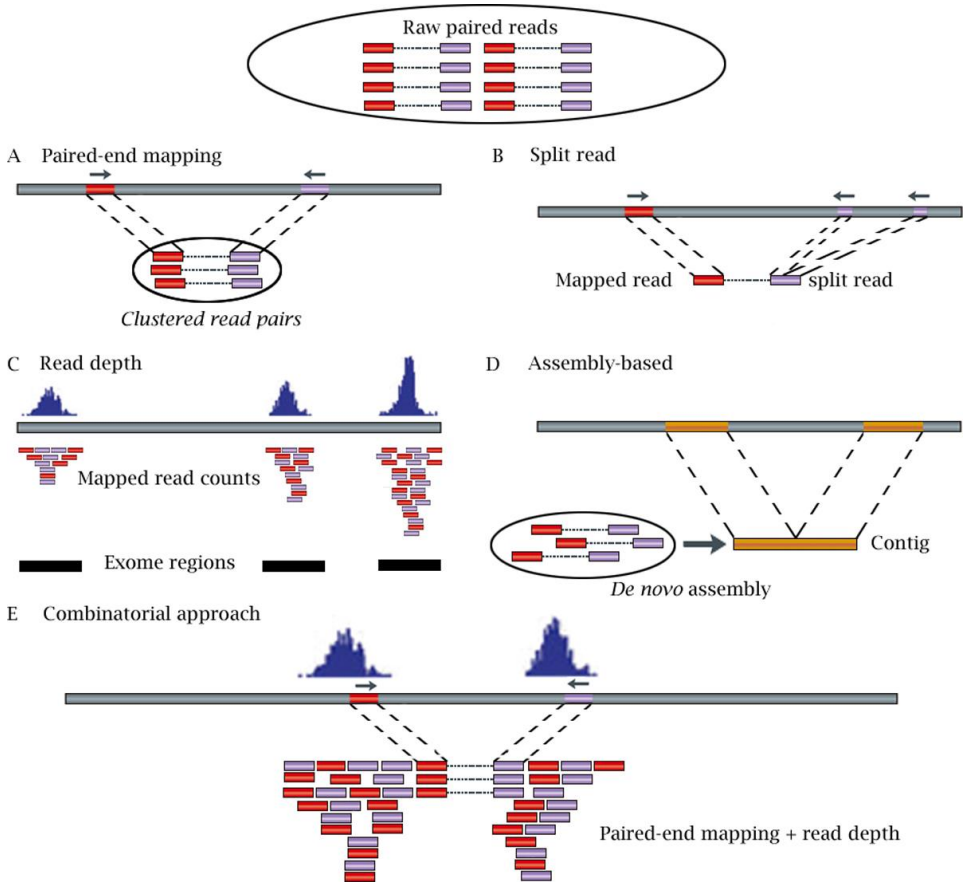
They are common and **affect a large fraction of the genome**

- In total, SVs impact more base pairs than all single nucleotide differences

They are a major **driver of genome evolution**

- Speciation can be driven by rapid changes in genome architecture
- Genome instability and aneuploidy: hallmarks of solid tumor genomes

# SV/CNV detection




**A**. Paired-end mapping (PEM) strategy detects SVs/CNVs through **discordantly mapped reads**. A discordant mapping is produced if the distance between two ends of a read pair is **significantly different from the average insert size**.

**B.** Split read (SR)-based methods use **incompletely mapped read** from each read pair to identify small SVs/CNVs.

**C.** Read depth (RD) approach detects by **counting the number of reads mapped** to each genomic region. In the figure, reads are mapped to three exome regions.

**D.** Assembly (AS)-based approach detects CNVs by **mapping contigs** to the reference genome.

**E.** Combinatorial approach combines **RD and PEM** information to detect CNVs.

*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3846878/*

Zhao *et al. BMC Bioinformatics* 2013 **14**

# Structural variations - tools

**Breakdancer**

- Insertions, deletions, inversions, translocations
- Fast, simple to run

**Pindel**

- Insertions, deletions

**GASVPro**

- Combines read depth info along with discordant paired-read mappings
- Duplications, deletions, insertions, inversions and translocations

**SVMerge**

- Results from several different SV caller (Breakdancer, Pindel, SE Cluster, RDXplorer, RetroSeq)

**Mavis**

- post-processing of structural variant calls.
- http://mavis.bcgsc.ca/

# Structural variations - tools

**LUMPY**

- Integrates different sequence alignment signals (read-pair, split-read and read-depth)
- https://github.com/arq5x/lumpy-sv

**Manta**

- Calling structural variants, medium-sized indels and large insertions
- Very fast

**Delly**

- Integrates short insert paired-ends, long-range mate-pairs and split-read alignments
- Detects CNVs, deletion, tandem duplication events, inversions or reciprocal translocations

**tardis**

- Rapid discovery of structural variants
- Available as Docker image

# Structural variations – combination/evaluation

**SURVIVOR**

- Simulates SVs given a reference, number and size ranges for each SV
  insertions, deletions, duplications, inversions and translocations
  - bed file to report the locations of the simulated SVs

- Evaluates SV
  - VCF input
  - start & stop coordinates of the sim and ident SV within 1 kb (parameter)

- Filter and combine the calls from VCF files

https://www.nature.com/articles/ncomms14061

# Meta SV-caller

**Parliament2**

- For WGS data
- Runs a combination of tools:
  - Breakdancer
  - Breakseq2
  - CNVnator
  - Delly2
  - Manta
  - Lumpy
- Merges calls with SURVIVOR

https://github.com/dnanexus/parliament2

# Structural variations - challenges

**Often many false positives**

- Short reads + heuristic alignment + rep. genome **= systematic alignment artifacts (false calls)**
- Ref. genome errors (e.g., gaps, misassemblies)
- **ALL** SV mapping studies use strict filters for above

**The false negative rate is also typically high**

- Most current datasets have low to moderate ***physical*** coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment

# Long Read Technologies

(+) SVs in repetitive regions
(+) Can identify nested SVs

(-) Higher error rate
(-) Hard to align

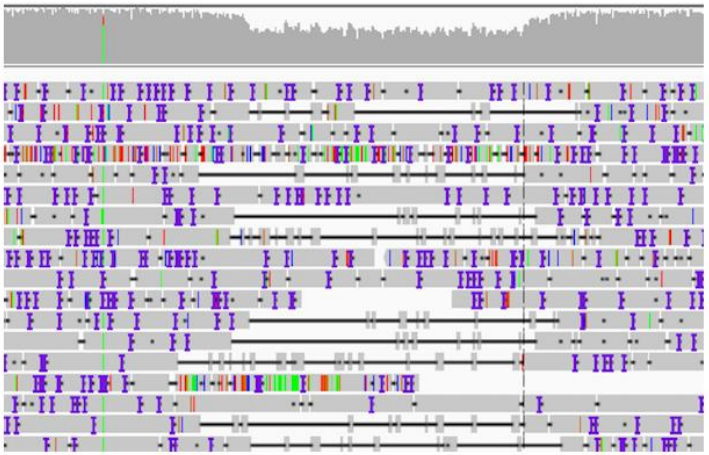

PACBIO®

# Improving long read alignment

- NGM – http://cibiv.github.io/NextGenMap/

1. Split the reads:
   - Translocations
   - Inversions
   - Duplications

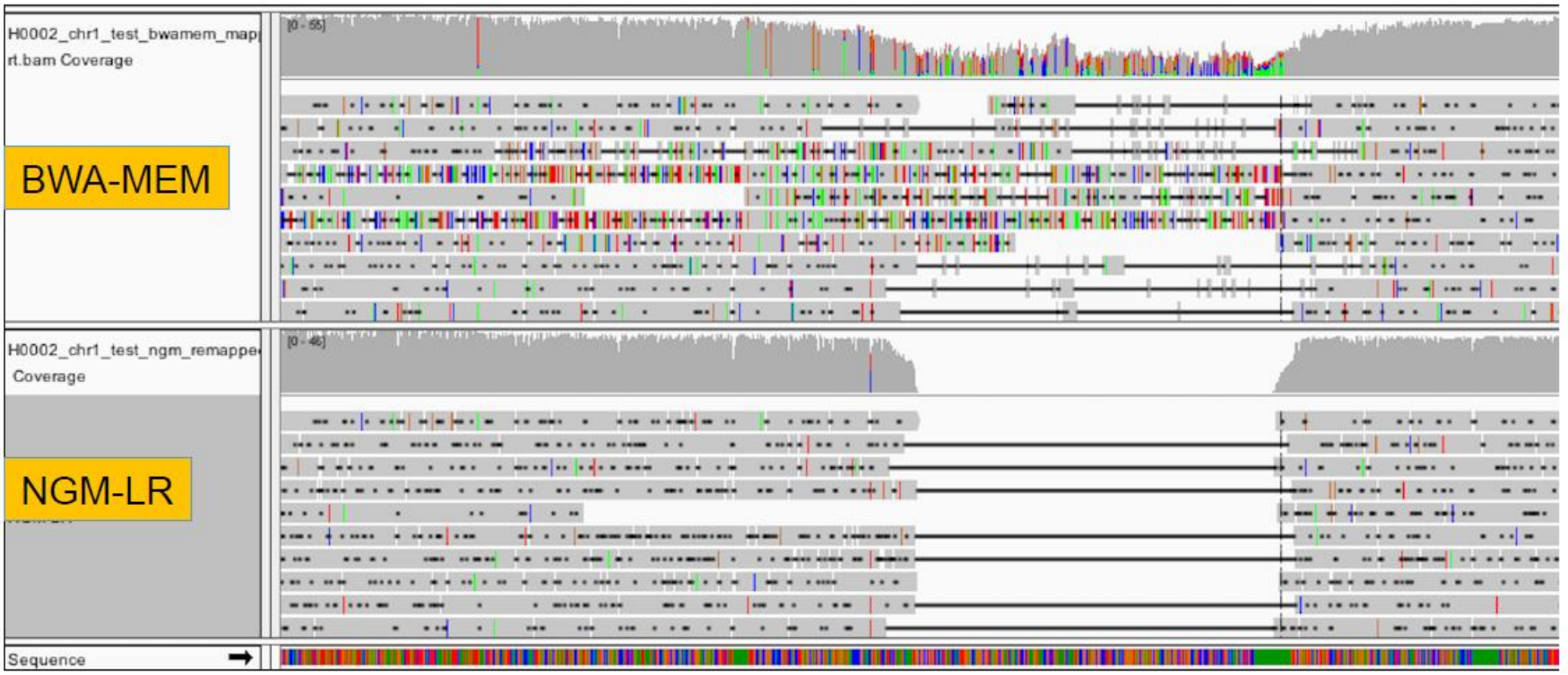
2. Improve alignment:
   - Insertions
   - Deletions



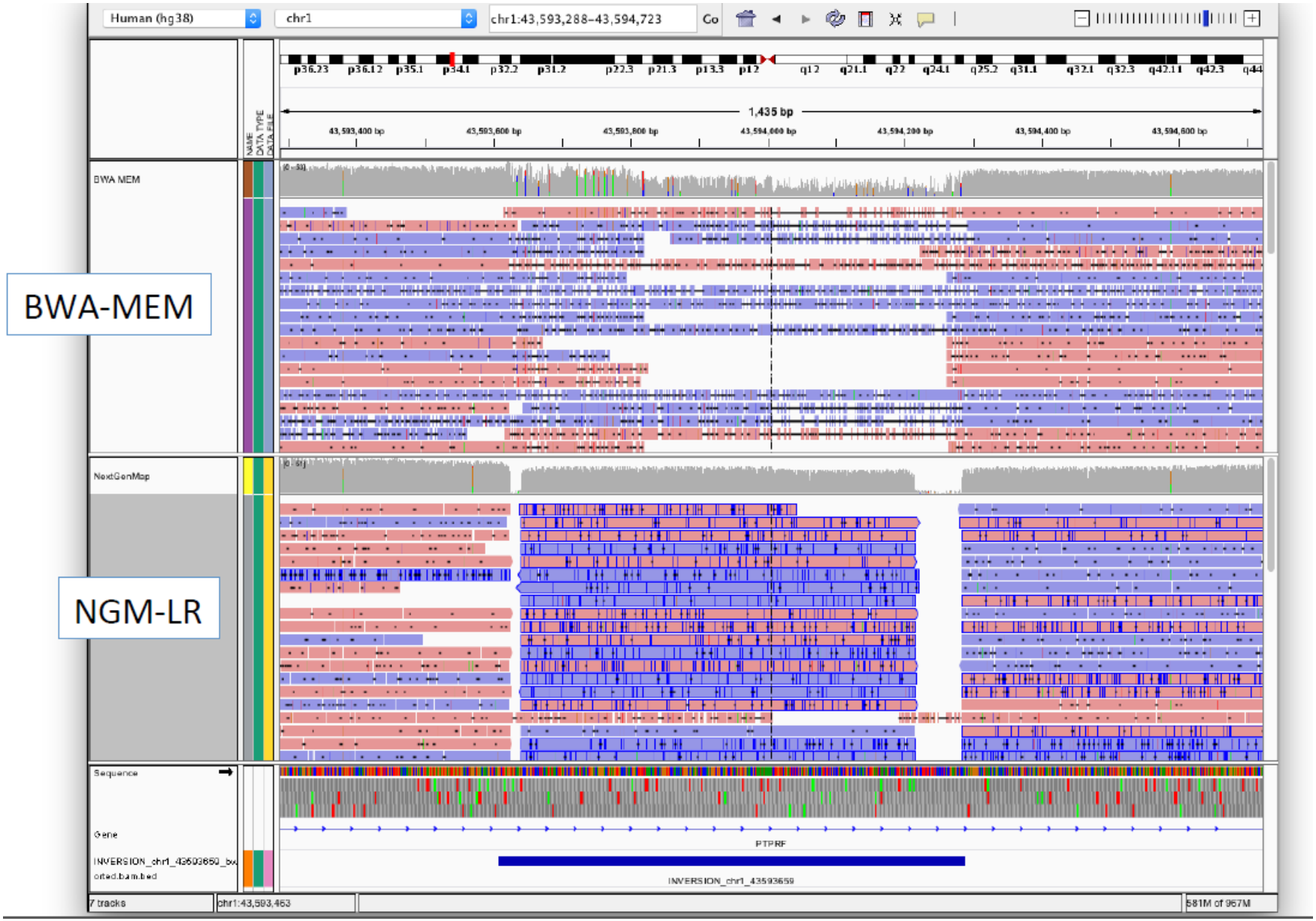
# Structural variations with 3rd gen sequencing

