AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Tools for variant analysis of next-generation genome sequencing

Part 1

Stephan Pabinger
stephan.pabinger@ait.ac.at

What to expect

- Finish the day 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
  - Alignment metrics reports
  - Base quality score recalibration
  - Variant calling
  - Variant annotation and filtering

You might experience some overlap with other modules of this workshop

My background

Bioinformatics

- Software development
- Web tools
- Pipeline design

Working with sequencing data

- DNASeq
- RNASeq
- MethylationSeq

Now at bioinformatics team @ Austrian Institute of Technology (AIT)
http://www.ait.ac.at/bioinformatics

Structure

Lecture 1

- File formats, Preprocessing
- Variant calling

Practical 1

- QC of mapping
- File manipulations and variant calling

Lecture 2

- Variant Calling
- Variant Annotation & Workflow Systems

Practical 2

- Variant Calling
- Annotation

Hints for the practicals – general hints

Results of trial and error

Hints – hints - hints

- Write every command in a file -> easy to create small scripts in Linux
- Use variables in scripts
- Write down the versions of used tools

- Document!

- Backup your scripts and raw data
- Use a version control system if available (or github)

GIT

Information

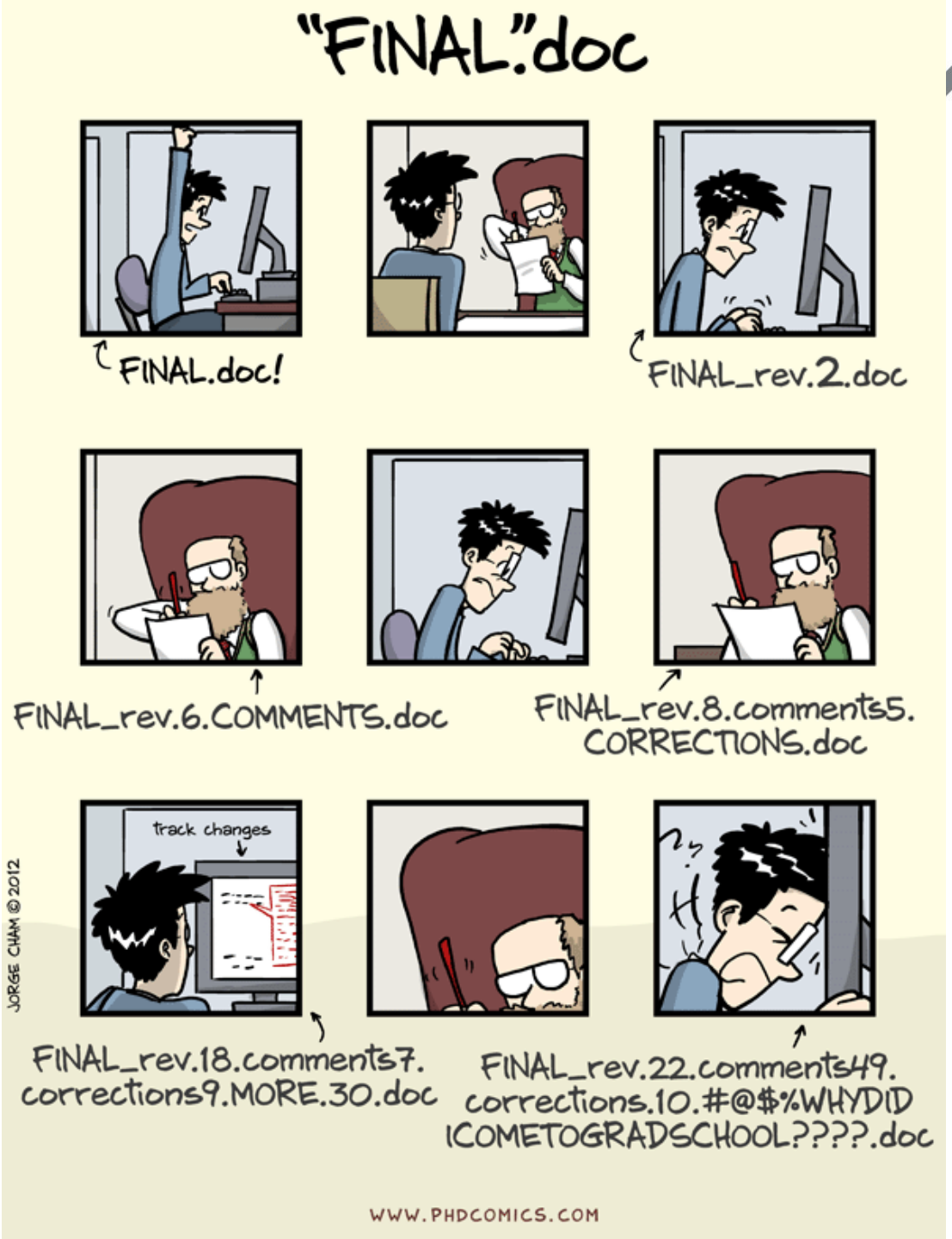
- Distributed revision control system
- Developed by Linus Torvalds (Linux developer)
- Local repositories & remote repositories

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“`

```
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 analys
[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`

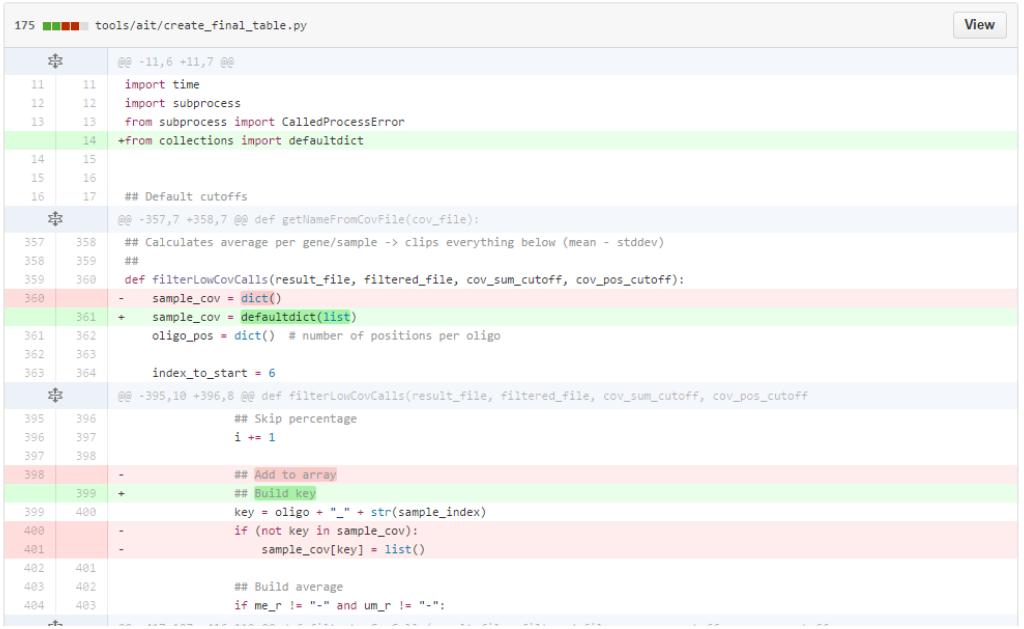
```
stephan@shaq /tmp/super_analysis $ git status
On branch master
Untracked files:
  (use "git add <file>..." to include in what will be committed)

        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



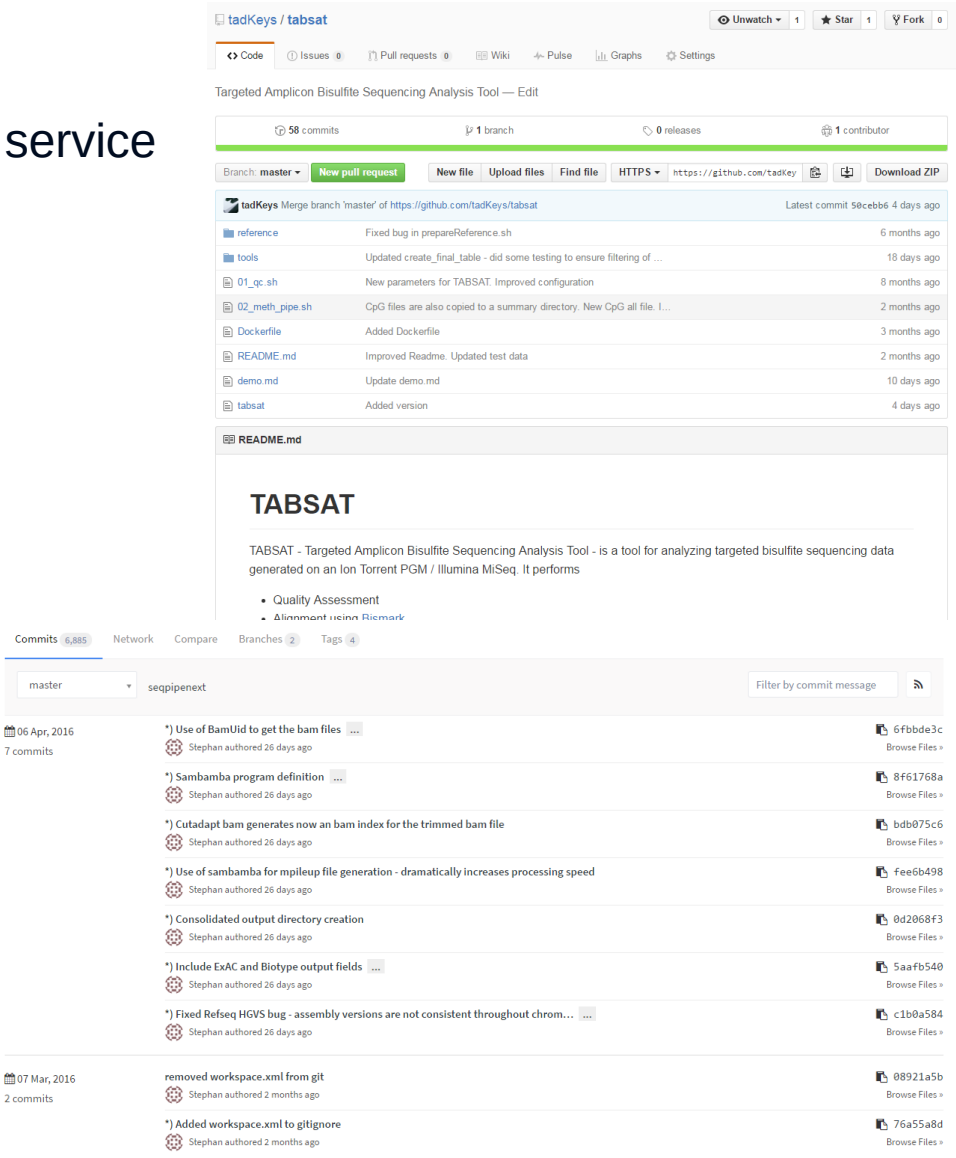
Github - Gitlab

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)

http://rogerdudler.github.io/git-guide/
http://kbroman.org/github_tutorial/
http://nyuccl.org/pages/GitTutorial/
https://swcarpentry.github.io/git-novice/

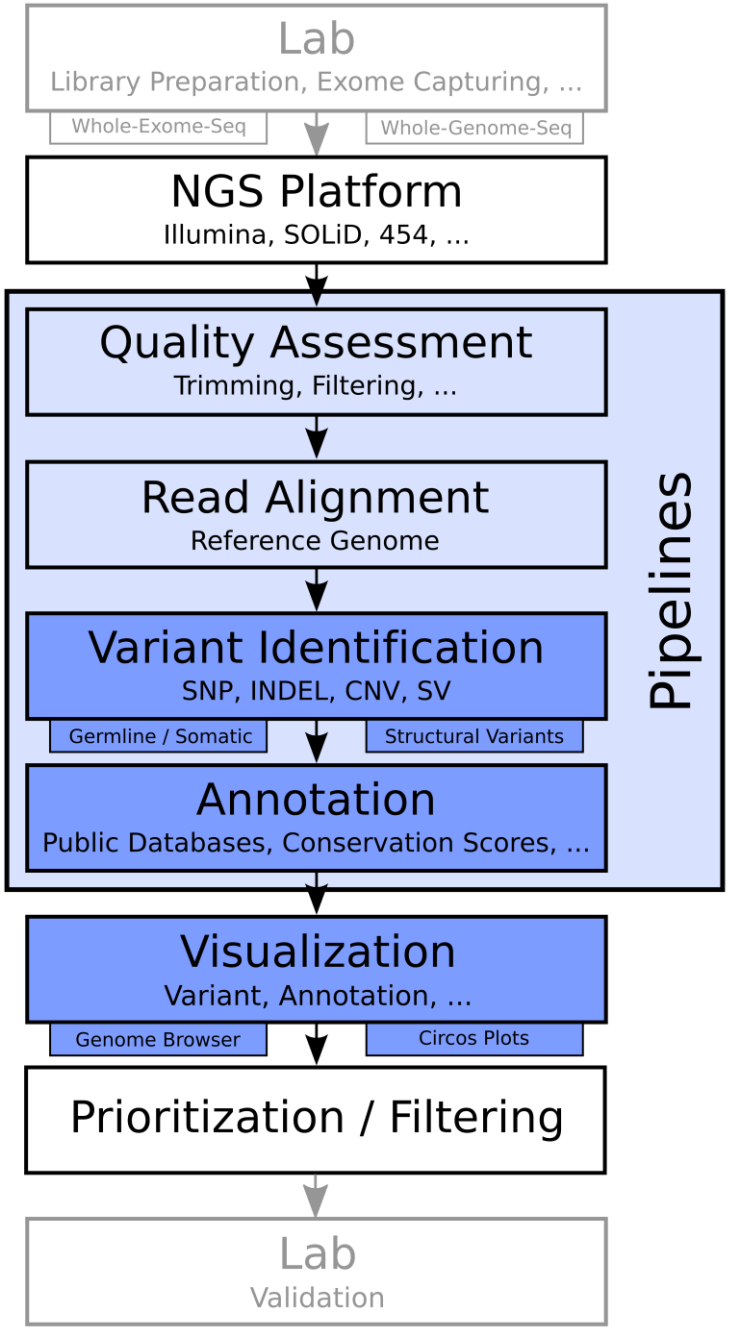
Windows

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

Linux & Mac

- Packages and GUIs available

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

Before we start

More useful information ...