NGM-LR + Sniffles: PacBio SV Analysis Tools

- ***1. NGM-LR***: Improve mapping of noisy long reads: improved seeding, convex gap scoring
- ***2. Sniffles***: Integrates evidence from split-reads, alignment fidelity, breakpoint concordance



# NGM-LR complex SV



# A robust benchmark for germline structural variant detection - 2019

- First benchmark set for identification of both false negative and false positive germline SVs
- 12,745 isolated, sequence-resolved insertion and deletion calls
- In general, the concordance for insertions is lower than the concordance for deletions, except among long-read callsets
- Excludes complex SVs

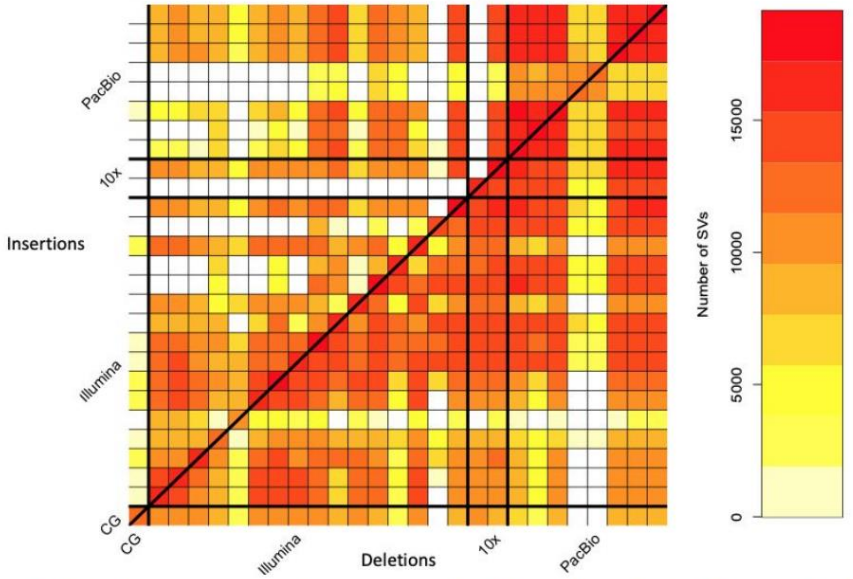


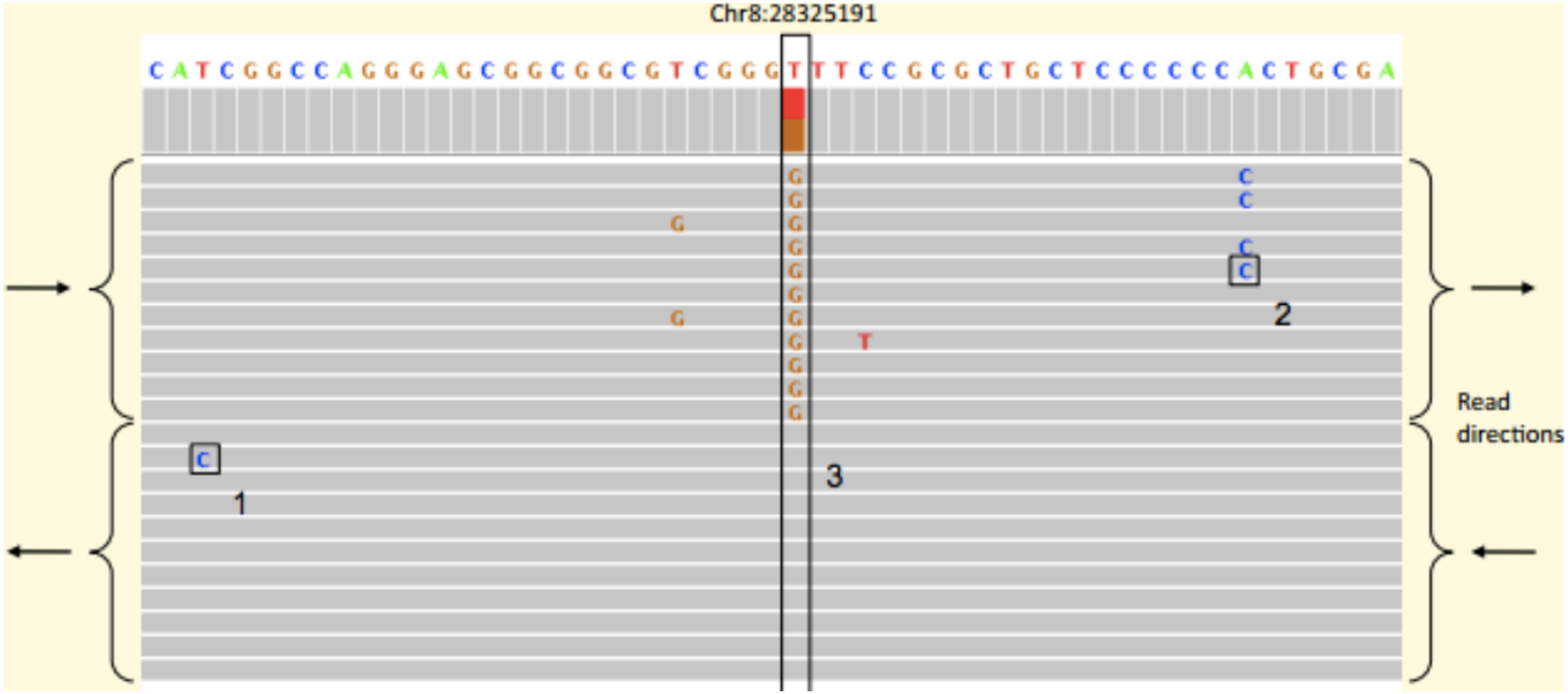

**Figure 1: Pairwise comparison of sequence-resolved SV callsets obtained from multiple technologies and SV callers for SVs ≥50bp from HG002**. Heatmap produced by SURVIVOR[40] shows the number of SVs overlapping between the individual SV caller and technologies split between insertions (upper left) and deletions (lower right). The diagonal highlights the overall number of SVs per SV caller. Overall we obtained a quite diverse picture of SVs calls supported by each SV caller and technology, highlighting the need for benchmark sets.

https://www.biorxiv.org/content/biorxiv/early/2019/06/11/664623.full.pdf

# What information is needed to decide if a variant exists?

- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

# Beware of Systematic Errors




- **Identification and correction of systematic error in high-throughput sequence data** Meacham et al. (2011) *BMC Bioinformatics.* 12:451
- **A closer look at RNA editing.** Lior Pachter (2012) *Nature Biotechnology.* 30:246-247

# Variant filtering

**The biggest problem is large numbers of FPs:**

- Based on bad alignments
- Can be systematic across samples,
  thus creating consistent SNPs across samples
- Sequencing errors
  should be accounted for by base quality + recalibration + marking of duplicates

**FPs and FNs, may result in:**

- Data drowning in noise & no result
- False results & erroneous result

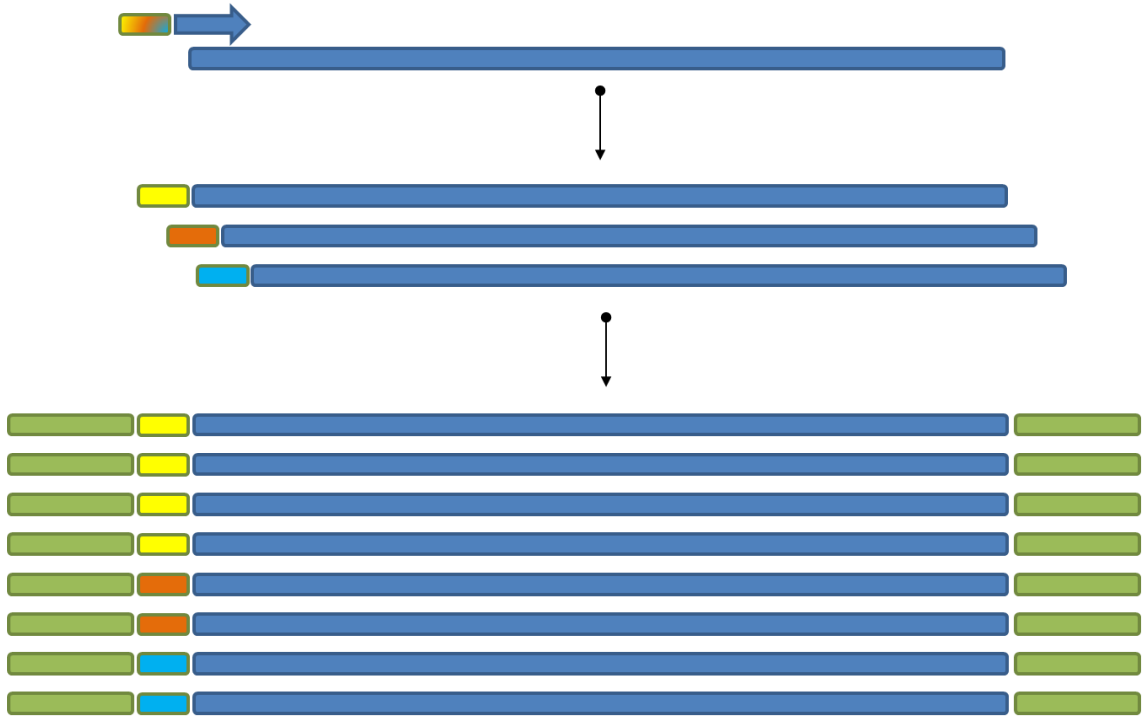
→ **Filter**

# Molecular barcoding

**Library generation**

STEP 1 - Barcoding

STEP 2 - Amplification



**Consensus sequence generation**


random barcode mix  unique barcodes  sequencing adaptors

Adapted from http://www.imgm.com

# Variant filtering – how to

QUAL (depends on MQ of reads and base qualities) is a useful measure

**But** - there will also be FP with high QUAL

**Signs of suspicious variants**

- Poorly mapped reads (ambiguity)
- MQ: Root Mean Square of MAPQ of all reads at locus
- MQ0: Number of MAPQ 0 reads at locus
  - check biased support for the REF and ALT alleles
- ReadPosRankSum: Read position rank sum test
  If alternate allele is only at ends of read → indicative for error
- Strand bias
- FS: Fisher strand test
  If reference carrying reads are balanced between strands, alternate carrying reads should be as well

More information: https://www.broadinstitute.org/gatk/guide/tagged?tag=VQSR

AUSTRIAN INSTITUTE OF TECHNOLOGY

# Tools for VCF

# vcflib

C++ library for parsing and manipulating VCF files

- Comparison of VCF files
- Filtering and subsetting
- Order VCF files
- Break multiple alleles into single files
- Prints statistics about variants

**https://github.com/ekg/vcflib**

# VCFtools

Easily accessible methods for working with complex genetic variation data

C++

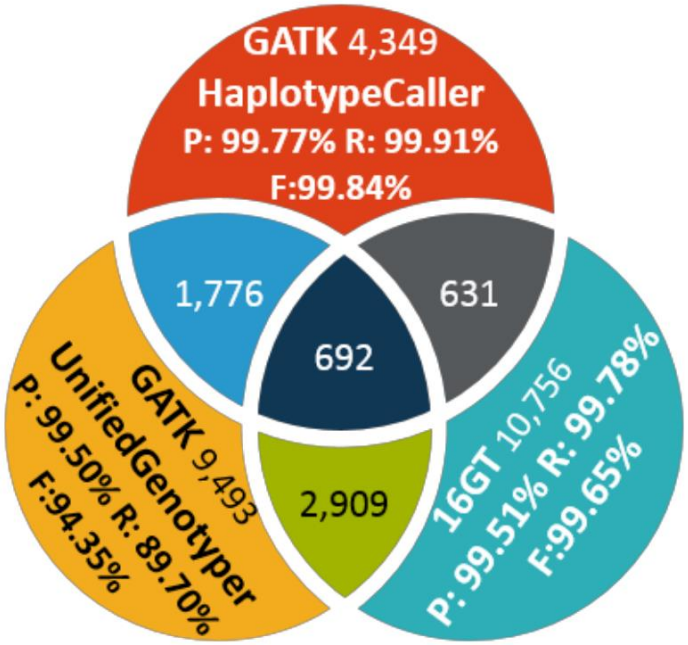
- Basic file statistics
- Filtering
- Comparing two files
- Sequencing depth information

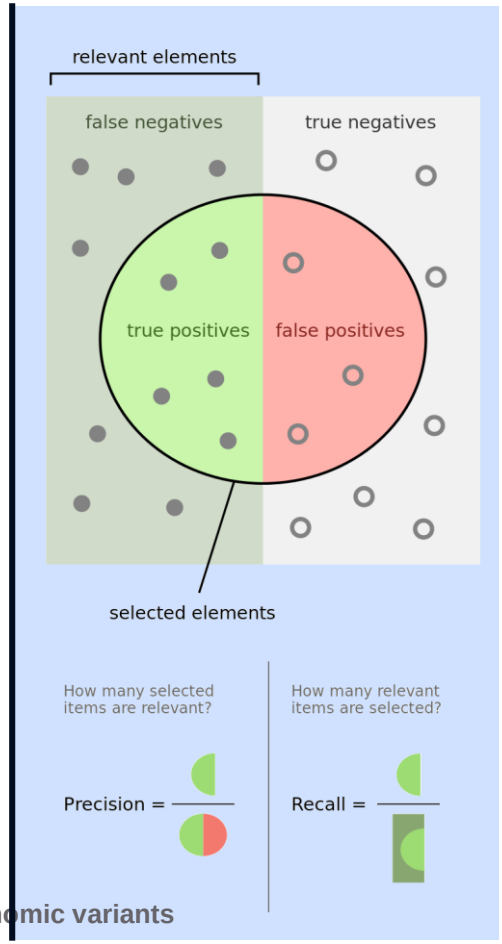
**http://vcftools.sourceforge.net/**

# Skyhawk

- Neural network-based discriminator
- Mimics expert review on clinically significant genomics variants

- Helps removing FP results




**Skyhawk: An Artificial Neural Network-based discriminator for reviewing clinically significant genomic variants**
Ruibang Luo, Tak-Wah Lam, Michael Schatz; https://www.biorxiv.org/content/10.1101/311985v2

# SPALTER

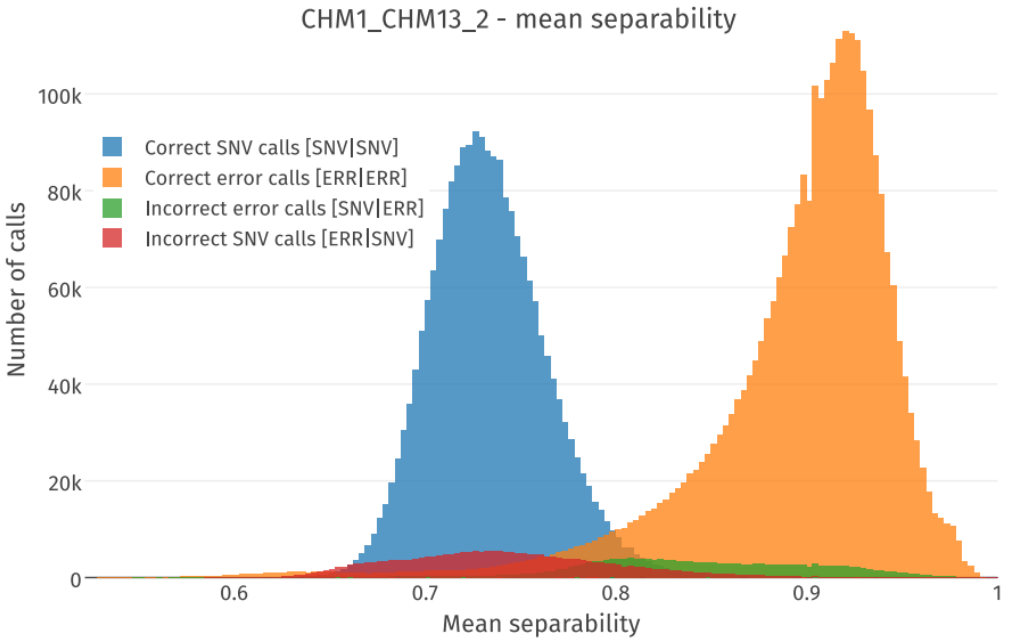
Spalter: A Meta Machine Learning Approach to Distinguish True DNA Variants from Sequencing Artefact

Allows to use different models

- Logistic regression
- SVM
- Decision trees

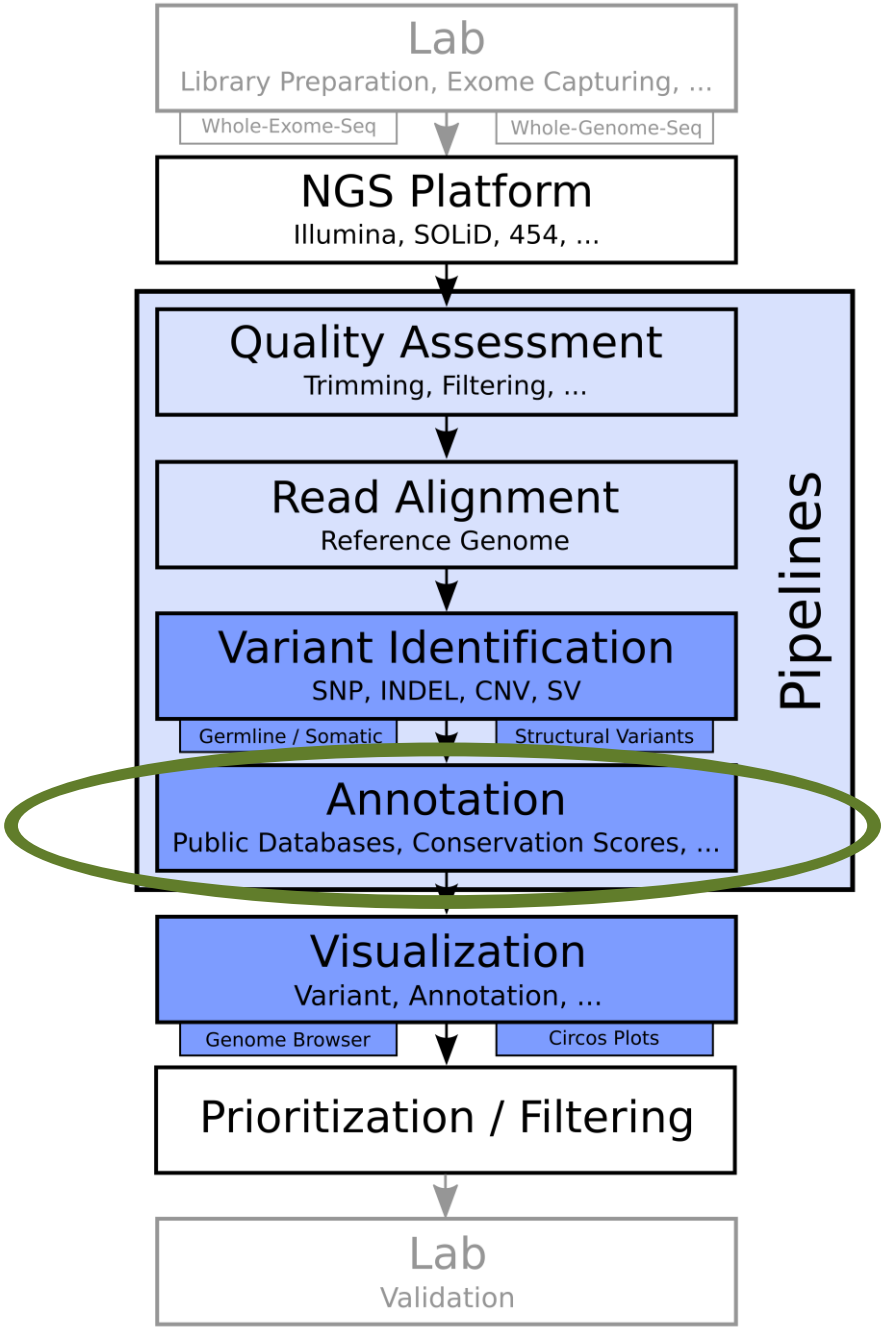
True SNVs: 85.2% detected
Correct SNVs: 91.7%

True Artefacts: 93.6% detected
Correct Artefacts: 88.4%

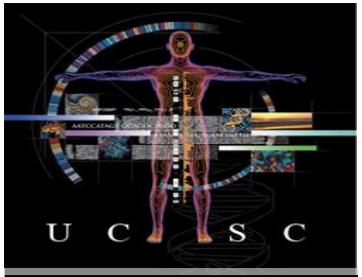



CHM: complete hydatidiform mole cell lines (i.e. completely homozygous)

# Variant annotation


Lab
Library Preparation, Exome Capturing, ...
Whole-Exome-Seq
Whole-Genome-Seq
NGS Platform
Illumina, SOLiD, 454, ...
Quality Assessment
Trimming, Filtering, ...
Read Alignment
Reference Genome
Variant Identification
SNP, INDEL, CNV, SV
Germline / Somatic
Structural Variants
Annotation
Public Databases, Conservation Scores, ...
Pipelines
Visualization
Variant, Annotation, ...
Genome Browser
Circos Plots
Prioritization / Filtering
Lab
Validation

# Annotation provides context for interpretation




Conservation
Repeat elements
Genome Gaps
Cytobands
Gene annotations
“Mappability”
DeCIPHER
ISGA

dbSNP
Short Genetic Variations

gnomAD browser

ClinVar

OMIM
Online Mendelian Inheritance in Man

Genetic variation

..CCTCATGCATGGAAA..
...CCTCATGTATGGAAA..
...CCTCATGCATGGAAA..
...CCTCATGCATGGAAA..
...CCTCATGTATGGAAA..
...CCTCATGCATGGAAA..
..CCTCATGTATGGAAA..
.

1000 Genomes
A Deep Catalog of Human Genetic Variation

ENCODE

Chromatin marks
DNA methylation
RNA expression
TF binding

Pfam

KEGG
Kyoto Encyclopedia of Genes and Genomes

HGMD®

Human Protein Reference Database

# Variant annotation

After variant calling → **many** variants

- Synonymous vs. nonsynonymous
- Frameshift mutation?
- Impact of variant?

**Annotation**

- Basis for filtering and prioritizing potential disease-causing mutations
- Most tools focus on the annotation of SNPs
- Many provide database links to various public variant databases (dbSNP...)
- Functional prediction of the variants
  - Sequence-based analysis
  - Region-based analysis
  - Structural impact on proteins

# Variant annotation

Two broad categories of annotations

**Annotations depending on gene models**

- Coding/non-coding
- If coding: synonymous / non-synonymous
- If non-synonymous → what is the impact on protein structure (Polyphen, SIFT, etc)

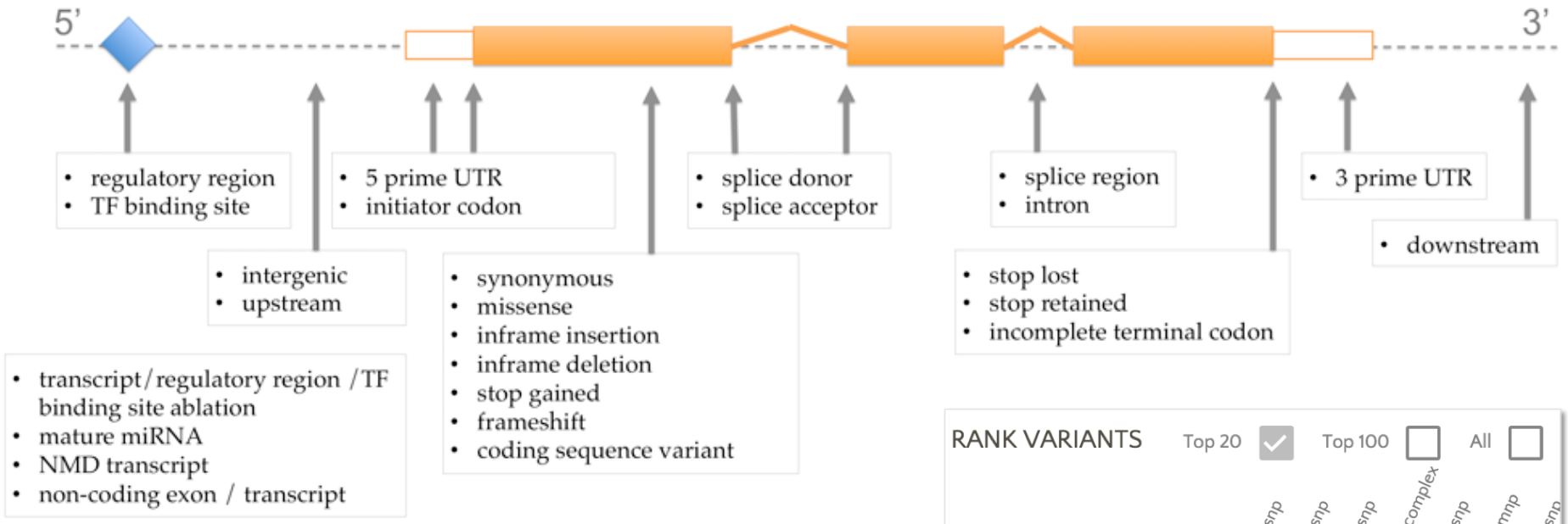
**Annotations that do not depend on gene models**