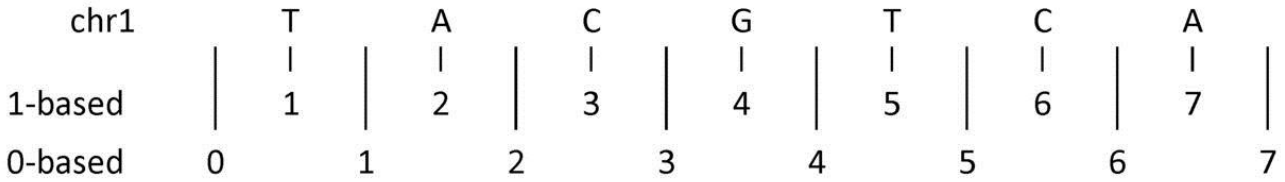
Coordinate system

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

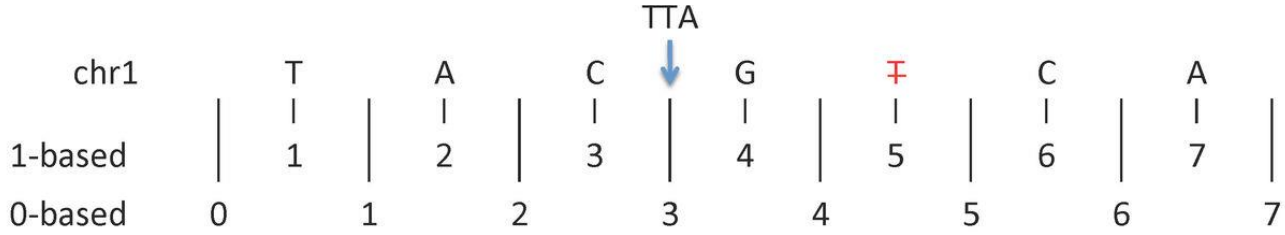
- BED – 0 based
- GFF – 1 based
- Ensembl uses a one-based coordinate system - UCSC use a zero-based coordinate system

	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/

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%

Widely used tools – also used in the practicals

AUSTRIAN INSTITUTE OF TECHNOLOGY

SAMtools

- Utilities for manipulating SAM files

GATK

- Genome Analysis Toolkit - wide variety of tools, variant discovery and genotyping, quality assurance

Picard

- Utilities for manipulating SAM files

Genetic variations

Genetic variation

A regular diploid human cell contains 46 chromosomes (23 pairs)

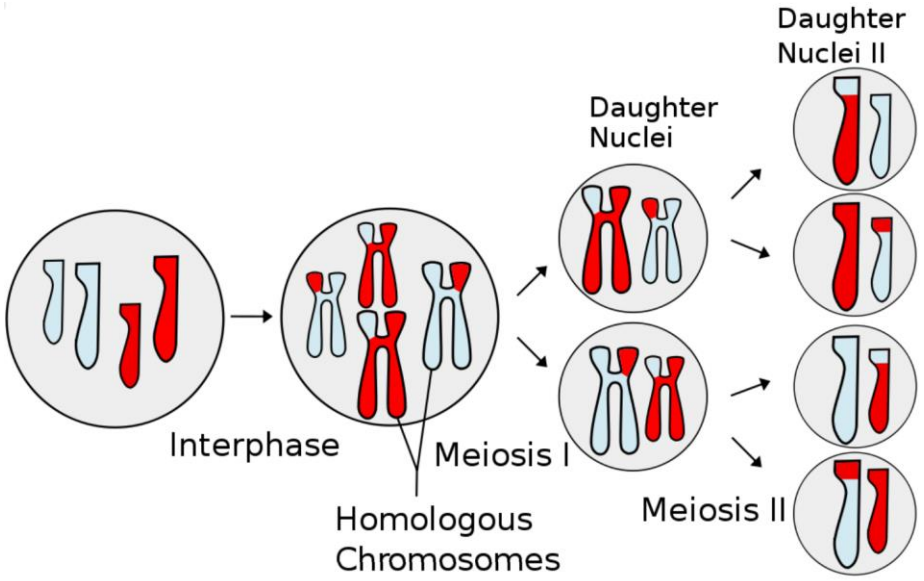
- 22 pairs + sex chromosomes XX(female) XY(male)
- One set of chromosomes inherited from each parent

Germline

- Meiosis → four genetically unique haploid gametes that each contain a unique mixture of the genetic code of the maternal and paternal chromosomes of the cell

Somatic mutation

- Not inherited from parent
- Acquired from spontaneous mutations during DNA replication



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

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, …
- ~550m submissions, ~150m variants (stats only for human) [v147]

Problems

- high FP rate
- not many validated SNPs (~40%)

→ careful when filtering out variants based on dbSNP information

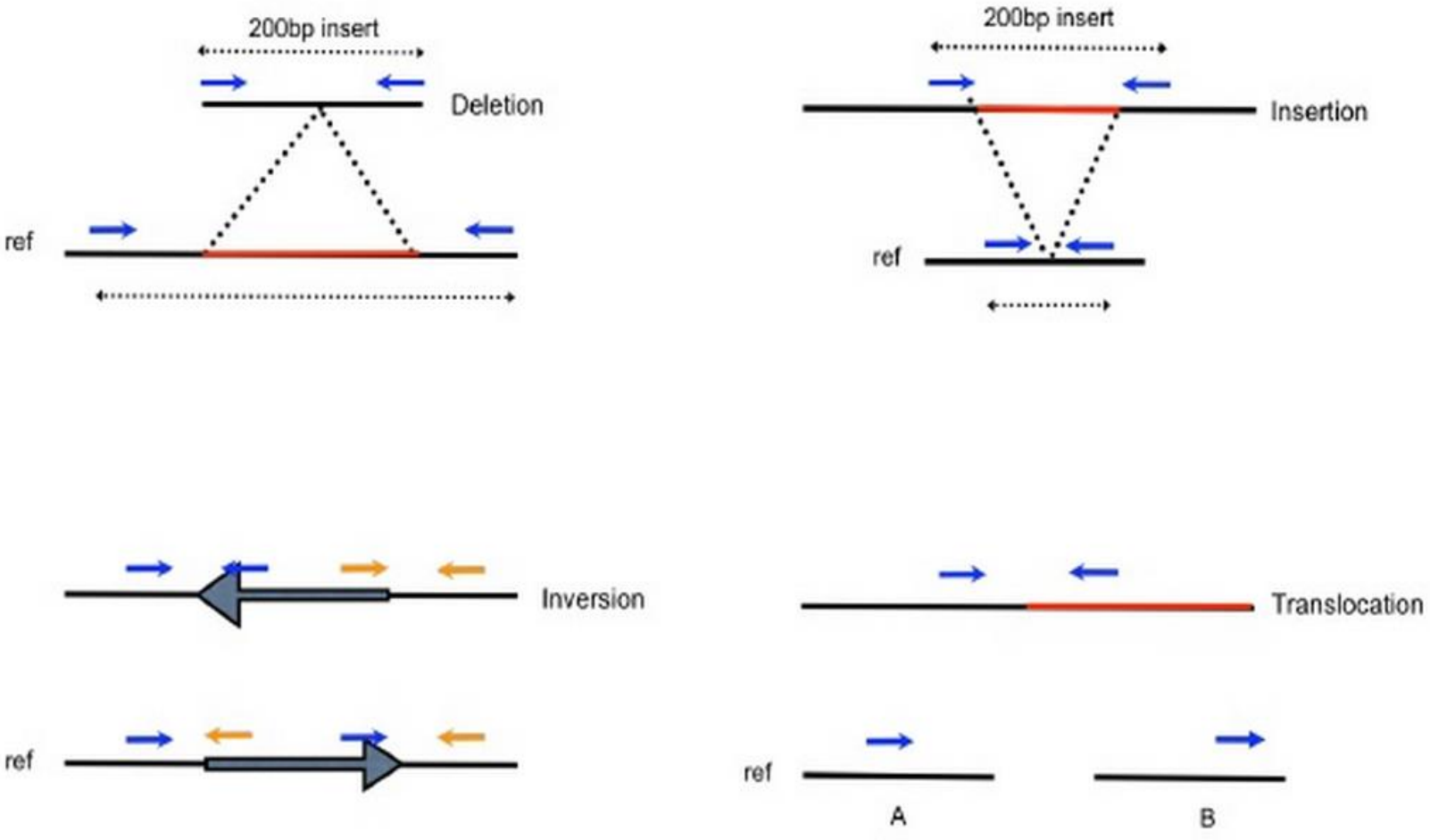
Types of genetic variations

Structural variations (SV)

- Variation in structure of an organism's chromosome

- Insertions
  Deletions
  CNV
  Inversions
  Translocations

Types of genetic variations



Types of genetic variations

Structural variations

dbVar

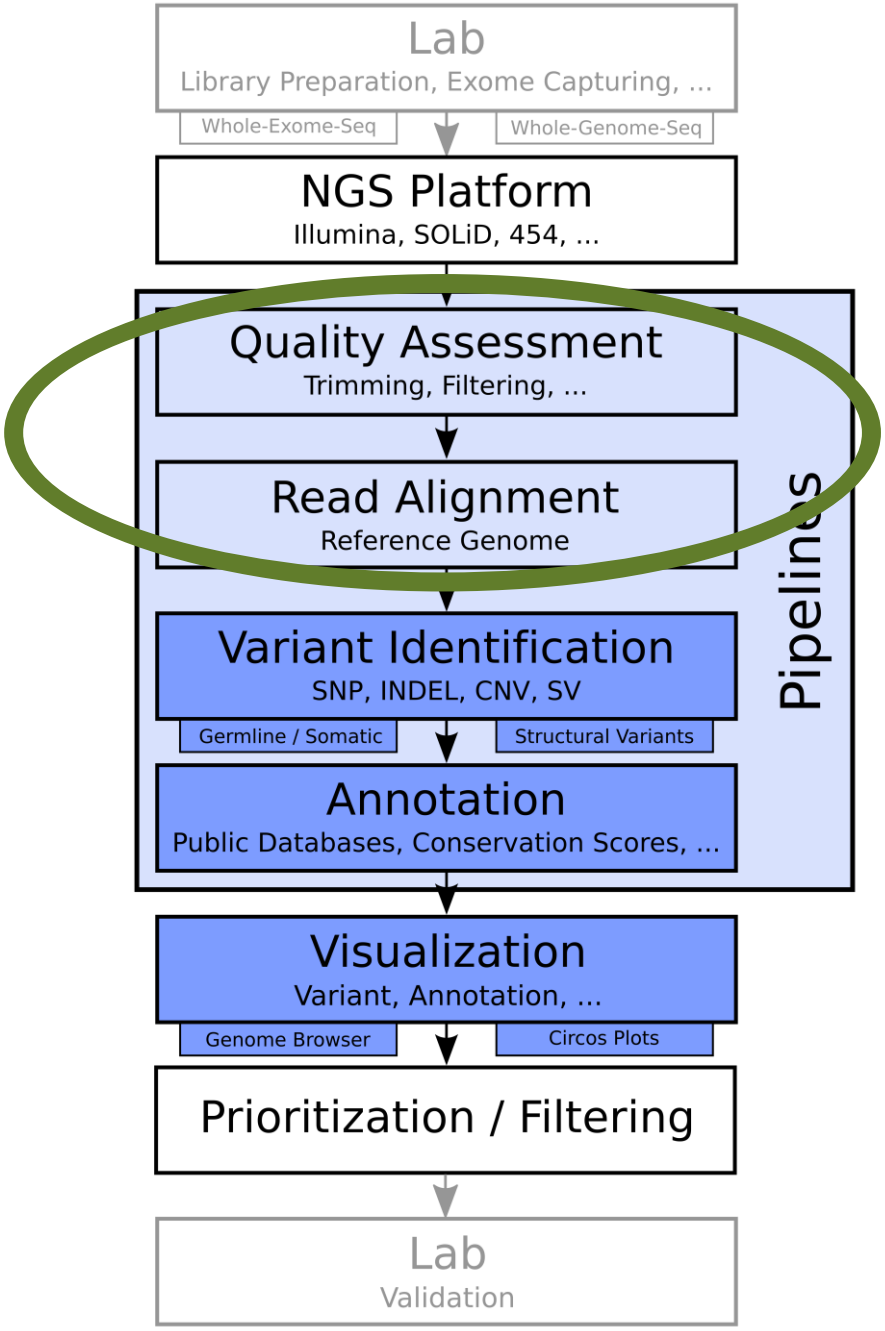
- Database of genomic structural variation
- http://www.ncbi.nlm.nih.gov/dbvar/

Database of Genomic Variants archive

- Repository that provides archiving, accessioning and distribution
- Available in all species
- http://www.ebi.ac.uk/dgva