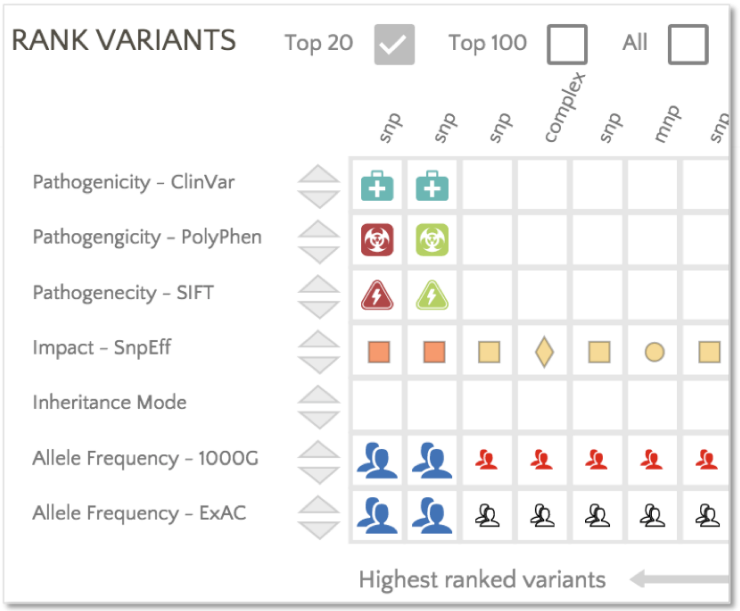
- Variant frequency in different database / different populations
- Degree of conservation across species

# Interpretation of Variants

DNA Sequencing -> Identified variants -> **Interpretation**?

Solution: effect prediction



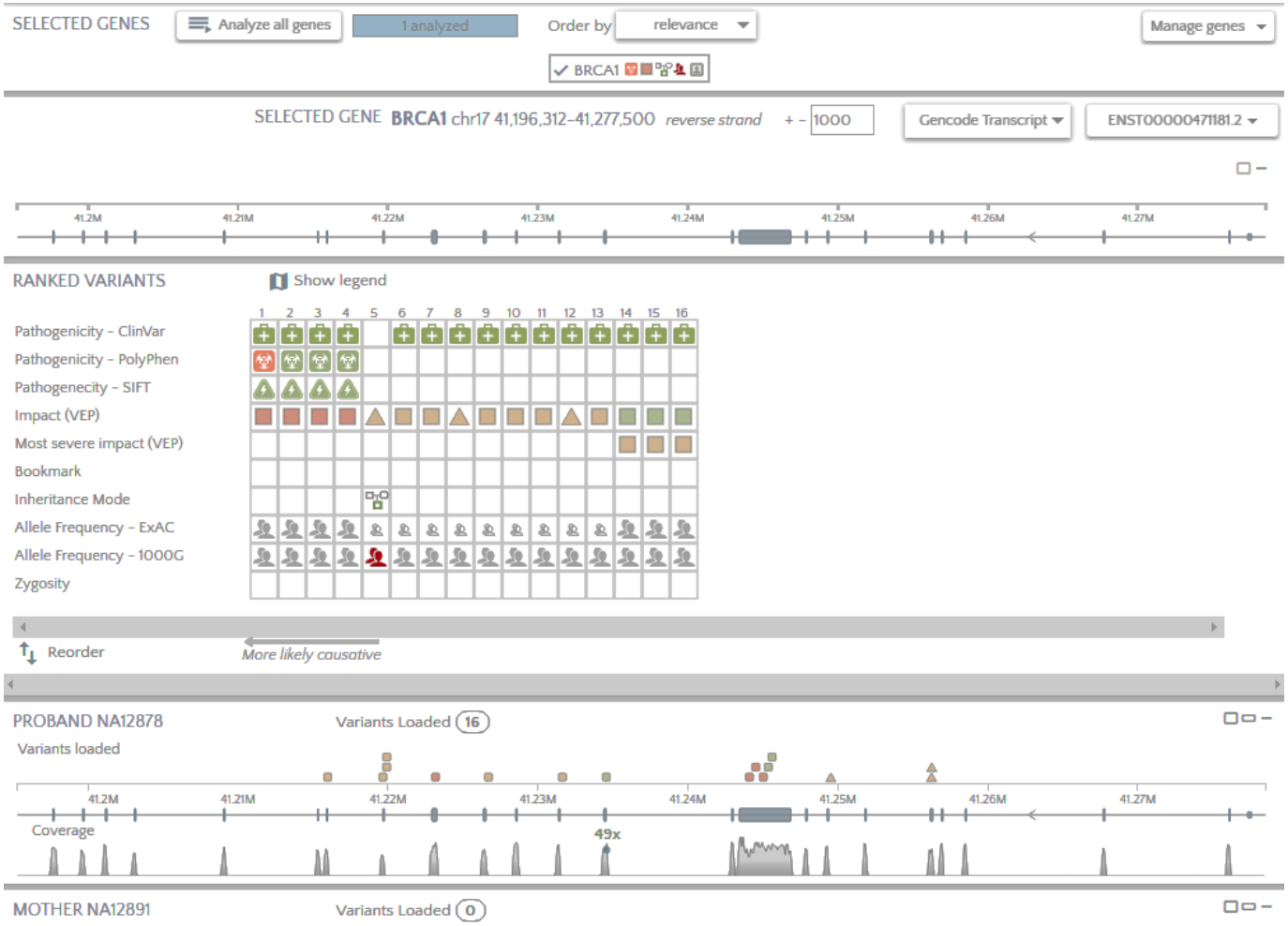

http://www.ensembl.org/info/genome/variation/consequences.jpg

http://iobio.io/public/images/blog/vepblog10.png

# Gene annotation

**Gene.iobio.io**

- Interactive
- Load custom data



# dbSNP

(www.ncbi.nlm.nih.gov/SNP)

- Single Nucleotide Polymorphism Database
- Central repository for SNPs and INDELs
- Information for variants: Population, Sample Size, allele frequency, genotype frequency, heterozygosity, …

**Submissions**

- ~550m submissions, ~150m variants (stats only for human) [v147, 2016]
- ~1800m submissions, ~660m variants (stats only for human) [v151, 2018]

https://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi

# dbSNP

AUSTRIAN INSTITUTE OF TECHNOLOGY

### Minor Allele Frequency (MAF)

Minor Allele Frequency is the allele frequency for the 2nd most frequently seen allele. dbSNP aggregates the minor allele frequency for each refSNP cluster over multiple submissions to help users distinguish between common polymorphisms and rare variants.

Consider a variation with the following alleles and allele frequencies:

Reference Allele = G; frequency = 0.600
Alternate Allele = C; frequency = 0.399
Alternate Allele = T; frequency = 0.001

Based on the MAF guideline mentioned above, the minor allele is "C," so the minor allele frequency (MAF) is 0.399. Allele "T" with frequency 0.001 is considered a rare allele rather than a minor allele.

http://www.ncbi.nlm.nih.gov/books/NBK174586/

# Variant annotation - tools

| Standalone | WEB |
| --- | --- |
| Installation | No installation |
| Mostly command line | Often easy to use |
| Depends on performance of local infrastructure | Depends on performance of public server |
| Local data transfer | Transfer data via WWW |
| Batch submission | Often no batch submission |
| No legal issues | Legal issues ... |
| Download of additional files often required | No download of additional files / databases |

# Variant annotation - tools

**ANNOVAR**

- Annotates SNPs, INDELs, block substitutions as well as CNVs.
- Gene-based, region-based and filter-based annotation
- Many preconfigured databases

**SeattleSeq Annotation server**

- Online tool
- Human SNPs and INDELs

**snpEff**

- Integrated within Galaxy and GATK.
- SNPs and INDELs

**MARRVEL**

- **M**odel organism **A**ggregated **R**esources for **R**are **V**ariant **E**xp**L**oration
- Web-based → provides many information (ClinVar, Domain, OMIM)

# Variant annotation - tools

**SVScore**

- *in silico* structural variation (SV) impact prediction
- use the precomputed SNP scores from CADD

**VEP**

- Variant Effect Predictor
- Ensembl
- Support plugins
- SNPs, insertions, deletions, CNVs or structural variants

**StructMAn** (https://academic.oup.com/nar/article/44/W1/W463/2499349)

- Annotation in structural context (analysis the spatial location)
- Web-based

# Variant annotation - tools

**ONCOTATOR**

- Web-application for annotating human variants --- cancer research
- Can also be downloaded and installed locally

**Exomiser**

- Find potential disease causing variants (annotation done by Jannovar)
- Uses VCF & HPO phenotypes
- http://www.sanger.ac.uk/science/tools/exomiser

**LOFTEE**

- VEP plugin to identify LoF (loss-of-function) variation
- Stop-gained, splice site disruption, frameshift

**Vcfanno**

- New tool for parallel annotation (8,000 variants per second)
- https://github.com/brentp/vcfanno

# ExAC

- If a disease phenotype is rare, the causal variant should also be similarly rare
- ExAC reports the allele frequency from diverse ancestries

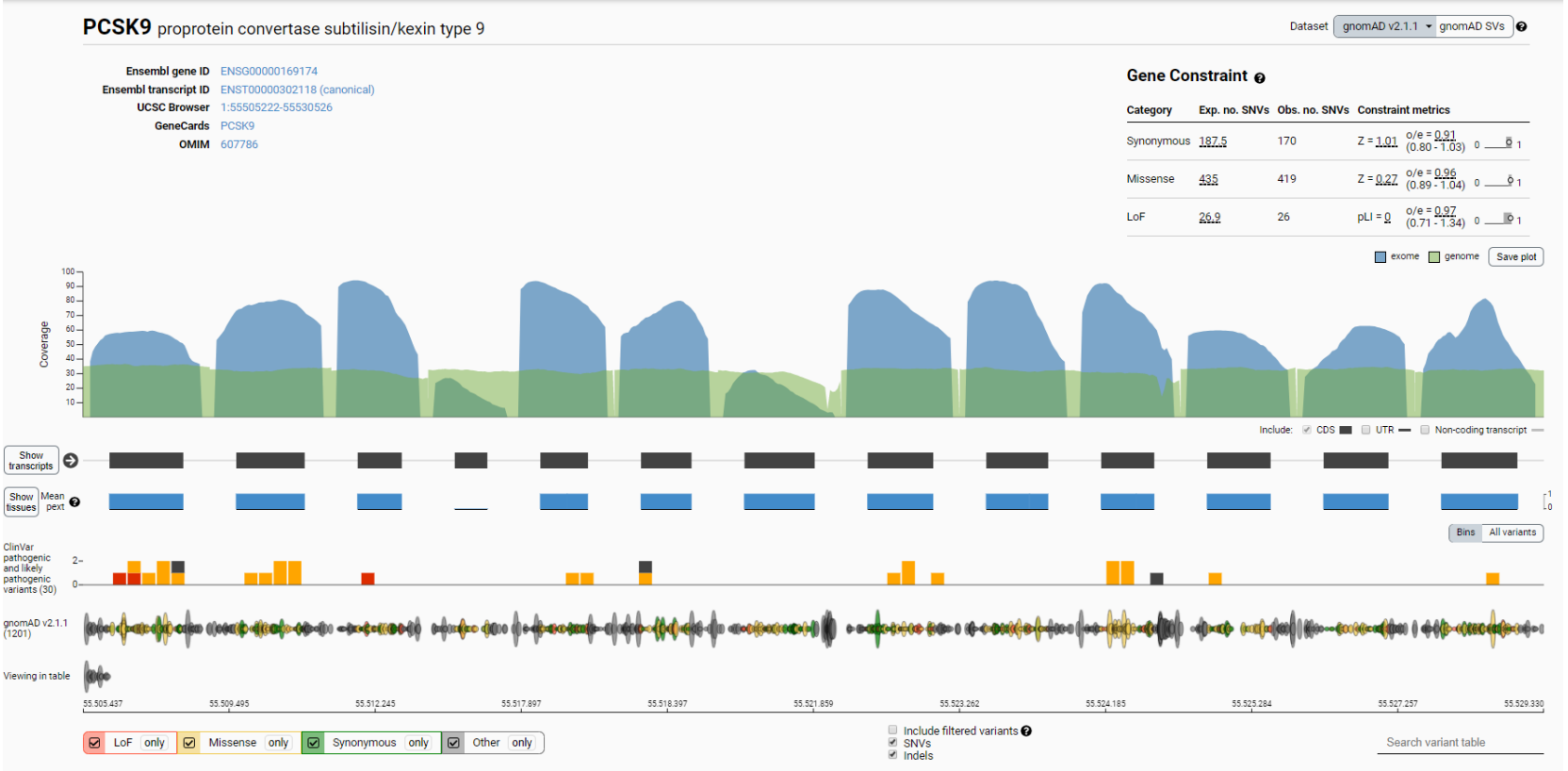
## Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2*], Eric V. Minikel[1,2,5*], Kaitlin E. Samocha[1,2,5,6*], Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P. Birnbaum[2], Jack A. Kosmicki[1,2,6,13], Laramie E. Duncan[1,2,6], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2], Emma Pierce-Hoffman[1,2], Joanne Berghout[14,15], David N. Cooper[16], Nicole Deflaux[17], Mark DePristo[18], Ron Do[19,20,21,22], Jason Flannick[2,23], Menachem Fromer[1,6,19,20,24], Laura Gauthier[18], Jackie Goldstein[1,2,6], Namrata Gupta[2], Daniel Howrigan[1,2,6], Adam Kiezun[18], Mitja I. Kurki[2,25], Ami Levy Moonshine[18], Pradeep Natarajan[2,26,27,28], Lorena Orozco[29], Gina M. Peloso[2,27,28], Ryan Poplin[18], Manuel A. Rivas[2], Valentin Ruano-Rubio[18], Samuel A. Rose[6], Douglas M. Ruderfer[19,20,24], Khalid Shakir[18], Peter D. Stenson[16], Christine Stevens[2], Brett P. Thomas[1,2], Grace Tiao[18], Maria T. Tusie-Luna[30], Ben Weisburd[2], Hong-Hee Won[31], Dongmei Yu[6,25,27,32], David M. Altshuler[2,33], Diego Ardissino[34], Michael Boehnke[35], John Danesh[36], Stacey Donnelly[2], Roberto Elosua[37], Jose C. Florez[2,26,27], Stacey B. Gabriel[2], Gad Getz[18,26,38], Stephen J. Glatt[39,40,41], Christina M. Hultman[42], Sekar Kathiresan[2,26,27,28], Markku Laakso[43], Steven McCarroll[6,8], Mark I. McCarthy[44,45,46], Dermot McGovern[47], Ruth McPherson[48], Benjamin M. Neale[1,2,6], Aarno Palotie[1,2,5,49], Shaun M. Purcell[19,20,24], Danish Saleheen[50,51,52], Jeremiah M. Scharf[2,6,25,27,32], Pamela Sklar[19,20,24,53,54], Patrick F. Sullivan[55,56], Jaakko Tuomilehto[57], Ming T. Tsuang[58], Hugh C. Watkins[44,59], James G. Wilson[60], Mark J. Daly[1,2,6], Daniel G. MacArthur[1,2] & Exome Aggregation Consortium†

**Large-scale reference data sets of human genetic variation are critical for the medical and functional interpretation of DNA sequence changes. Here we describe the aggregation and analysis of high-quality exome (protein-coding region) DNA sequence data for 60,706 individuals of diverse ancestries generated as part of the Exome Aggregation Consortium (ExAC). This catalogue of human genetic diversity contains an average of one variant every eight bases of the exome, and provides direct evidence for the presence of widespread mutational recurrence. We have used this catalogue to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation; identifying 3,230 genes with near-complete depletion of predicted protein-truncating variants, with 72% of these genes having no currently established human disease phenotype. Finally, we demonstrate that these data can be used for the efficient filtering of candidate disease-causing variants, and for the discovery of human 'knockout' variants in protein-coding genes.**

# gNOMAD

- Aggregating and harmonizing both exome and genome sequencing data
- 125,748 exome sequences
- 15,708 whole-genome sequences



# CADD

Combined Annotation Dependent Depletion

- Scoring the deleteriousness of SNVs/INDELS in human genome
- Integrates multiple annotations into one metric
  - 63 distinct annotations (e.g., GERP, phyloP; transcription factor binding, transcript information SIFT, and PolyPhen)
- Trained a support vector machine (SVM)
- Scores freely available for research

http://cadd.gs.washington.edu/info

# Visualization

# Visualization

Genome browsers - most widely-used tools

**Read**

- SAM/BAM
- VCF
- GTF/GFF/BED
- FASTA
- ...

**Able to**

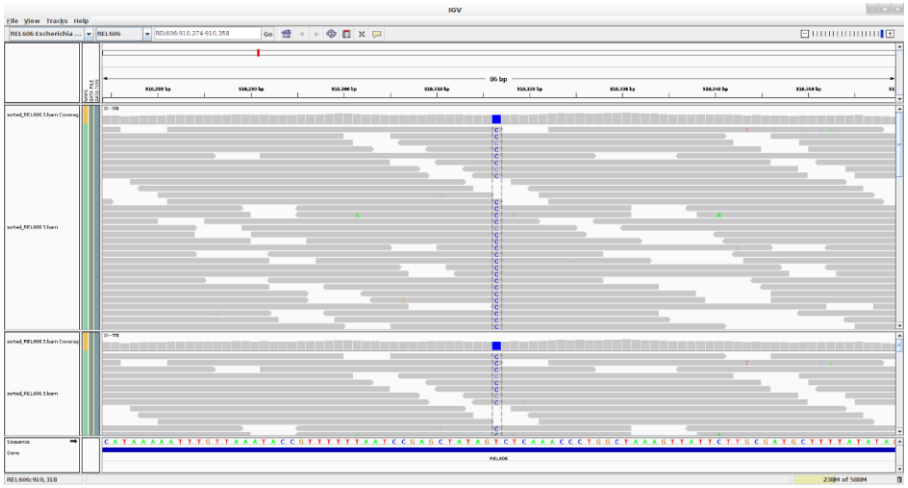
- Browse/zoom genome
- Display multiple samples / multiple tracks
- Colorize/mark features of your data (paired reads, SNPs, ...)

# Genome Browsers

**IGV** (Integrative Genomics Viewer)

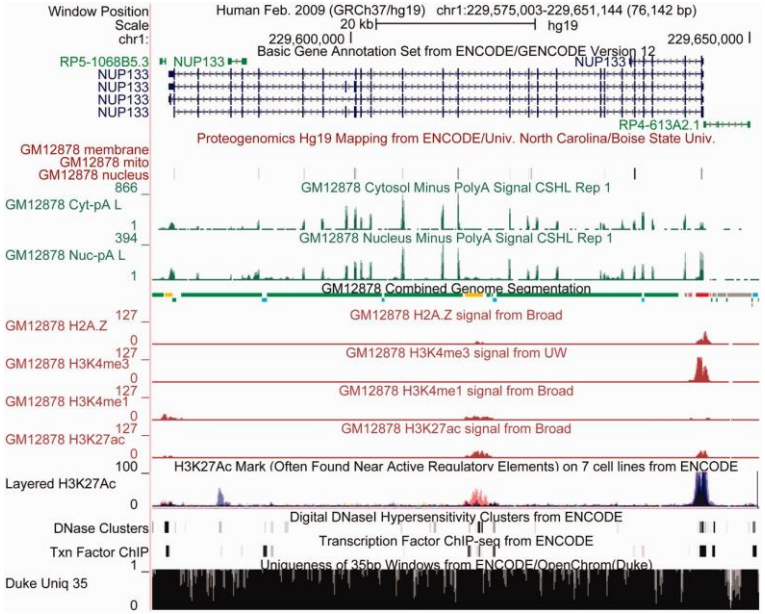
- Widely used viewer
- Java based – standalone tool
- Easy and fast to view own data
- IGV3 **supports long-reads**

http://www.pacb.com/blog/igv-3-improves-support-pacbio-long-reads/



**UCSC**

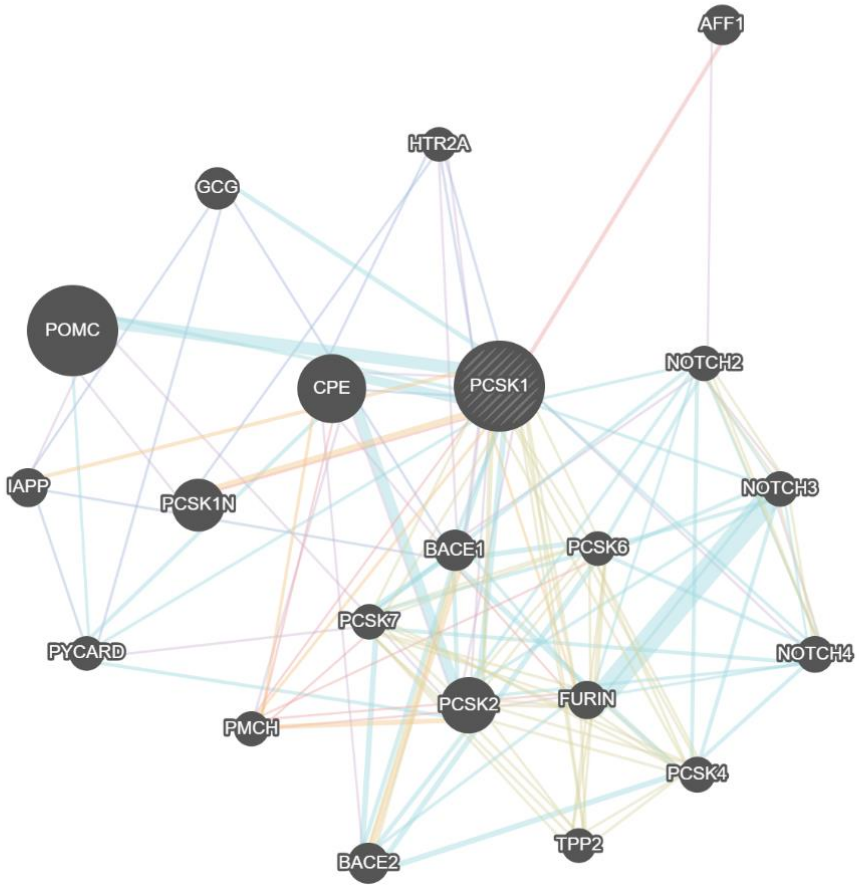
- Web based tool
- Offers many different annotation tracks
- Needs some configuration to display own data