AUSTRIAN INSTITUTE OF TECHNOLOGY

Genome reference


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

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

	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

AUSTRIAN INSTITUTE OF TECHNOLOGY

FASTA & FASTQ

Cleaning up fastq files

Before mapping make sure

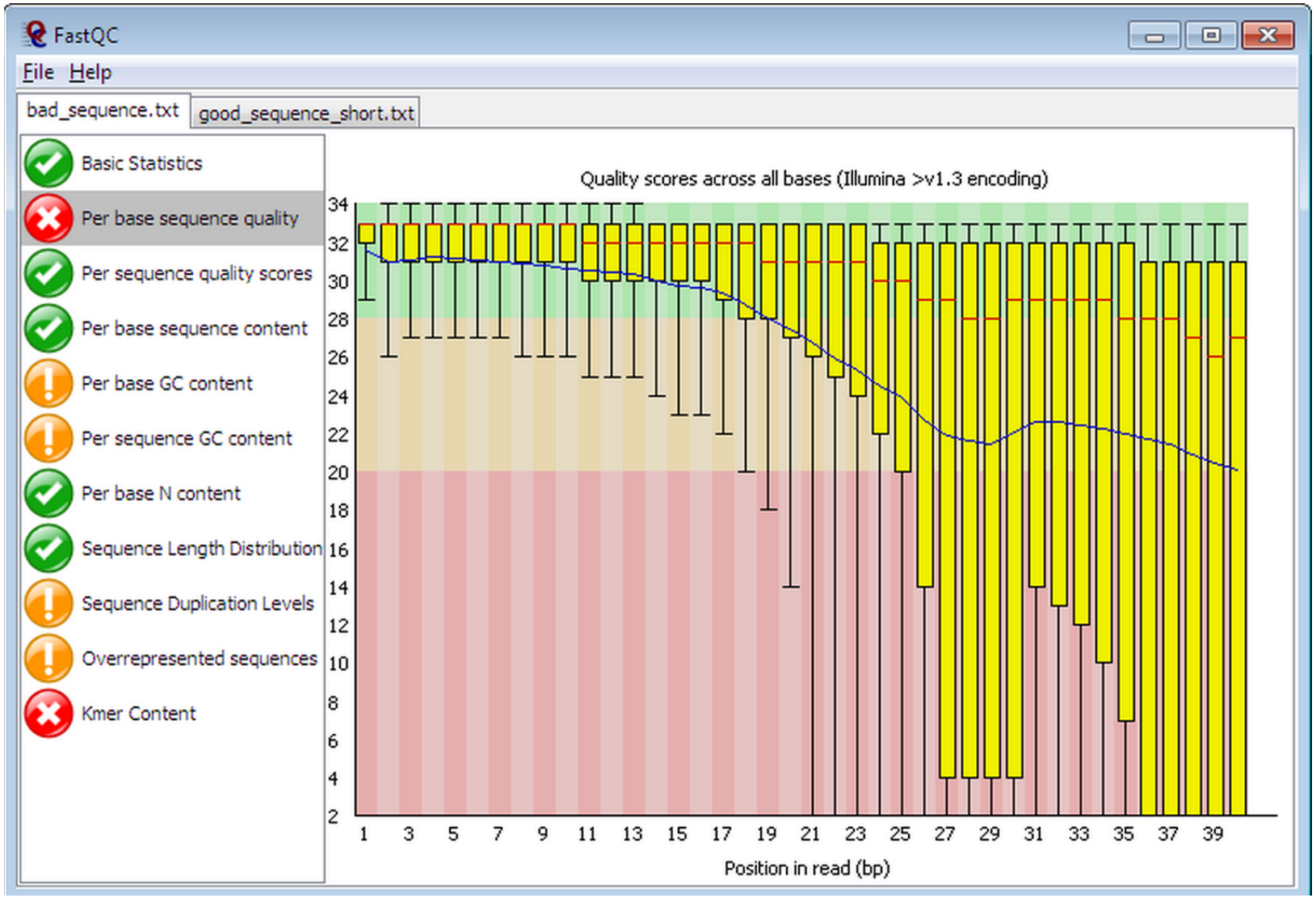
- Non genomic sequences are removed (barcodes, …)
- Adapter sequences are removed
- Clean contaminations (PRINSEQ, DeconSeq)
- Trim Ns
- Trim bad quality reads

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

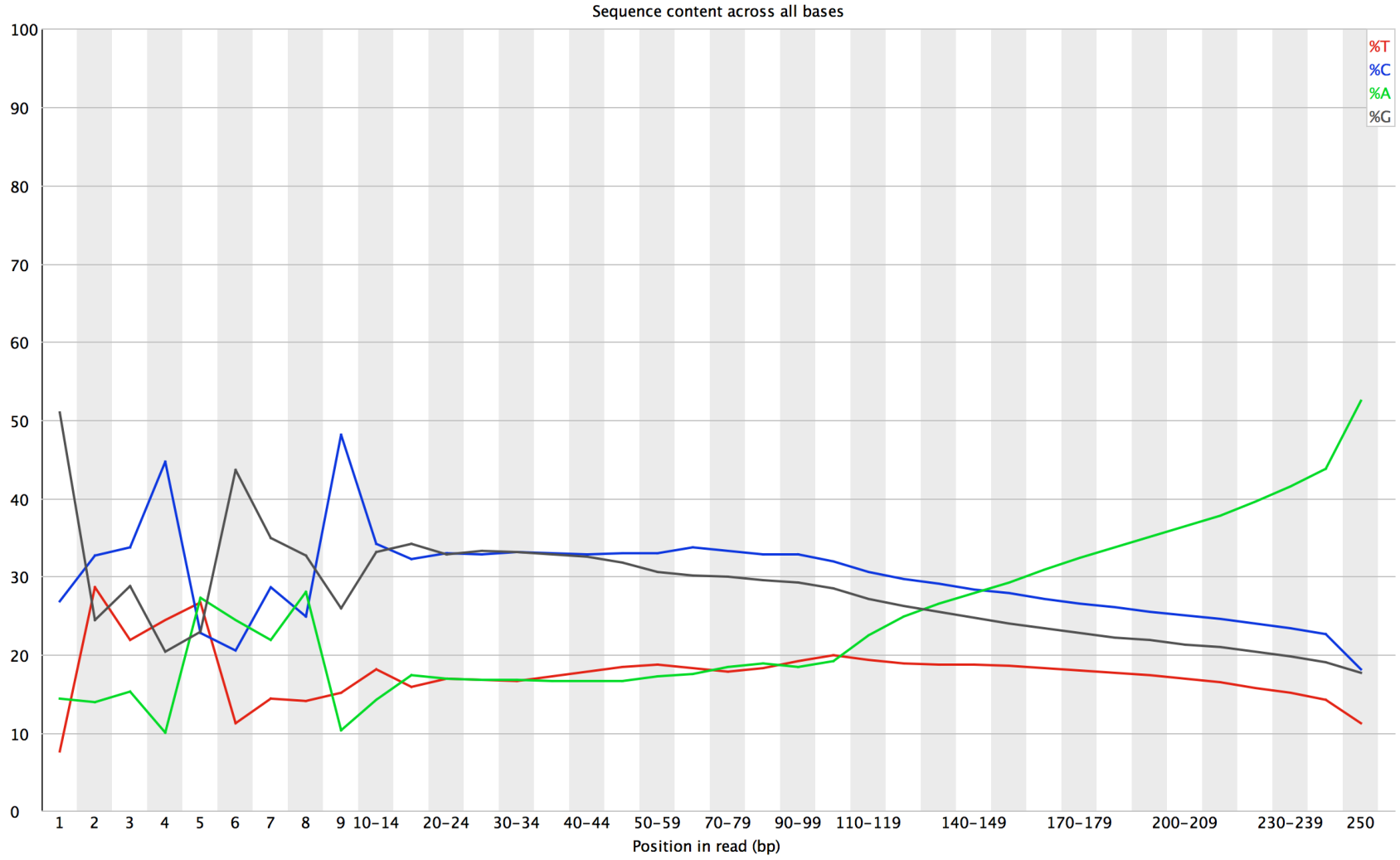
Tools for FASTQ manipulation

- 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
- ...

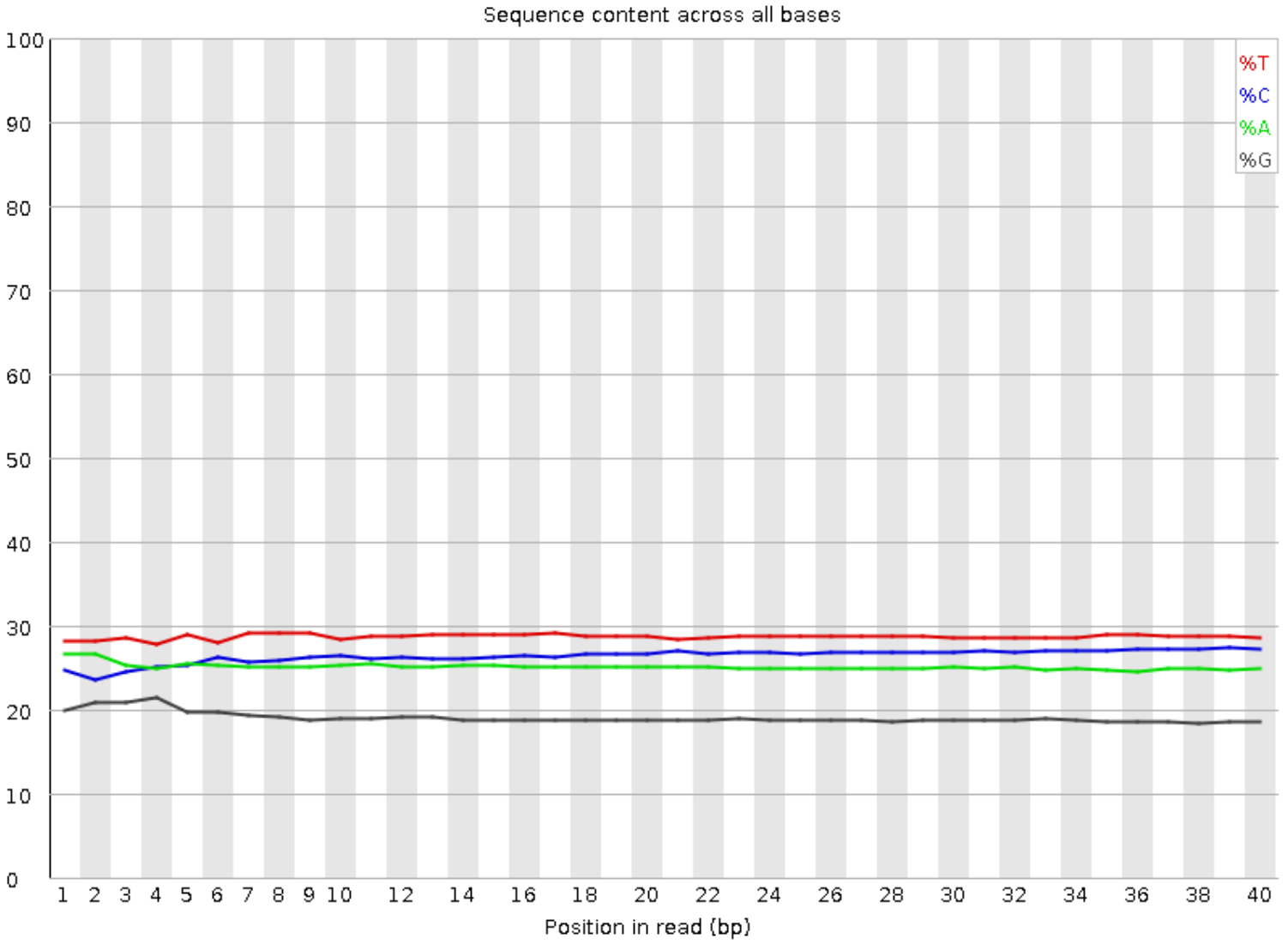
FASTQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