# GeneMANIA - Network of genes

- Enter multiple genes

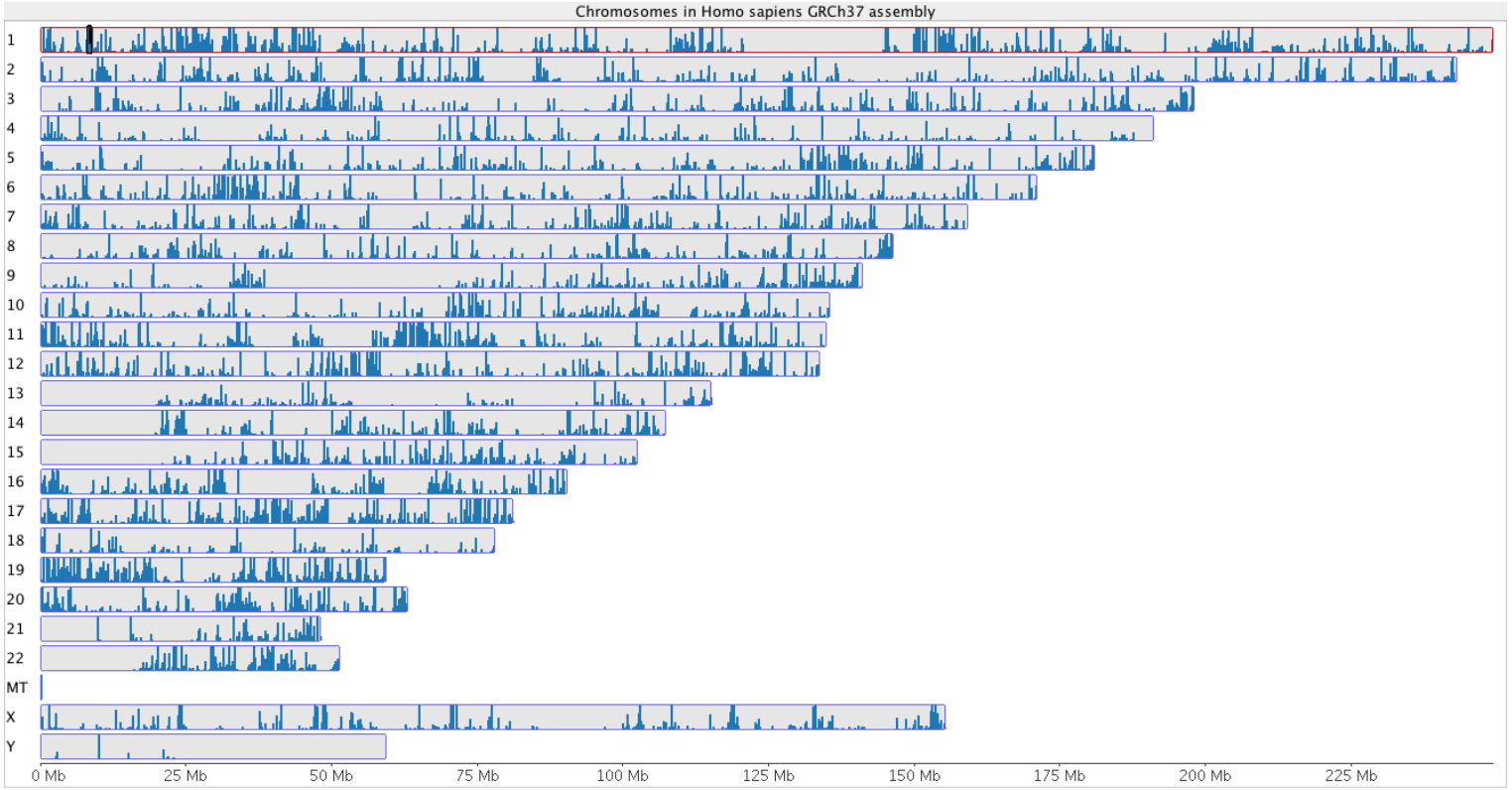

- Interpret their context



# Coverage visualization

Coverage histogram for chromosomes

- http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/

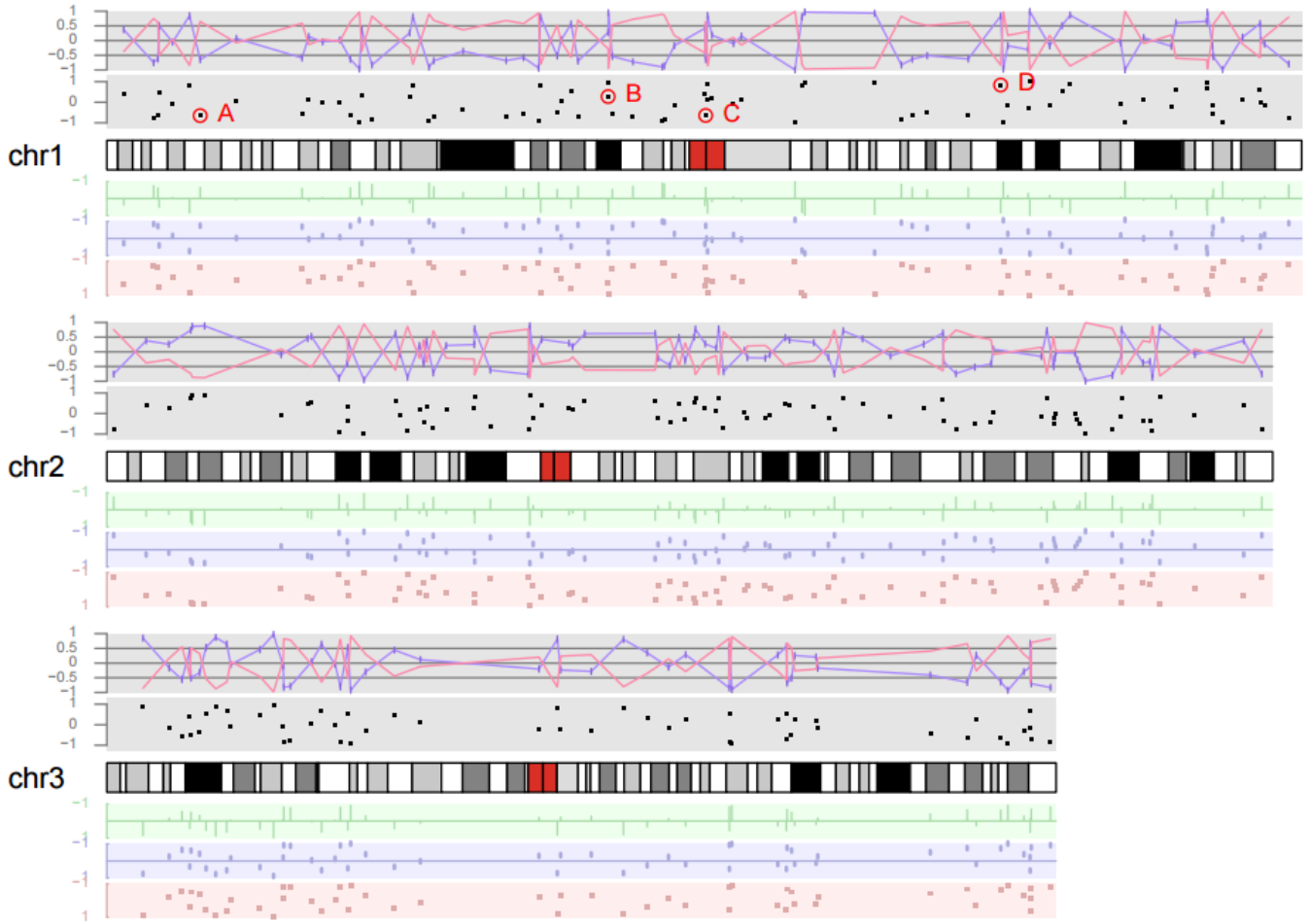



http://seqanswers.com/forums/attachment.php?attachmentid=2118&d=1364889859

# Genome-wide visualization

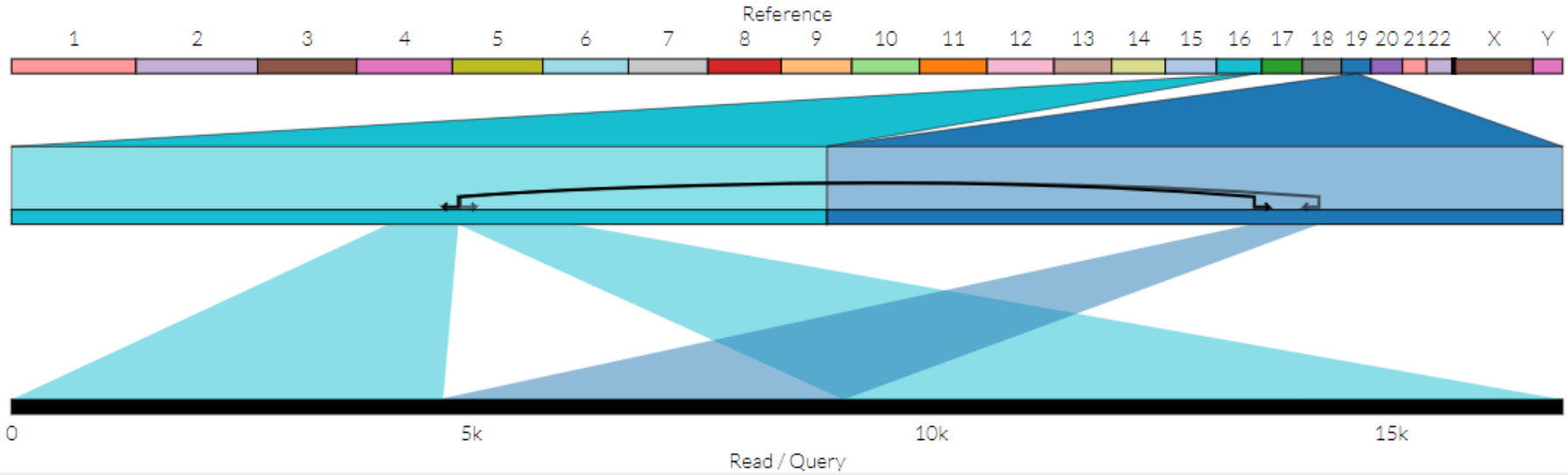
## karyoploteR

- Customizable karyotypes with arbitrary data



# Visualization of structural variants

Ribbon: Visualizing complex genome alignments and structural variation



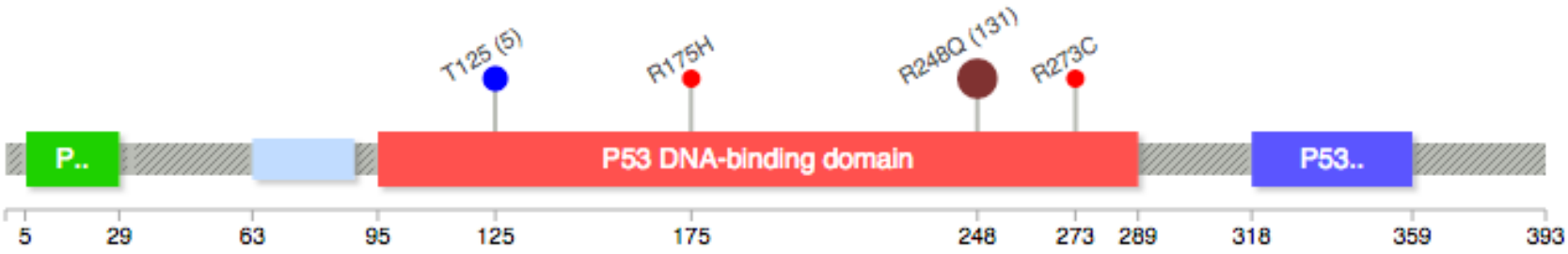

Maria Nattestad, Chen-Shan Chin, Michael C. Schatz; https://www.biorxiv.org/content/10.1101/082123v1

# Lollipop

Lollipop-style mutation diagrams for annotating genetic variations

- https://github.com/pbnjay/lollipops/blob/master/README.md

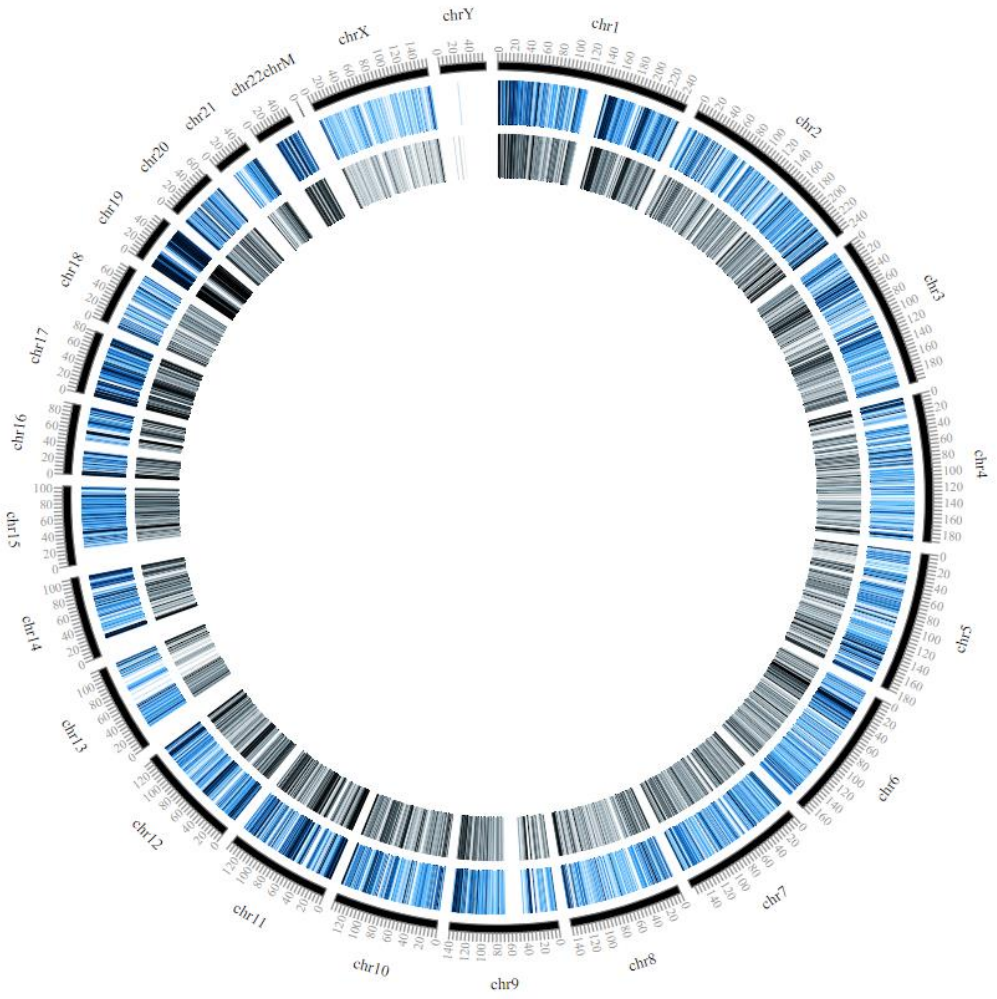
```
./lollipops -labels TP53 R248Q#7f3333@131 R273C R175H T125@5
```



# CircosVCF

http://legolas.ariel.ac.il/~tools/CircosVCF/

- Interactive tool
- Web-based



# Analysis pipelines and workflow systems

# Pipelines

**Bcbio pipeline**

- Python toolkit providing best-practice pipelines
- DNASeq and RNASeq pipelines

**Nextflow**

- Many pipelines available (https://github.com/nextflow-io/awesome-nextflow)
- Supports Docker

**ngs_backbone**

- NGS analysis as well as with sanger sequences
- BWA, GATK, blast  --- read cleaning, ORF annotation

# Galaxy

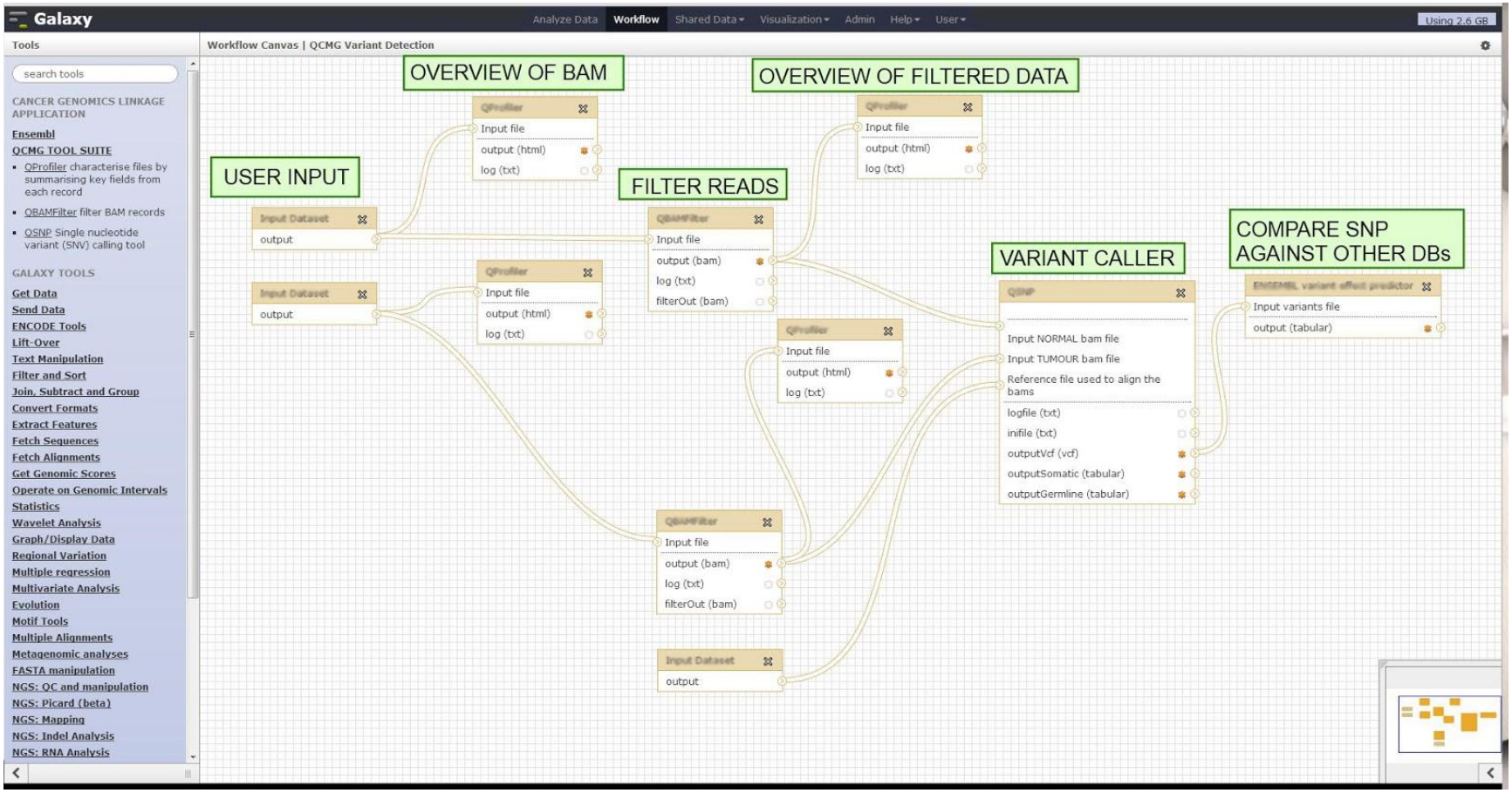
"Galaxy is an open, web-based platform for data intensive biomedical research"

- Workflow & data integration platform
- Computational biology for users without programming experience
- Includes wrappers for many tools
- Store history of workflows → reproducibility

- Public instances to analyze the data
- Existing workflows for DNASeq & RNASeq ...

- Can be locally installed and used

# Galaxy

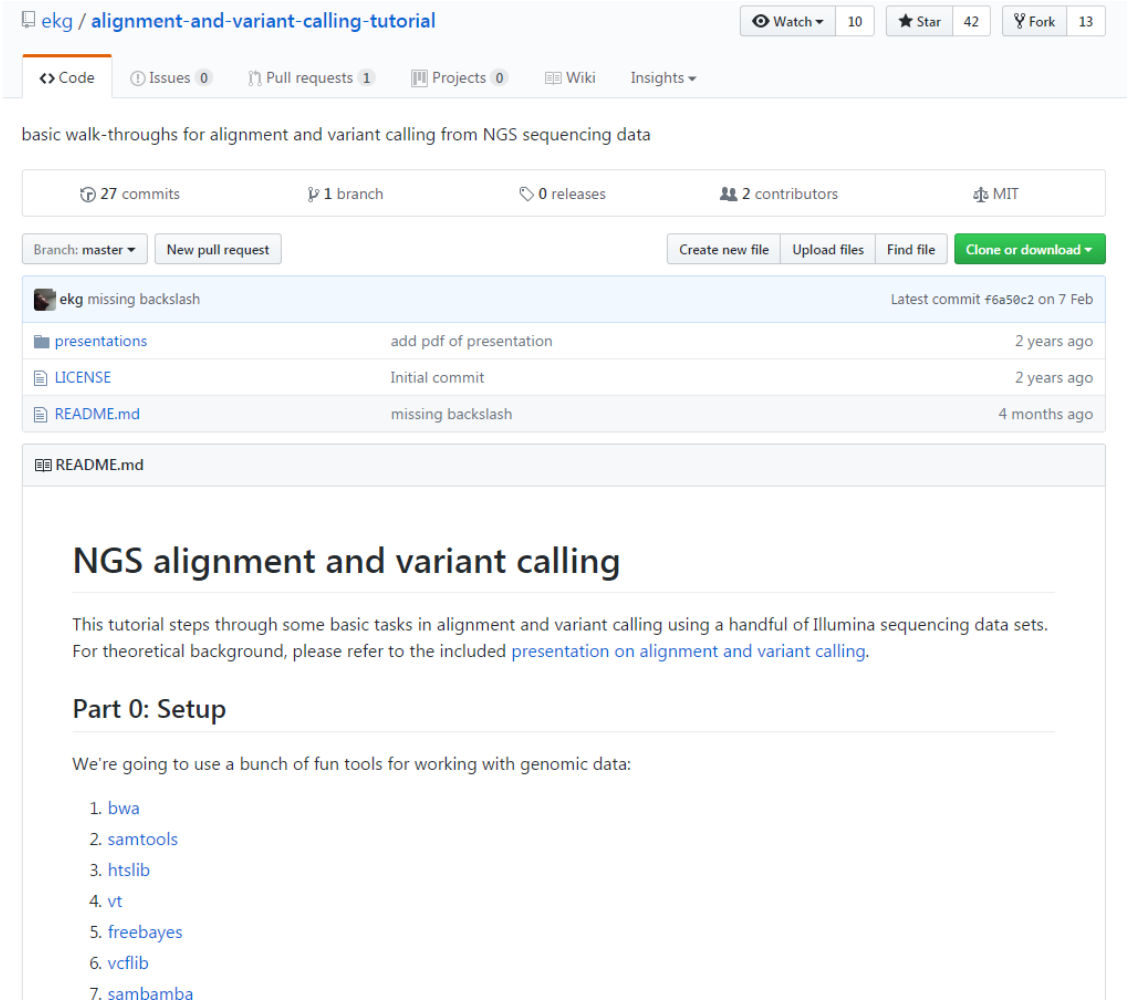



„The Cancer Genomics Linkage Application”

# Other useful information

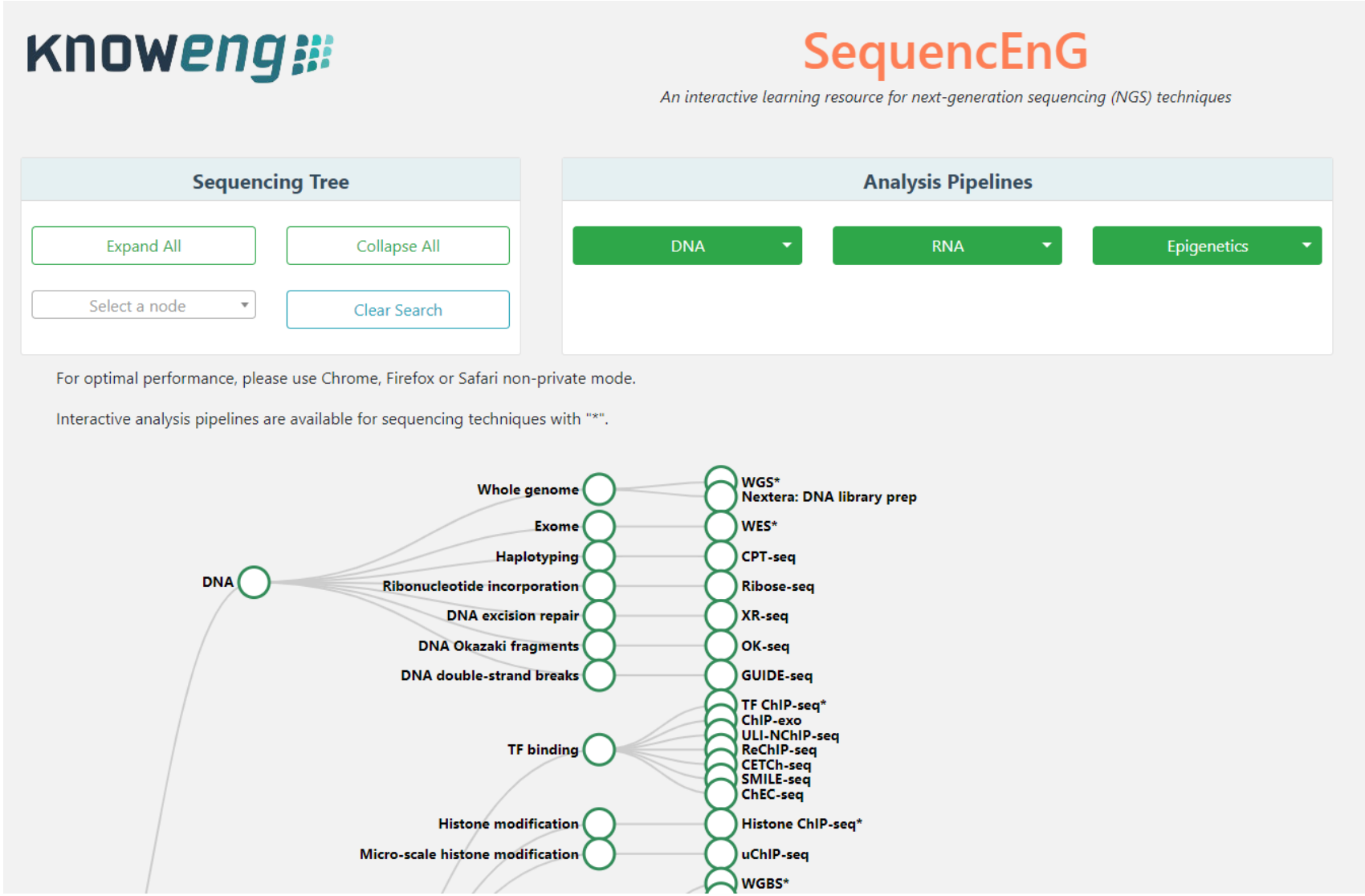
# Useful information on how-to perform variant calling

https://github.com/ekg/alignment-and-variant-calling-tutorial



# SequencEnG

http://education.knoweng.org/sequenceng/index.html



# Genome in a Bottle (GIAB)

Develop the **technical infrastructure** (reference standards, reference methods, and reference data) to enable **translation of whole human genome sequencing to clinical practice**.

- Github repository:
  https://github.com/genome-in-a-bottle
- Pilot genome Reference Material
  - genomic DNA (NA12878)
  - derived from a large batch of the Coriell cell line GM12878
  - high-confidence SNPs, INDEL, and homozygous reference regions
- Four new GIAB reference materials available
- http://jimb.stanford.edu/giab/

# IGSR: International Genome Sample Resource

- Provides ongoing support for the 1000 Genomes Project data
- Usability of the 1000 Genomes reference data
- Data repository (raw, mapped, variant calling)

IGSR and the 1000 Genomes Project


The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR.