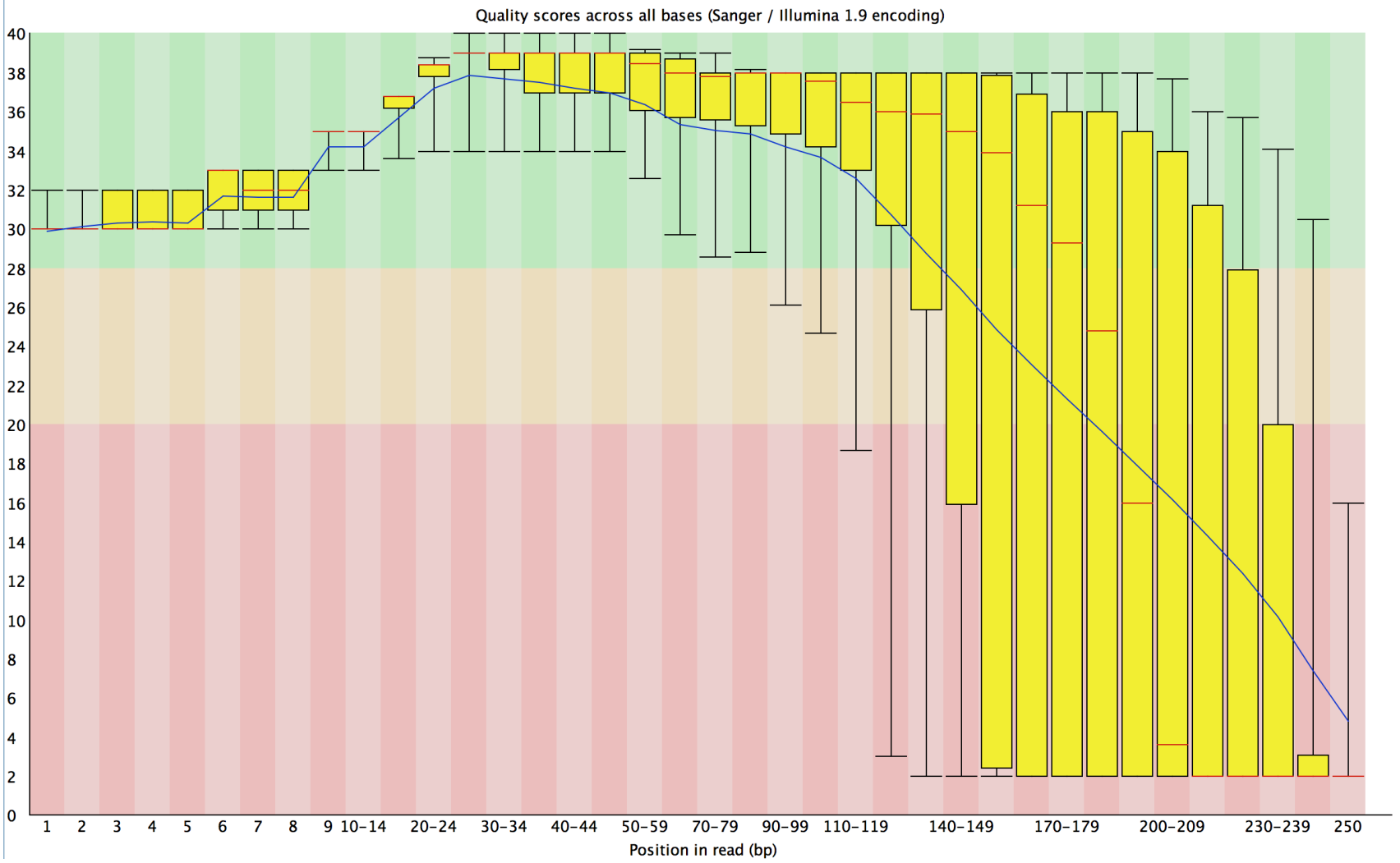
Sequencing problem



Good sequencing run



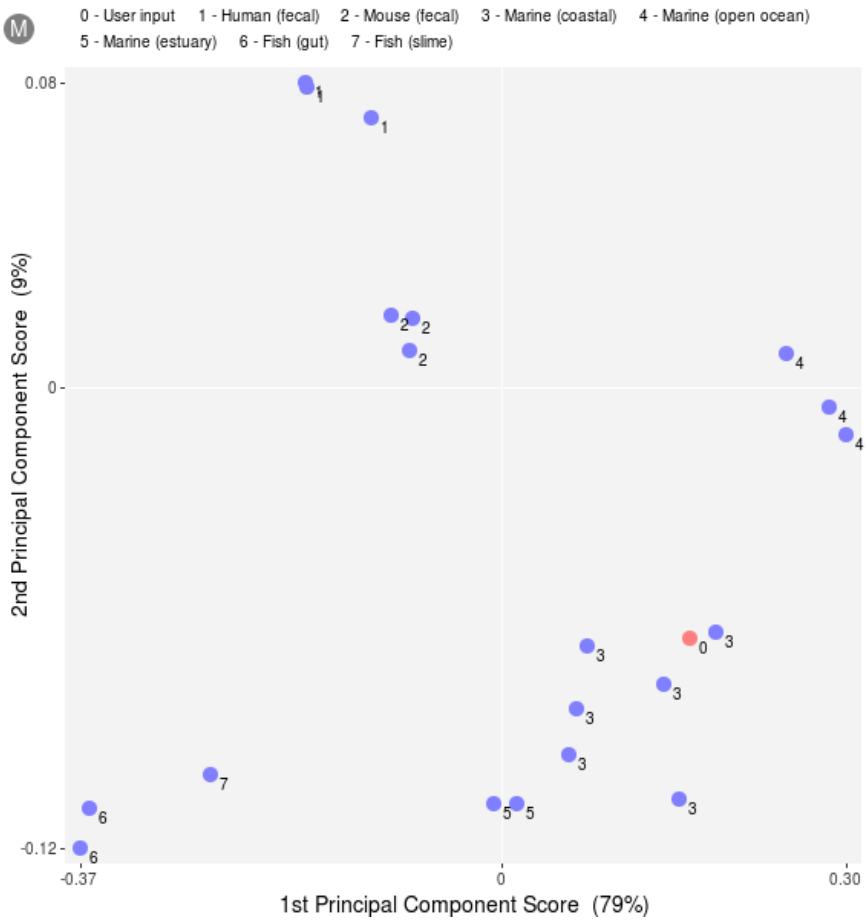
Sequencing problem



Check for contamination

PRINSEQ

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

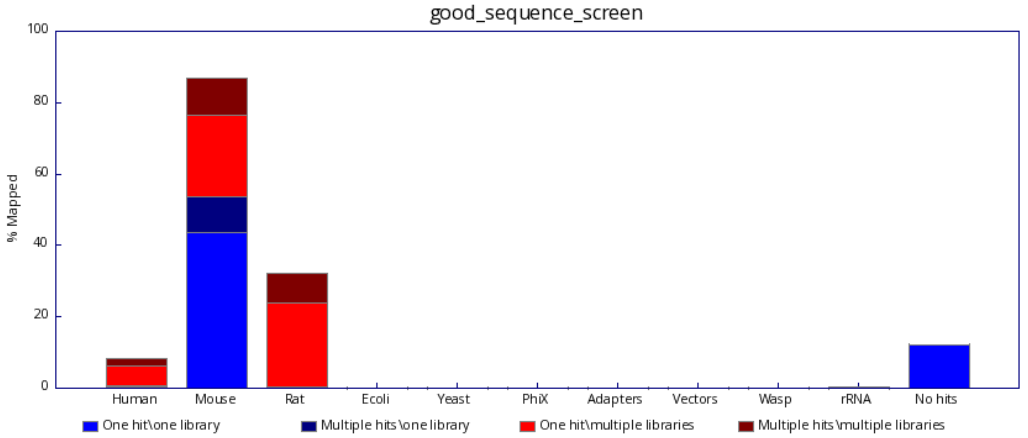


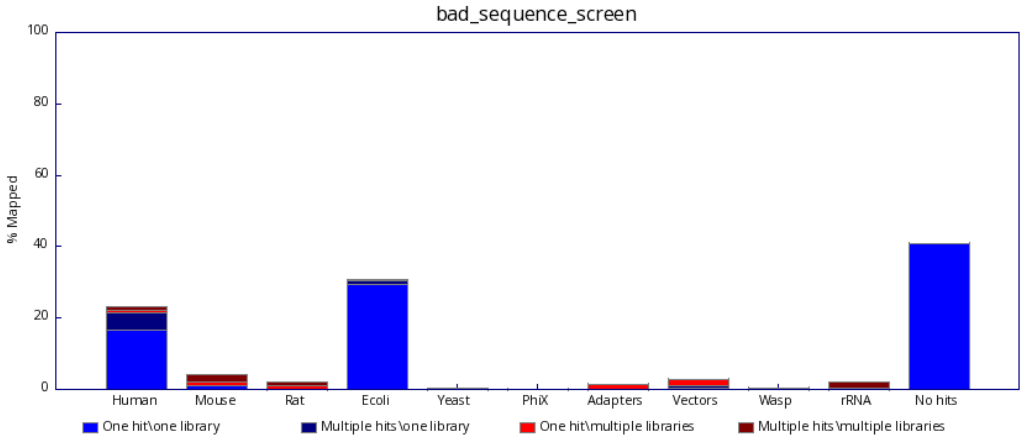
Check for contamination

FastQ Screen

Screen against a set of sequence databases:

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

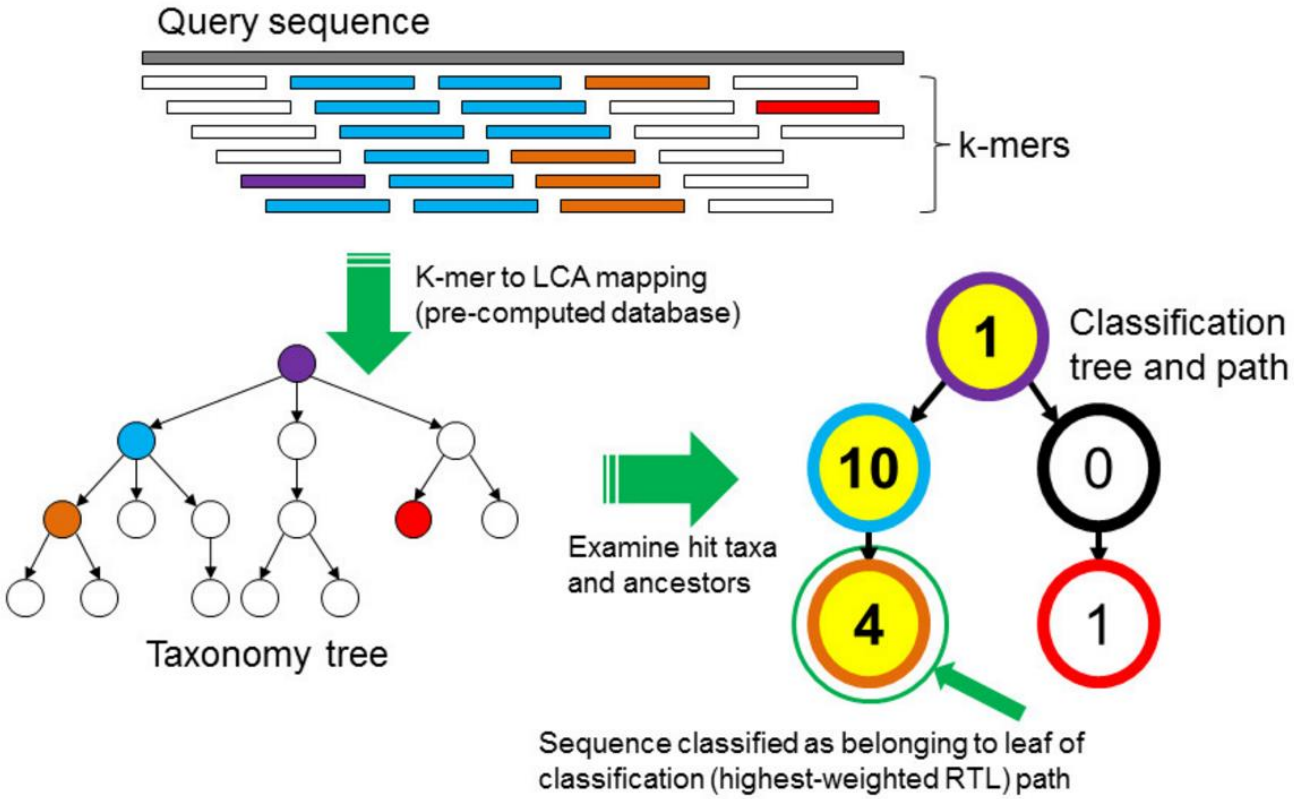




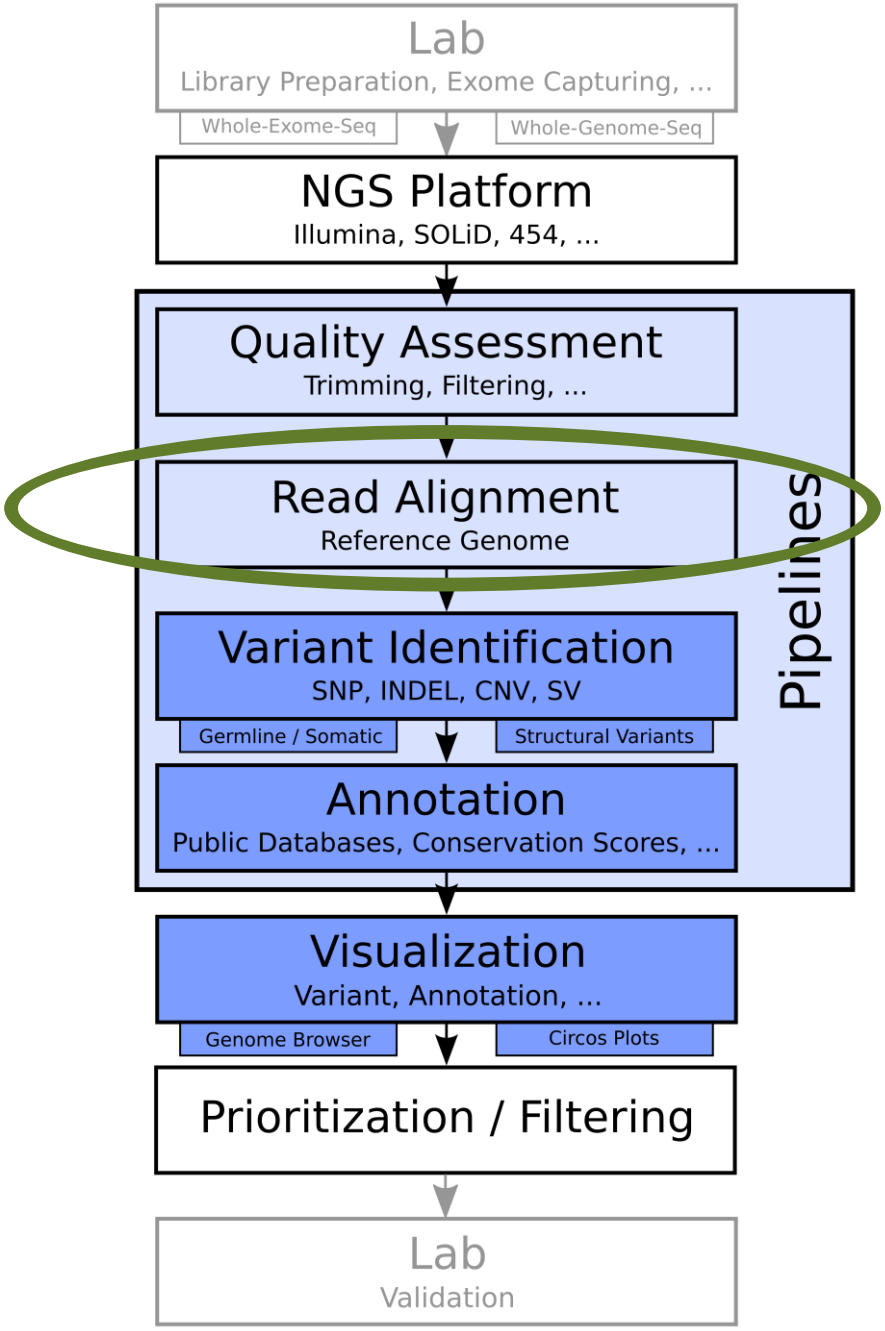
Check for contamination

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

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



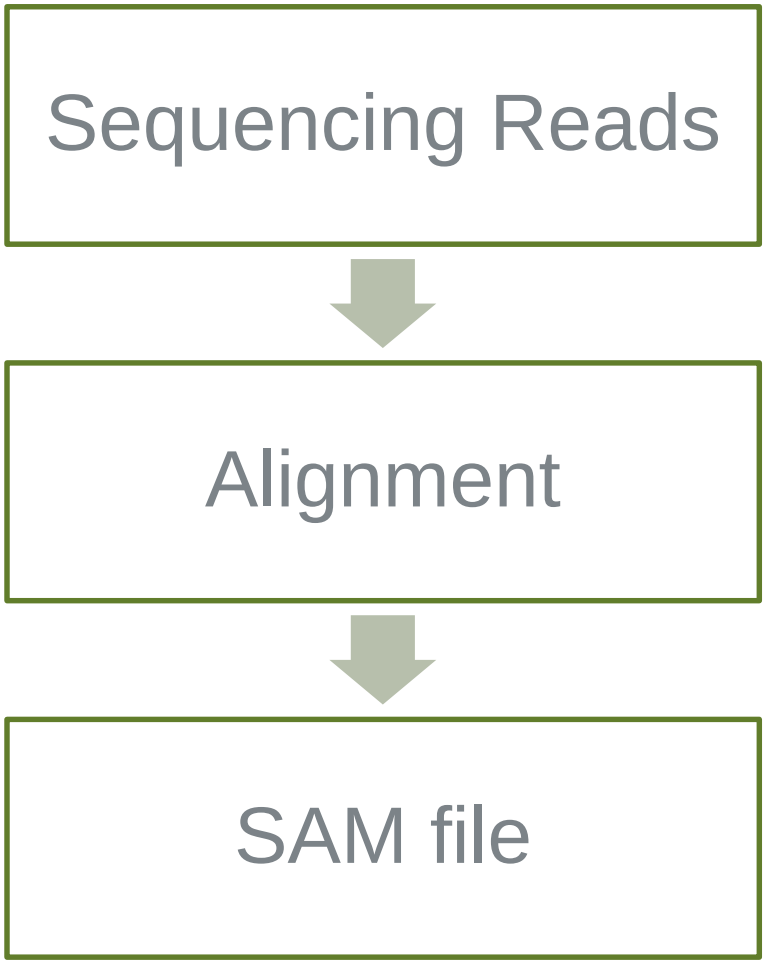
Mapping and QC


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

AUSTRIAN INSTITUTE OF TECHNOLOGY

SAM

Sequence Alignment/Map Format


Sequencing Reads
Alignment
SAM file

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

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

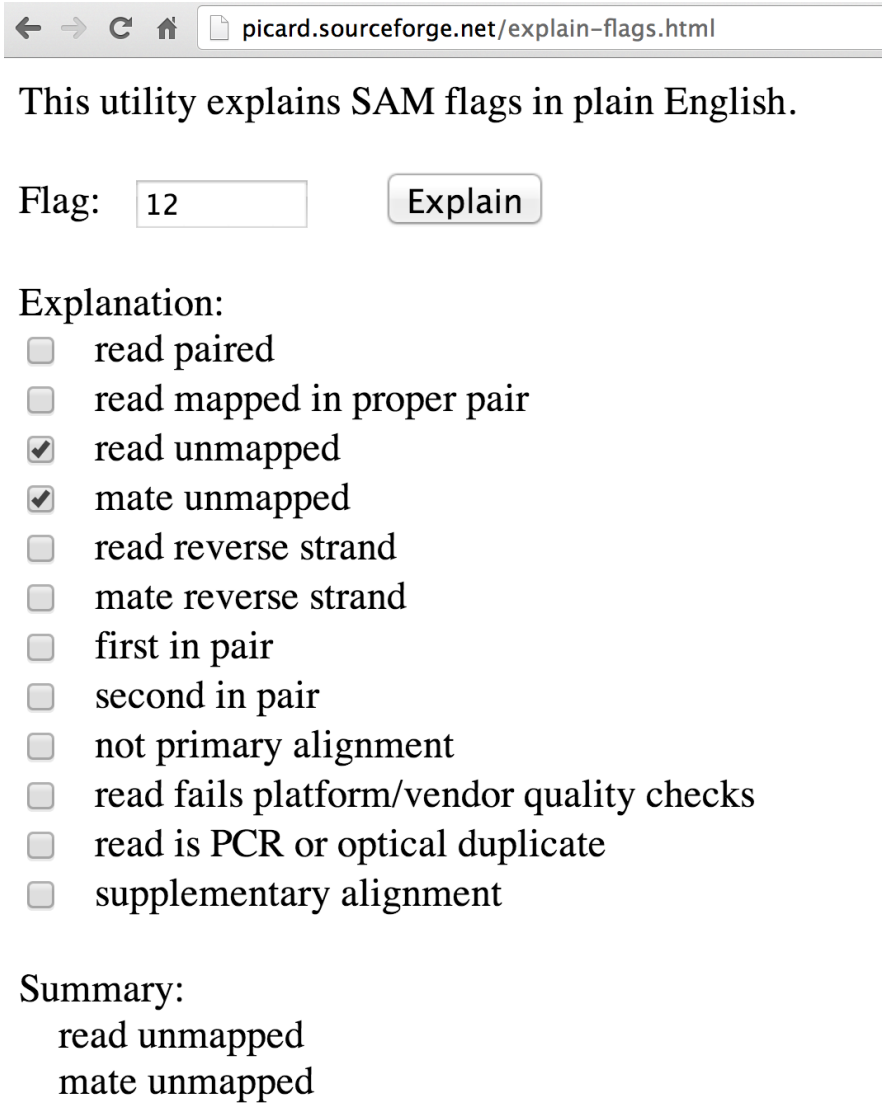
1 = 0001 -> PE read
4 = 0100 -> Unmappable
5 = 0101 -> Unmapped PE

2 FLAG

- Bitwise
- Picard (online) – explain flags

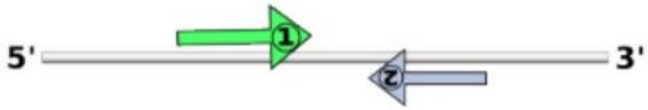
Examples

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



2 FLAG

Reads mapped in proper pair

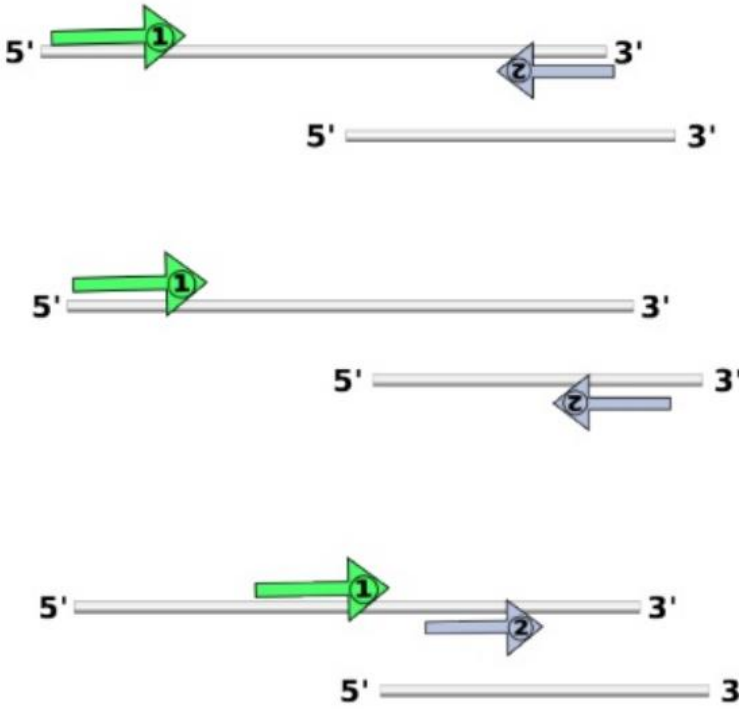


Flag: 67 Explain

Explanation:

- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [ ] mate reverse strand
- [x] first in pair

Reads not mapped in proper pair



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

Read `ACGC-TGCAGTTATATAGG`
Ref `ACTCAGTG--GT`
Cigar 4M1D3M2I2M7S

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

Read Group Tag (RG)

RG - Meta information

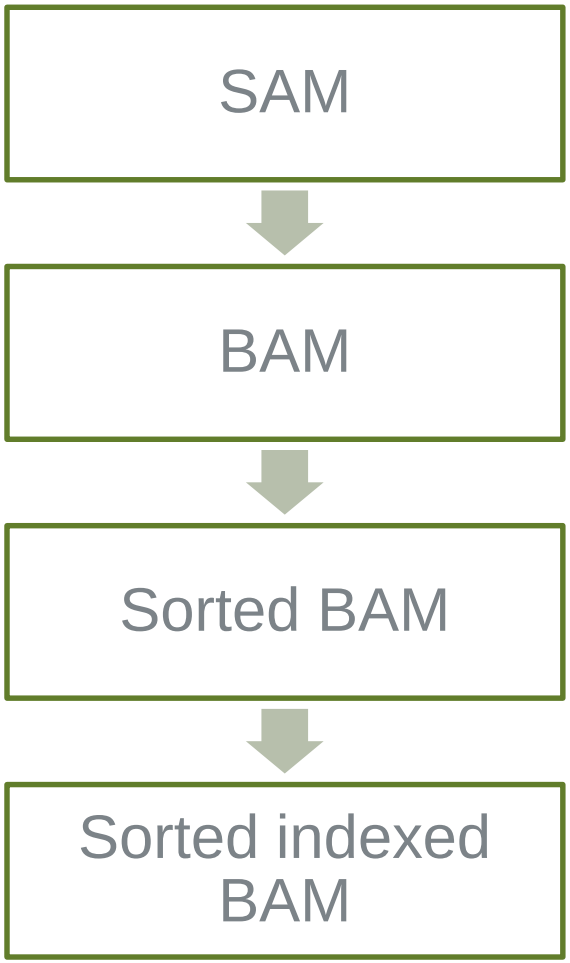
- ID: unique e.g., SRA number (Sequence read archive)
- PL: Sequencing platform
- PU: Platform unit (run name / flowcell-barcode-lane)
- LB: Library name
- PI: Insert size (Predicted mean insert size)
- SM: Sample (Individual)
- CN: Sequencing center

PICARD

```
java -jar picard/AddOrReplaceReadGroups.jar I=exampleWO_RG.bam O=
exampleW_RG.bam SORT_ORDER=coordinate RGID=superID RGLB=superLib
RGPL=illumina RGSM=superSample
```

BAM – Binary SAM

SAM / BAM



Convert (view)

sort

index

BAM

AUSTRIAN INSTITUTE OF TECHNOLOGY

SAM

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

BAM (Binary SAM)

- Compressed -> saves disk space
- Can be sorted & indexed - for quick viewing/searching
- 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