# Recommendations

- Choose sequencing system according to your needs
- Use transparent analysis systems
- Optimize analysis settings to use-case
- Check technical properties of variants (coverage, strand, qualities, …)
- Look at variants in genome browser

# Where can you get help and information?

**Biostar**

- A high quality question & answer Web site.

**SEQanswers**

- A discussion and information site for next-generation sequencing.

**http://omictools.com/**

- An informative directory for multi-omic data analysis

**Rosalind (http://rosalind.info/)**

- Platform for learning bioinformatics through problem solving
- Also used for a coursera course https://www.coursera.org/course/bioinformatics

**Collection of helps**

http://www.acgt.me/blog/2015/11/1/where-to-ask-for-bioinformatics-help-online

# Useful information

**List of one liners**

https://github.com/stephenturner/oneliners

## Basic awk & sed

Extract fields 2, 4, and 5 from file.txt:

```
awk '{print $2,$4,$5}' input.txt
```

Print each line where the 5th field is equal to ‘abc123’:

```
awk '$5 == "abc123"' file.txt
```

Print each line where the 5th field is *not* equal to ‘abc123’:

```
awk '$5 != "abc123"' file.txt
```

Print each line whose 7th field matches the regular expression:

```
awk '$7  ~ /^[a-f]/' file.txt
```

Print each line whose 7th field *does not* match the regular expression:

```
awk '$7 !~ /^[a-f]/' file.txt
```

Get unique entries in file.txt based on column 1 (takes only the first instance):

# SAM and BAM filtering oneliners

https://gist.github.com/davfre/8596159

bamfilter_oneliners.md Raw

SAM and BAM filtering one-liners

@author: David Fredman, david.fredmanAAAAAA@gmail.com (sans poly-A tail)
@dependencies: http://sourceforge.net/projects/bamtools/ and http://samtools.sourceforge.net/

Please comment or extend with additional/faster/better solutions.

BWA mapping (using piping for minimal disk I/O)

```
bwa aln -t 8 targetGenome.fa reads.fastq | bwa samse targetGenome.fa - reads.fastq\
| samtools view -bt targetGenome.fa - | samtools sort - reads.bwa.targetGenome

samtools index reads.bwa.targetGenome.bam
```

Count number of records (unmapped reads + each aligned location per mapped read) in a bam file:

```
samtools view -c filename.bam
```

Count with flagstat for additional information:

```
samtools flagstat filename.bam
```

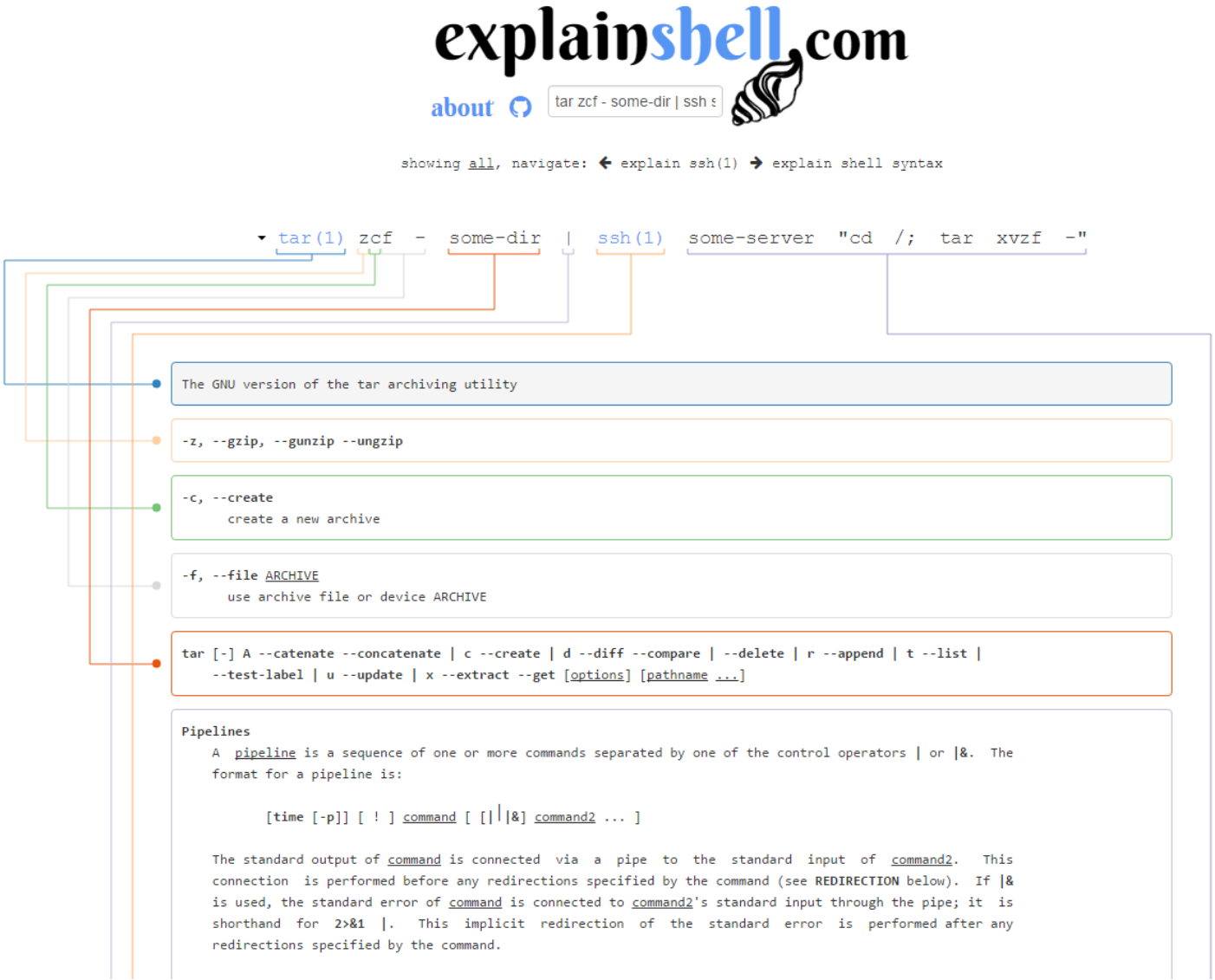
Count the number of alignments (reads mapping to multiple locations counted multiple times)

# Collection of published “guides” for bioinformaticians

http://biomickwatson.wordpress.com/2013/11/05/collection-of-published-guides-for-bioinformaticians/

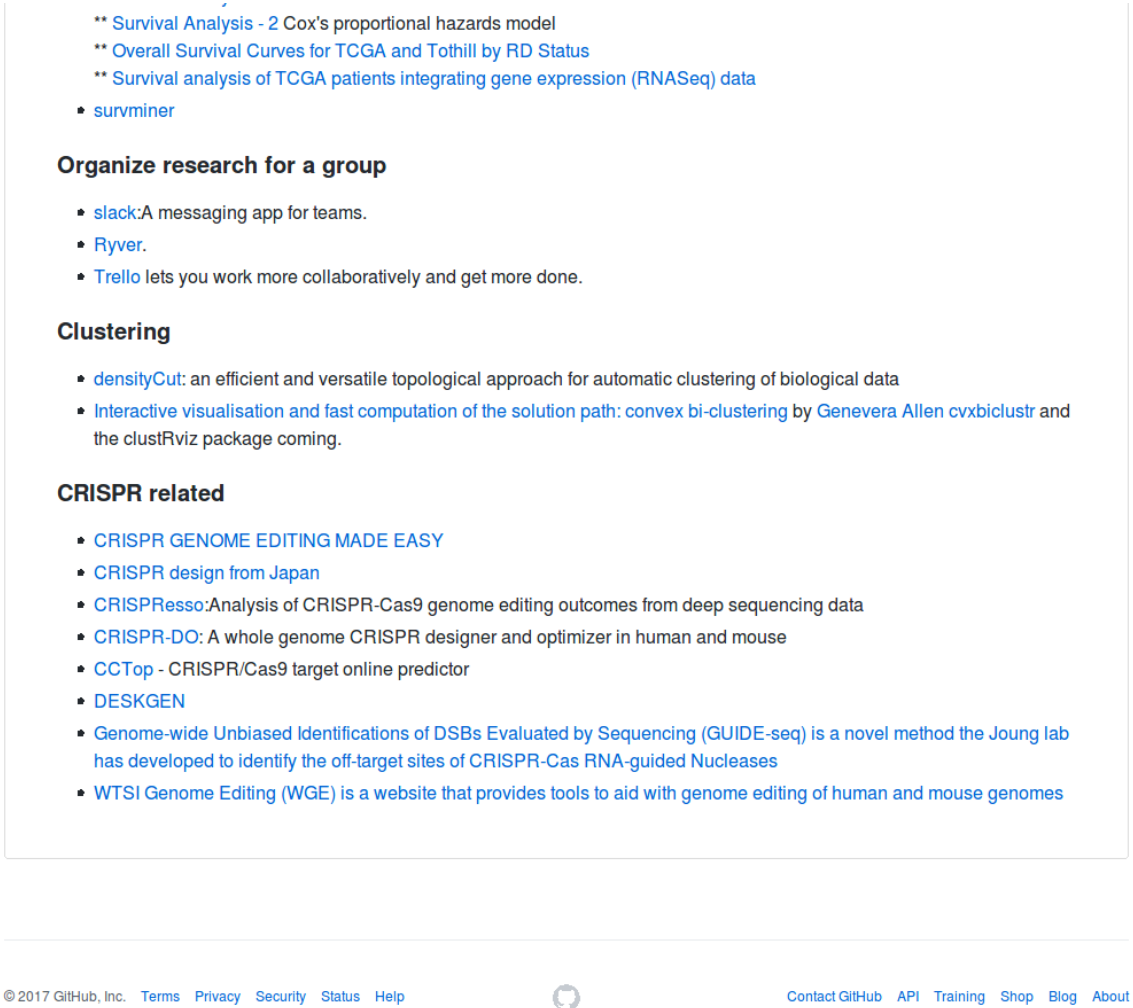
1. Loman N and Watson M (2013) So you want to be a computational biologist? *Nature Biotech* **31(11)**:996-998 [link]
2. Corpas M, Fatumo S, Schneider R. (2012) How not to be a bioinformatician. *Source Code Biol Med.* **7(1)**:3. [link]
3. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SHD, Huff K, Mitchell IM, Plumbley M, Waugh B, White EP, WIlson P (2013) Best Practices for Scientific Computing. *arXiv* http://arxiv.org/abs/1210.0530 [link]
4. Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. *PLoS Comput Biol* **9(10)**: e1003285. [link]
5. Bourne PE (2011) Ten Simple Rules for Getting Ahead as a Computational Biologist in Academia. *PLoS Comput Biol* **7(1)**: e1002001. [link]
6. Oshlack A (2013) A 10-step guide to party conversation for bioinformaticians. *Genome Biology* **14**:104 [link]
7. Via A, De Las Rivas J, Attwood TK, Landsman D, Brazas MD, et al. (2011) Ten Simple Rules for Developing a Short Bioinformatics Training Course. *PLoS Comput Biol* **7(10)**: e1002245. [link]
8. Via A, Blicher T, Bongcam-Rudloff E, Brazas MD, Brooksbank C, Budd A, De Las Rivas J, Dreyer J, Fernandes PL, van Gelder C, Jacob J, Jimenez RC, Loveland J

# Explain Shell - https://explainshell.com/



# Huge resource

https://github.com/crazyhottommy/getting-started-with-genomics-tools-and-resources



** Survival Analysis - 2 Cox's proportional hazards model
** Overall Survival Curves for TCGA and Tothill by RD Status
** Survival analysis of TCGA patients integrating gene expression (RNASeq) data

- survminer

### Organize research for a group

- slack:A messaging app for teams.
- Ryver.
- Trello lets you work more collaboratively and get more done.

### Clustering

- densityCut: an efficient and versatile topological approach for automatic clustering of biological data
- Interactive visualisation and fast computation of the solution path: convex bi-clustering by Genevera Allen cvxbiclustr and the clustRviz package coming.

### CRISPR related

- CRISPR GENOME EDITING MADE EASY
- CRISPR design from Japan
- CRISPResso:Analysis of CRISPR-Cas9 genome editing outcomes from deep sequencing data
- CRISPR-DO: A whole genome CRISPR designer and optimizer in human and mouse
- CCTop - CRISPR/Cas9 target online predictor
- DESKGEN
- Genome-wide Unbiased Identifications of DSBs Evaluated by Sequencing (GUIDE-seq) is a novel method the Joung lab has developed to identify the off-target sites of CRISPR-Cas RNA-guided Nucleases
- WTSI Genome Editing (WGE) is a website that provides tools to aid with genome editing of human and mouse genomes

AUSTRIAN INSTITUTE OF TECHNOLOGY

# Thanks