Quality check of alignment

Based on SAM/BAM files

Detect biases in the sequencing and/or mapping

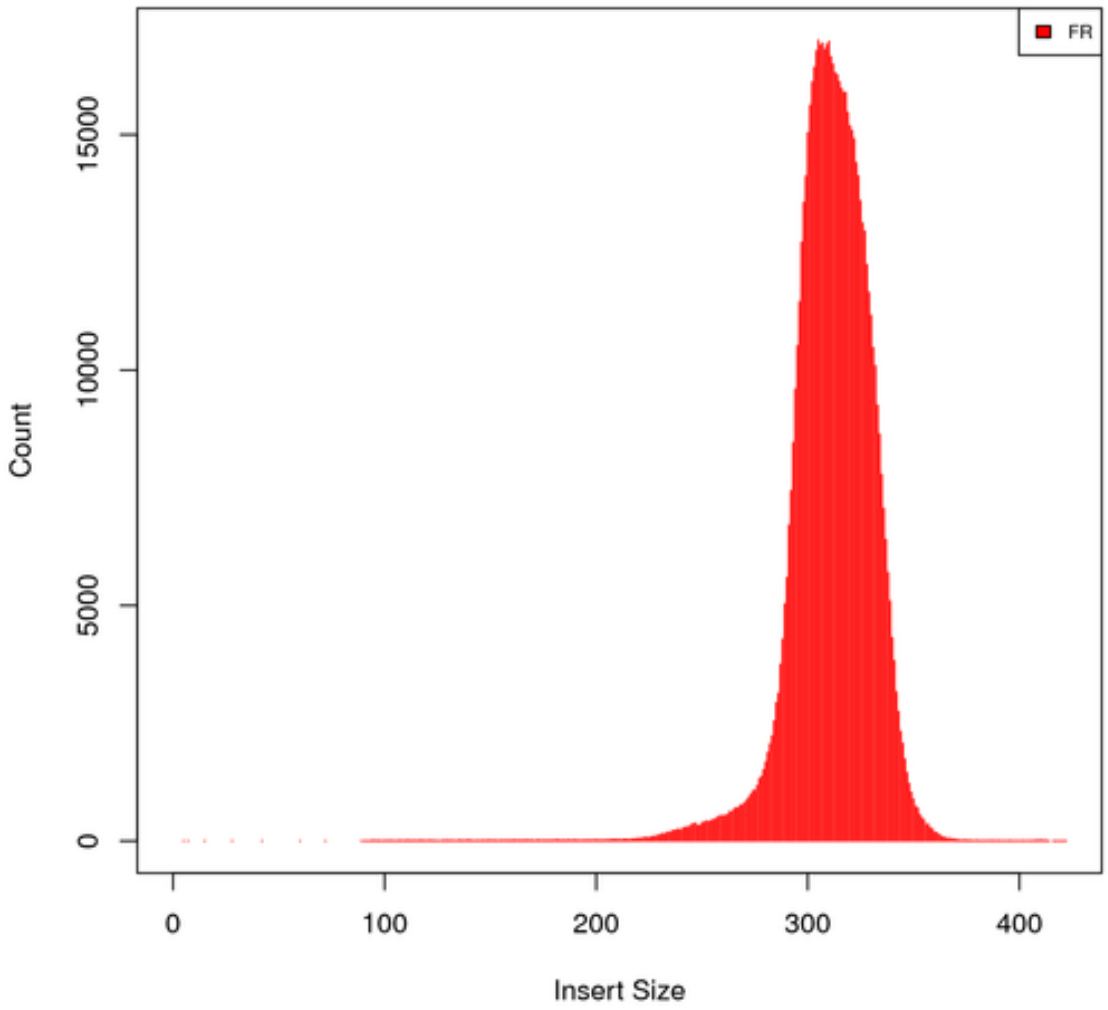
Metrics

- 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

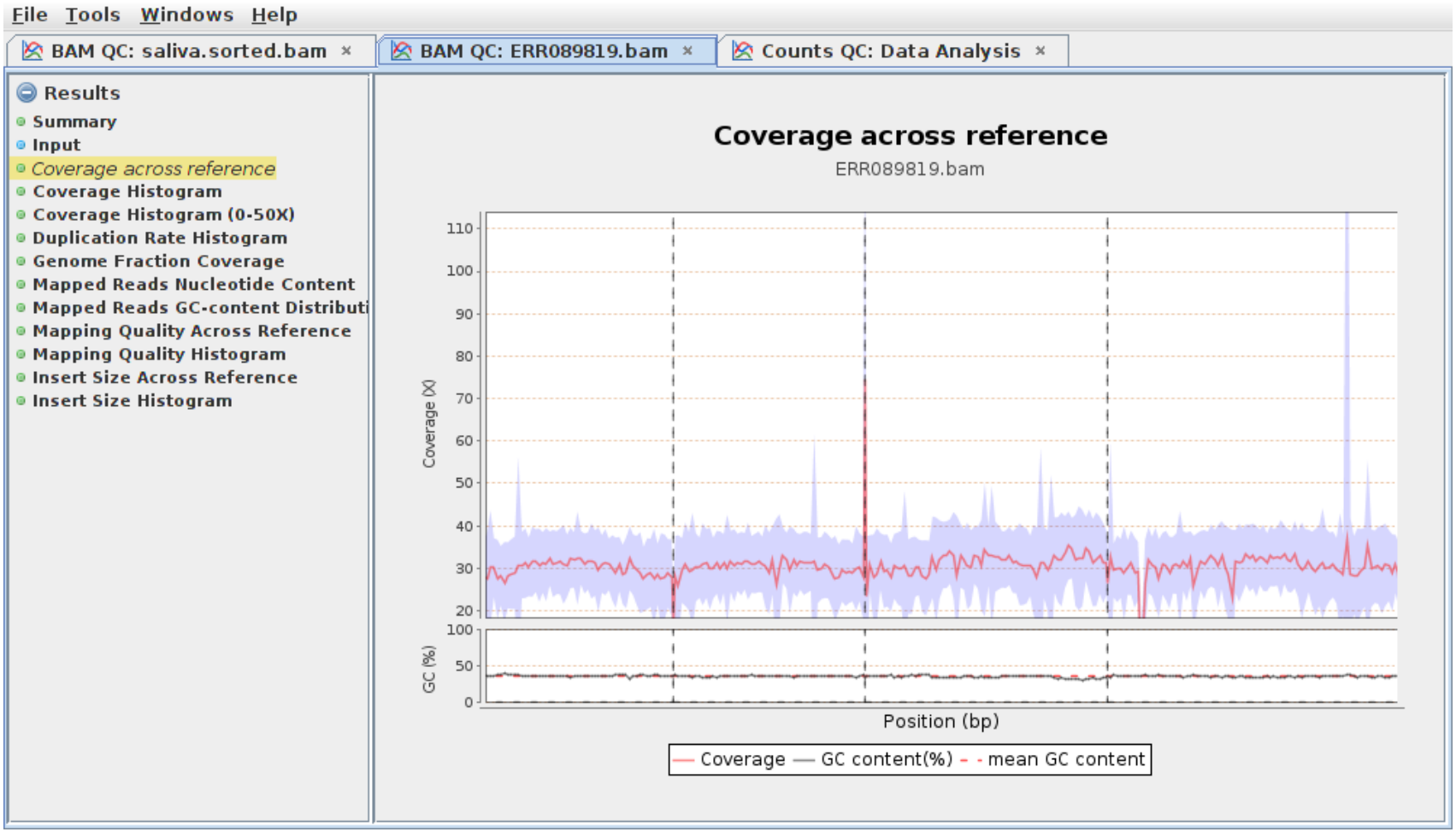
Tools

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

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

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

Problem

More steps during PCR amplification with little input material → more duplicates

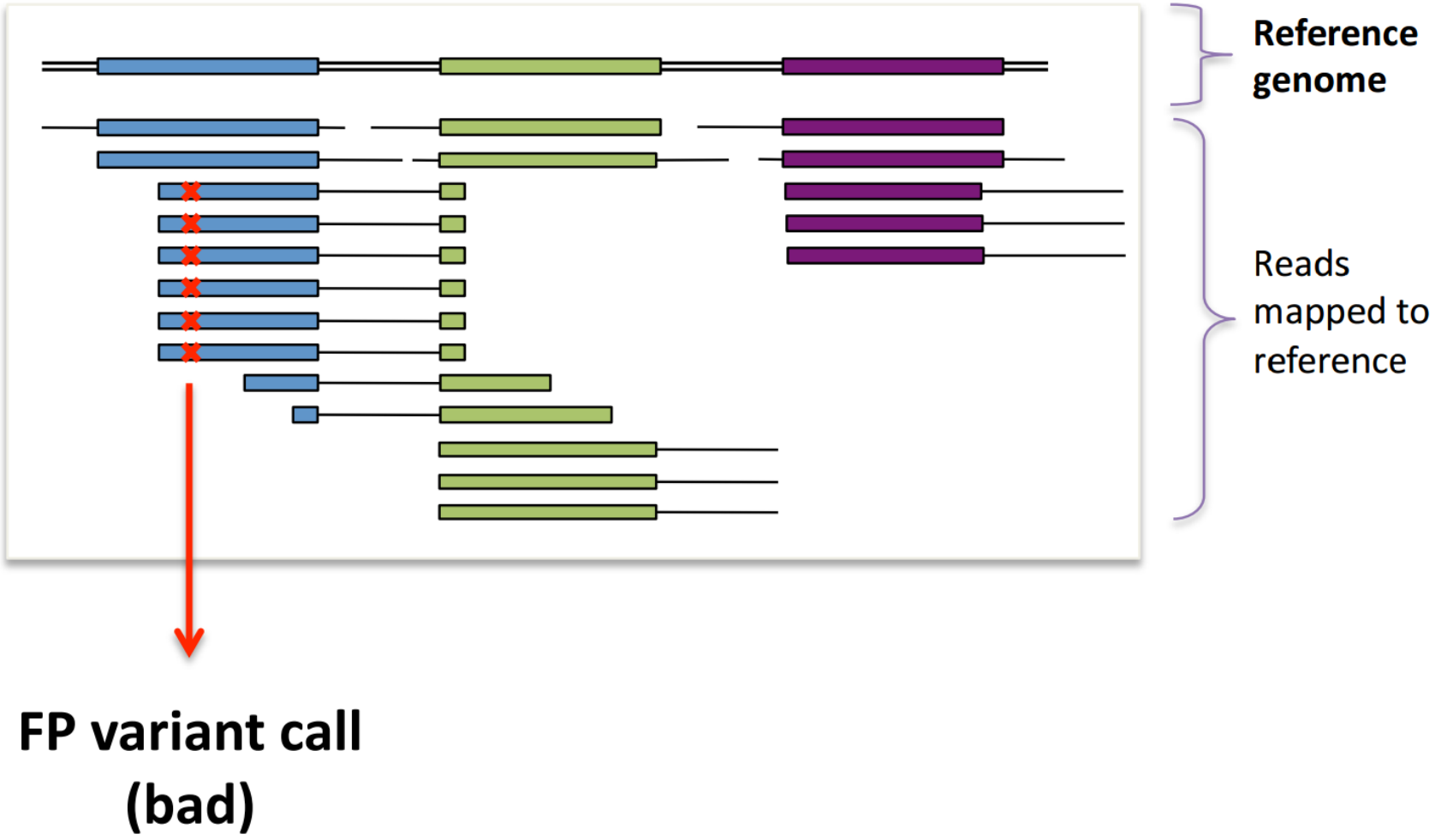
Read duplicates

Problems

- Can 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)

Info: http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/

Read duplicates

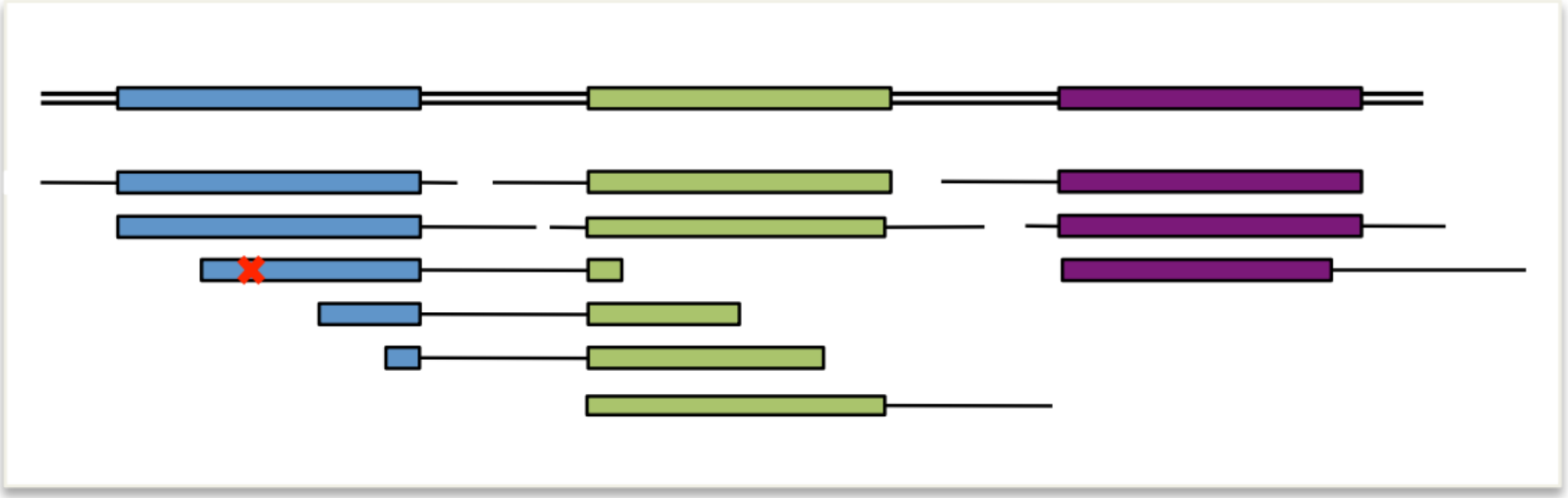


http://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-1-Map_and_Dedup.pdf

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

Would remove wrong 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

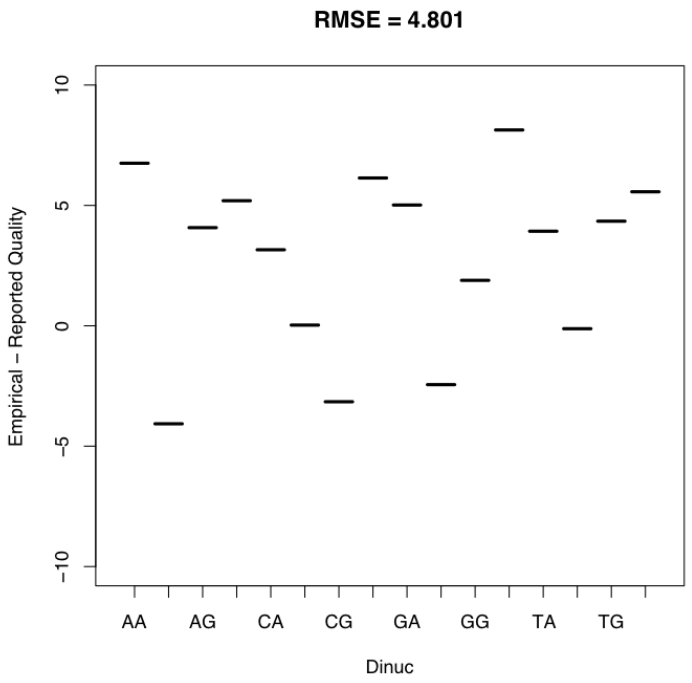
- For WES restrict to capture targets
  off-target sites are likely to have higher error rates

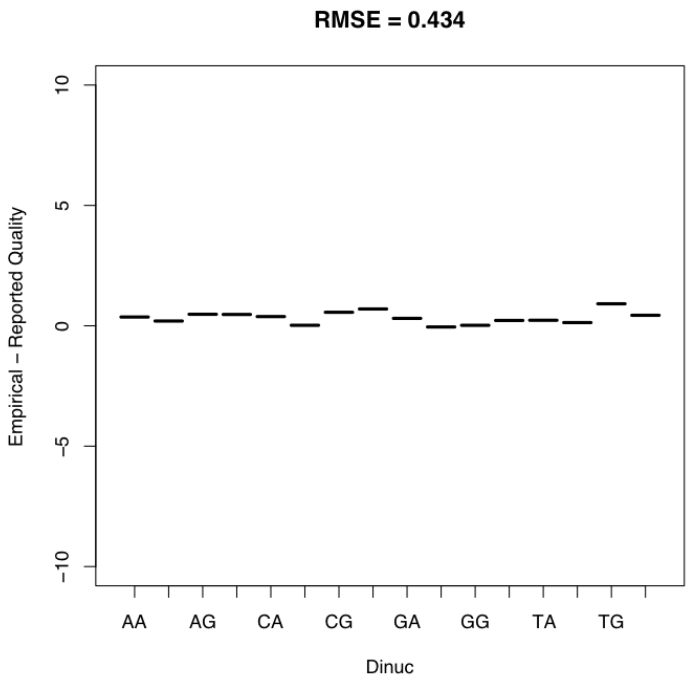
Organism with no known variants?

- 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

Residual Error by Dinucleotide



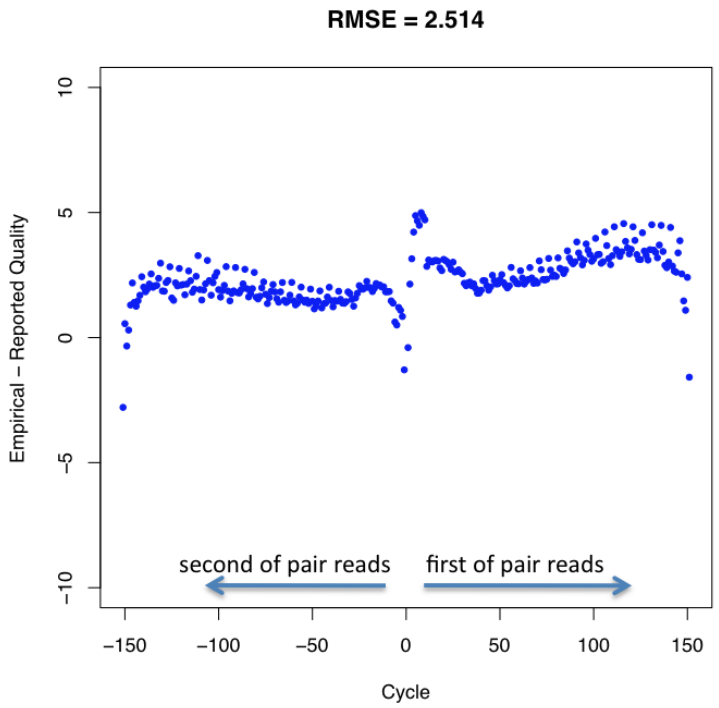


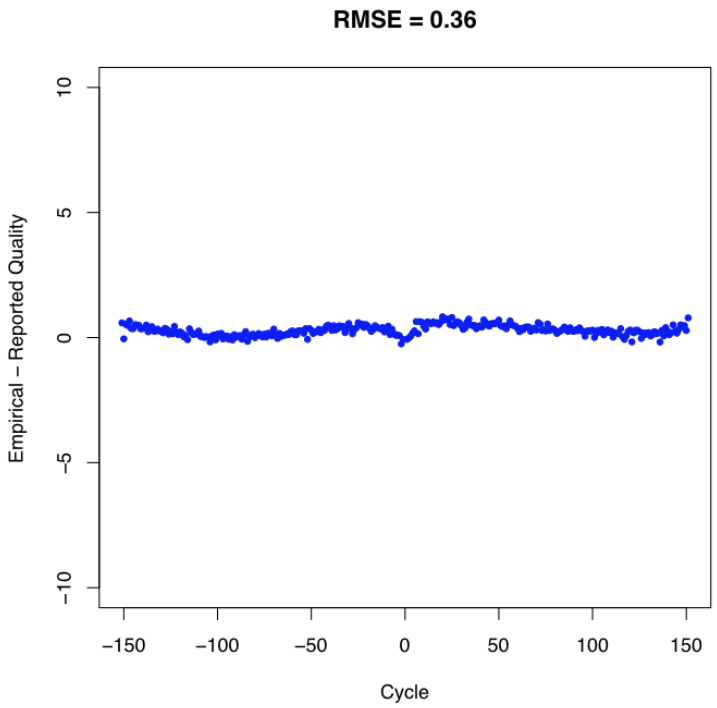
Original Data

After GATK Recalibration

RMSE = root-mean-square error

Residual Error by Machine Cycle





Original Data

After GATK Recalibration

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 <aln.bam> sorted
```

Index

```
samtools index <aln.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 BAM to SAM

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

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

```
samtools view -bSh <file.sam> | samtools sort - file_sorted
```

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

SAMBAMBA

„Multithreaded” SAMtools

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

- Binaries available

https://github.com/lomereiter/sambamba

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

https://github.com/exascience/elprep

PICARD

JAVA based tool

http://picard.sourceforge.net/

Tools

- BuildBamIndex
- CollectAlignmentSummaryMetrics
- FastqToSam
- MergeSamFiles
- MergeVcfs
- ...

AUSTRIAN INSTITUTE OF TECHNOLOGY

Variant Calling

What kind of data do you have?

Genotyping

- Single samples
- Family → finding causing mutation of rare disease
- Time series

Somatic mutations

- Primary tumor tissue
- Blood samples

→ Different callers / strategies for respective samples

What are variant calls?

Find differences to a reference (hg19, GRCh38)

Naive variant calling

- Check all the reads that cover base chr11:1234567
- Add up the bases at chr11:1234567
- e.g. 15 A's, 4 G's
- Is this an A/G heterozygous site or four sequencing errors?

Actual variant callers

- Estimate likelihood of a variant site vs a sequencing error
- Sequencing error rate
- Quality scores

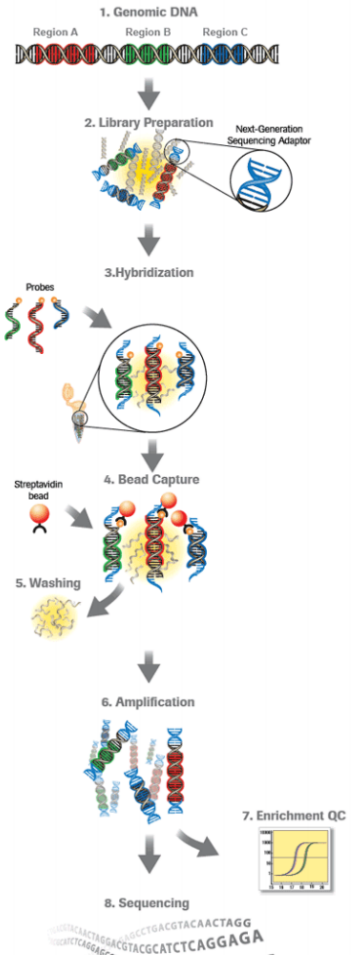
Possible reasons for a mismatch

- 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

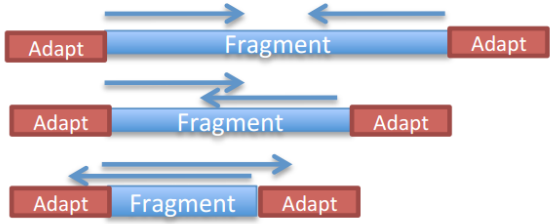
Difficulties



Sonication

Library prep (sequencing adaptors on)

Problem: **error in sonication** >> adaptor seq in reads >> **unmapped reads**



Hybridisation to probes → Possible biases in sequences that hybridise >> **coverage bias**

Bead capture → Possible biases in sequences that elute >> **coverage bias**

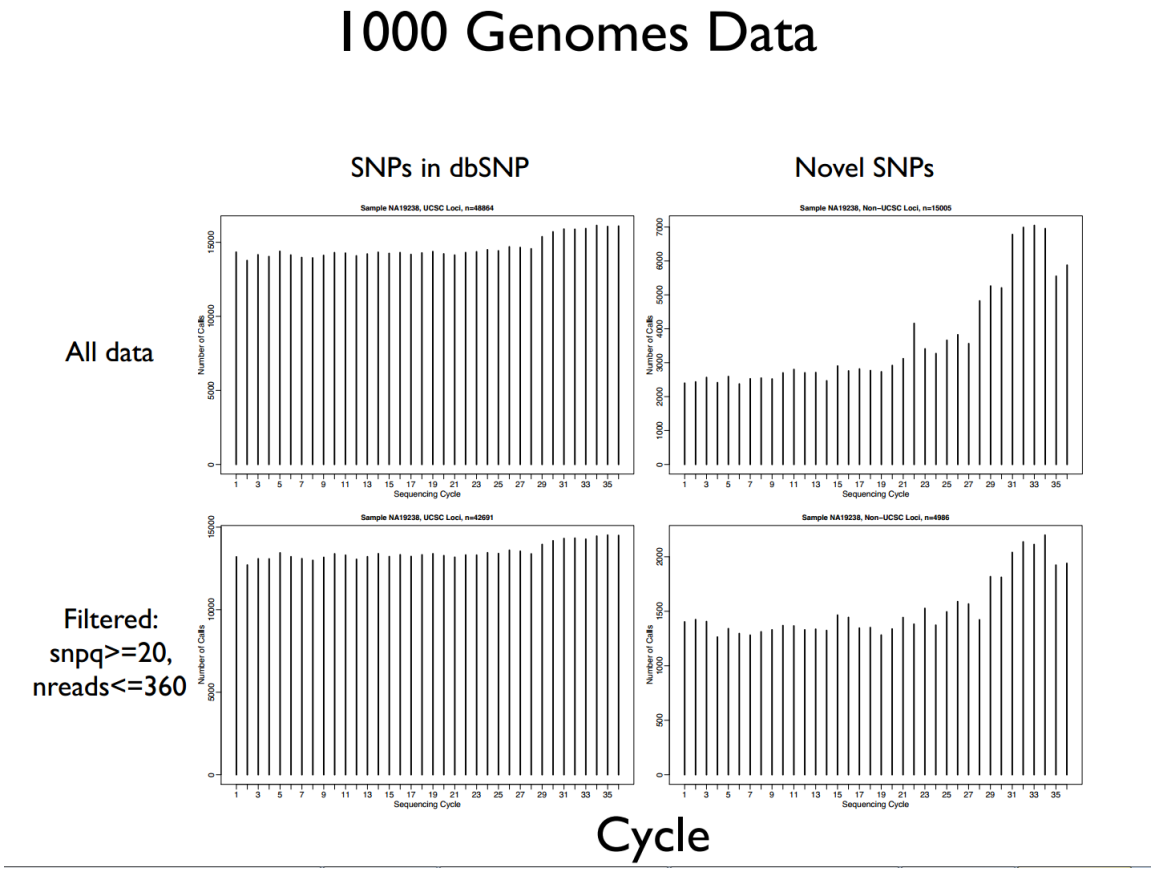
Amplification → Possible biases in sequences that amplify **>> sequence PCR duplicates**

Sequencing → Possible biases in sequences that bridge PCR >> **coverage bias**

https://wiki.uio.no/projects/clsi/images/8/80/VariantCallingCourse_sept2013_Part1.pdf

Difficulties

- Base calling gets more difficult the longer the read gets



http://www.biostat.jhsph.edu/~khansen/LecIntro1.pdf

Difficulties

- High depth - low quality regions → likely due to copy number or other larger structural events
- Repetitive regions → artifactual 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

SNV calling

Consider

- Base quality
- Proximity to Indel
- Homopolymer run
- Mapping qualities of the reads supporting the SNP
- Read length
- Paired reads
- Sequencing depth

Greatest challenge

Misalignments → incorrect SNP calls

Indel calling

In general – call Indels based on the I and D events in BAM file

Consider

- Misalignment of the read
- Homopolymer runs
- Length of reads
- Zygosity

Approach to remove FP

- Create new haplotype (new reference) and realign the reads to this ref
- Count number of reads supporting this new haplotype
- → computationally extensive

Somatic mutations

- Change in genetic structure **not** inherited from a parent
- Constantly happening in living organisms
- Can be used for detection of disease

- Detected by comparison to „normal“ cells

Cancer

- Used to characterize tumor cells

SV calling

Use paired end information to detect these events

- Deviations of the expected insert size
- Presence/absence of mate pairs
- Read depth for CNVs

AUSTRIAN INSTITUTE OF TECHNOLOGY

Genotype variant calling

Variant callers

Bayesian genotype likelihood 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

Ti/Tv ratio

Transition (Ti)

- purine <-> purine (A <-> G)
- pyrimidine <-> pyrimidine (C <-> T)

Transversion (Tv) purine <-> pyrimidine

A <->C, A <->T, G <->C, G <->T

Transition is more frequent than transversion

- Ti/Tv ~ 2.0 - 2.1 for genome wide
- Ti/Tv ~ 3.0 - 3.3 for exonic variations
- Ti/Tv = 2/4 = 0.5 = 2/4 =0.5 for random, uniform sequencing - sequencing error

Variant Calling Quality

Factors

- Coverage at position (DP)
- Number of reads supporting the call
- Strand bias
- Base qualities at variant position

AUSTRIAN INSTITUTE OF TECHNOLOGY

Practical 1

https://github.com/tadKeys/BioinformaticsAndGenomeAnalyses2016

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Tools for variant analysis of next-generation genome sequencing

Part 2

Stephan Pabinger
stephan.pabinger@ait.ac.at

Realignment around indels

Why?

- Alignments tend to accumulate FP SNPs near true indels
- SNPs are often less penalized compared to indels

Realignment principles

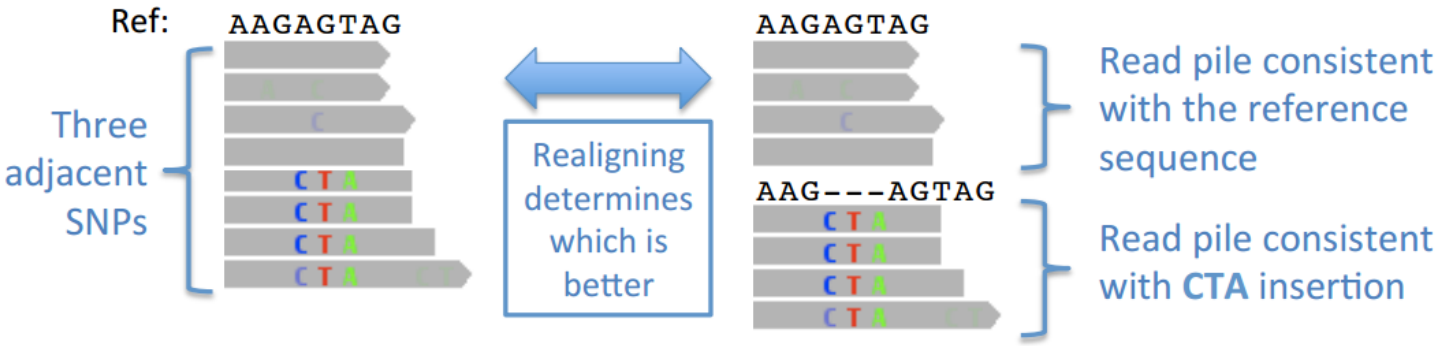
- Realign locally around indels → GATK

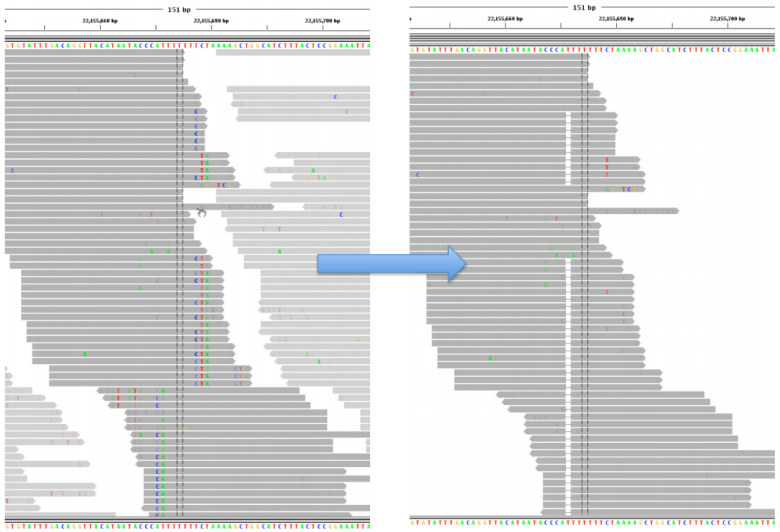
- Input set of known indel sites (dbSNP, 1000genomes)
- Indels from alignments
- Presence of mismatches & softclips (BAM file)

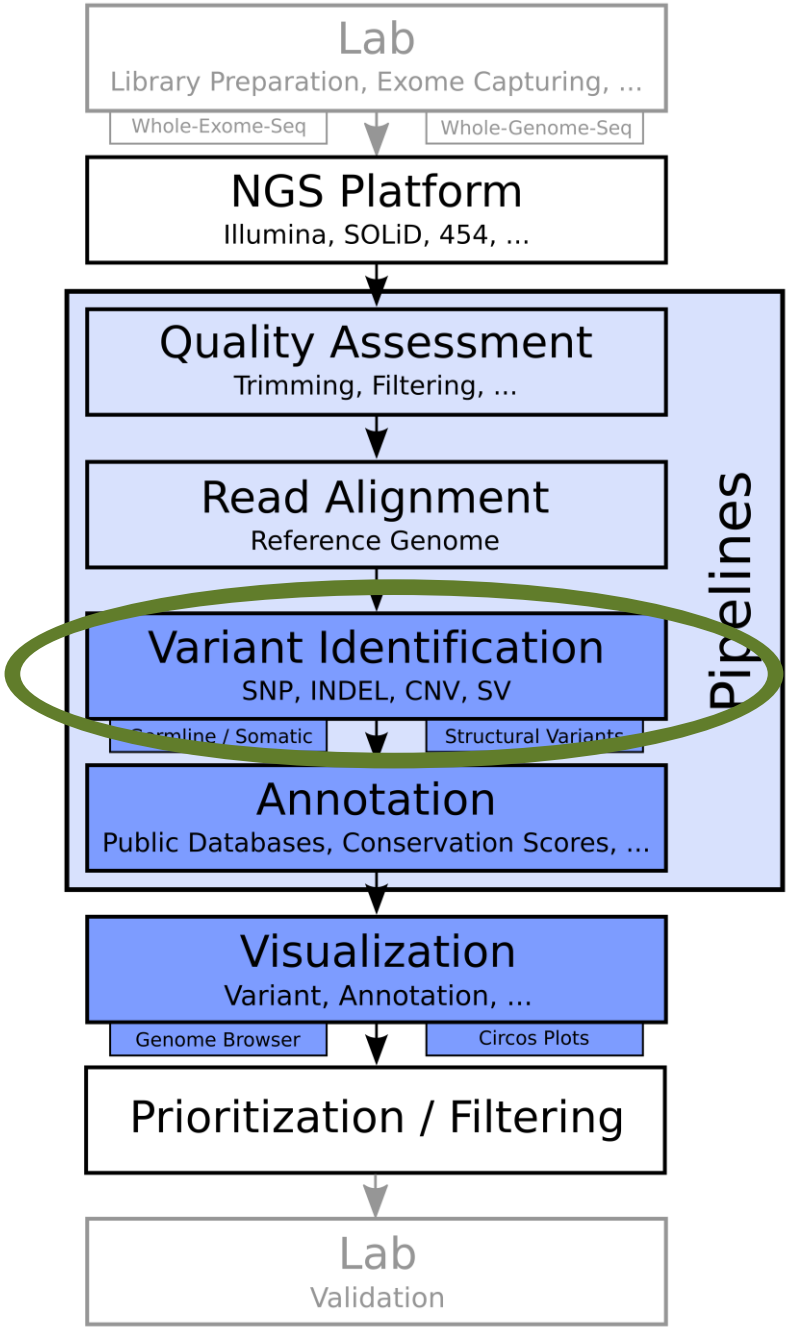
Realignment around indels

Realignment

- Model the indel haplotype
- → if score of alternative consensus is better than original use realigned





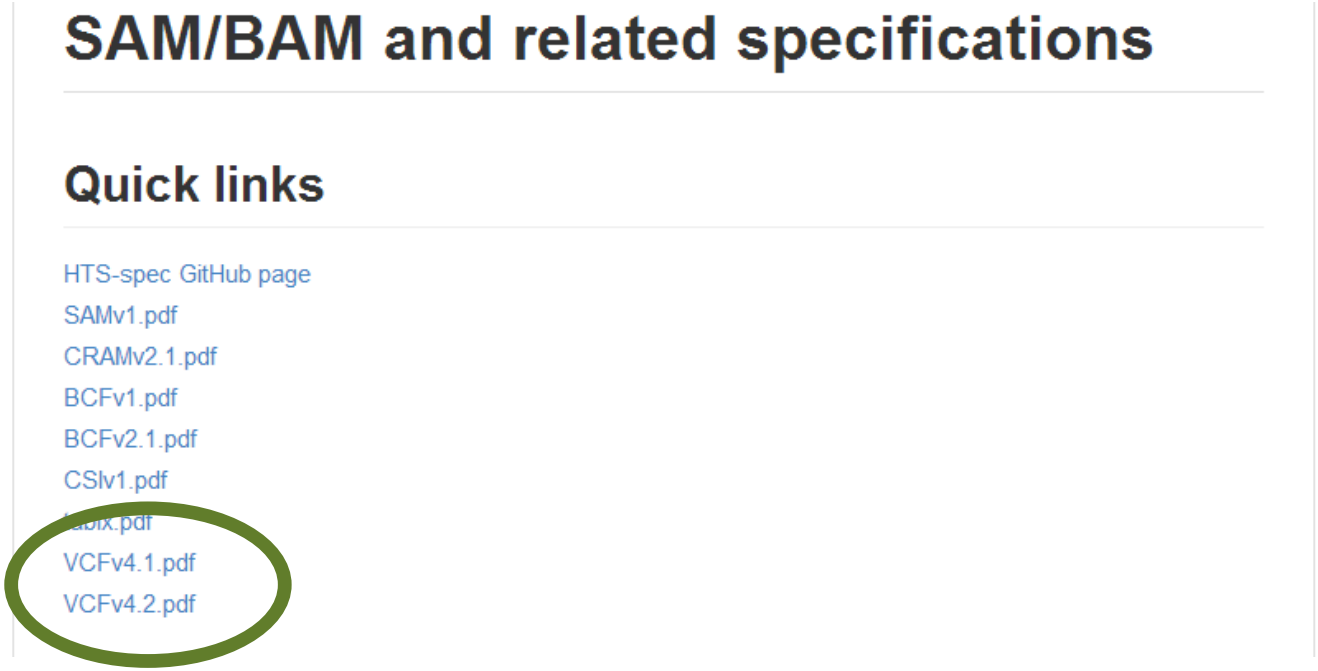

Lab
Library Preparation, Exome Capturing, ...
Whole-Exome-Seq
Whole-Genome-Seq
NGS Platform
Illumina, SOLiD, 454, ...
Quality Assessment
Trimming, Filtering, ...
Read Alignment
Reference Genome
Variant Identification
SNP, INDEL, CNV, SV
Structural Variants
Annotation
Public Databases, Conservation Scores,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

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 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

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

...

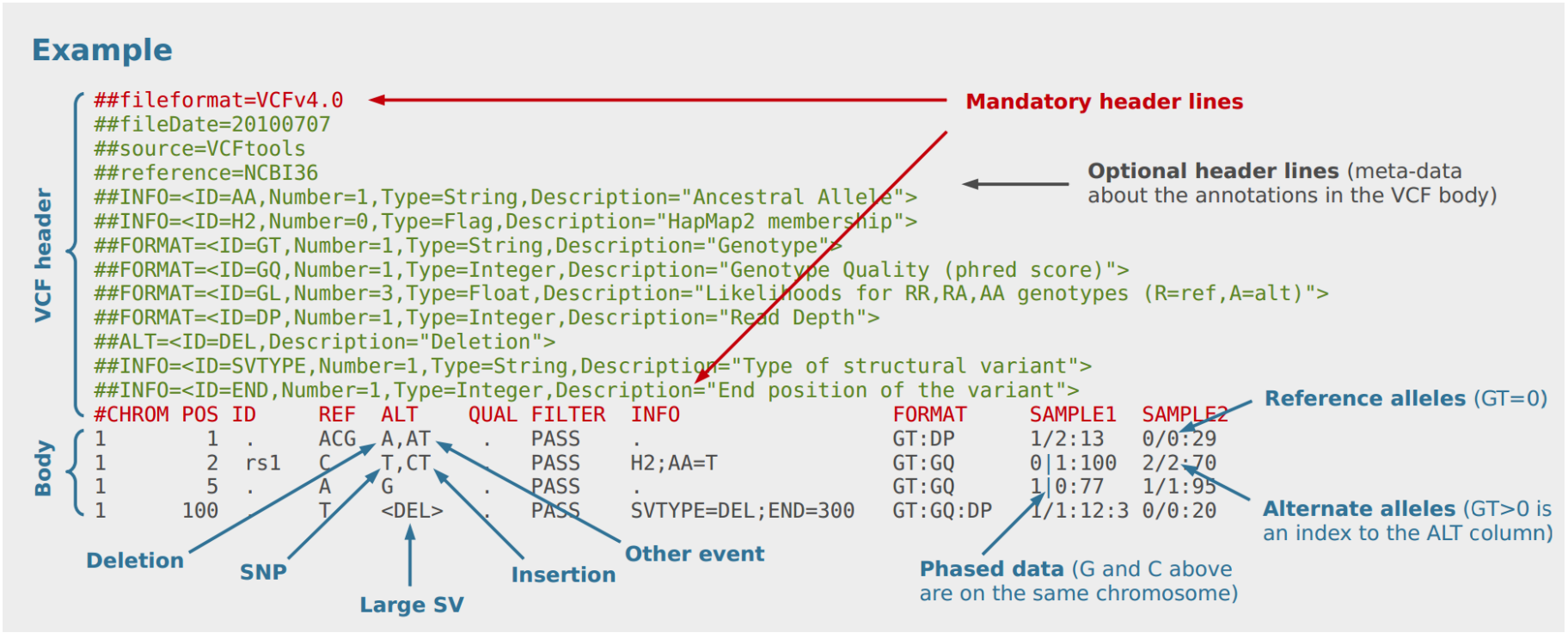
Genotype field

- 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



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

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?

AUSTRIAN INSTITUTE OF TECHNOLOGY

Types of variants

SNPs

Alignment
```
ACGT
ATGT
```

VCF representation

POS	REF	ALT
2	C	T

Insertions

Alignment
```
AC-GT
ACTGT
```

VCF representation

POS	REF	ALT
2	C	CT

Deletions

Alignment
```
ACGT
A--T
```

VCF representation

POS	REF	ALT
1	ACG	A

Complex events

Alignment
```
ACGT
A-TT
```

VCF representation

POS	REF	ALT
1	ACG	AT

Large structural variants

VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=300

Deletion in VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	2	.	TCG	T	.	PASS	DP=100

This is a deletion of two reference bases since the reference allele TCG is being replaced by just the T [the reference base]. Again there are only two alleles so I have the two following segregating haplotypes:

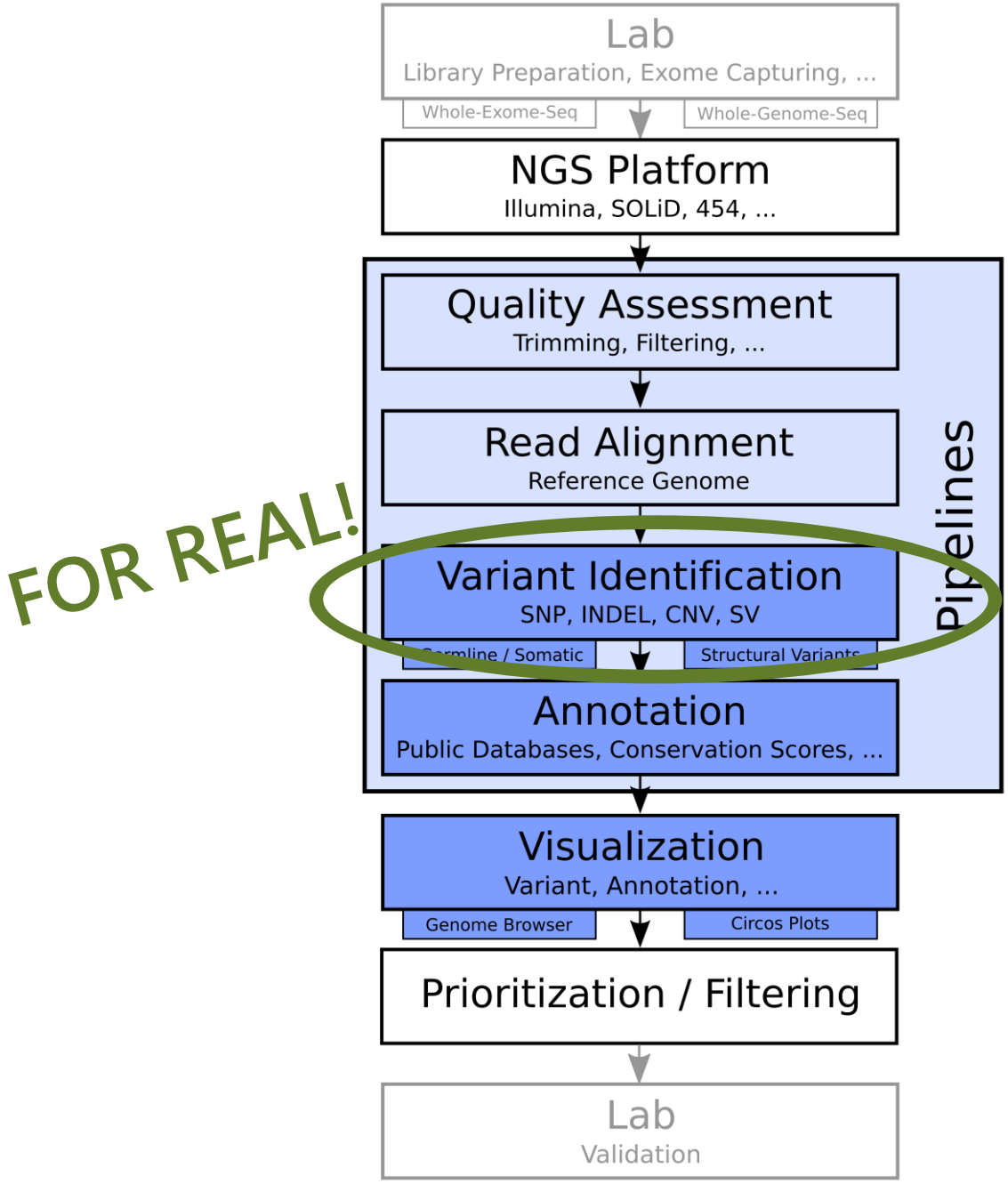
Example	Sequence	Alteration
Ref	`a T C G a`	T is the (first) reference base
1	`a T - - a`	following the T base is a deletion of 2 bases

Insertion in VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	3	.	C	CTAG	.	PASS	DP=100

This is a insertion since the reference base C is being replaced by C [the reference base] plus three insertion bases TAG. Again there are only two alleles so I have the two following segregating haplotypes:

Example	Sequence	Alteration
Ref	a t C - - - g a	C is the reference base
1	a t C T A G g a	following the C base is an insertion of 3 bases


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
Structural Variants
Annotation
Public Databases, Conservation Scores, ...
Pipelines
Visualization
Variant, Annotation, ...
Genome Browser
Circos Plots
Prioritization / Filtering
Lab
Validation
FOR REAL!

SAMtools

Variant calling

SAMtools variant calling

AUSTRIAN INSTITUTE OF TECHNOLOGY

Command

```
samtools mpileup -uf hg19.fasta deduprg.bam | bcftools call
-c -v -o samtools.vcf
```

Filter

```
bcftools/vcfutils.pl varFilter -Q 20 -d 10 -D 200
hs37d5_allseqs_bwa.raw.vcf
```

quality 20, read depth > 10; read depth < 200

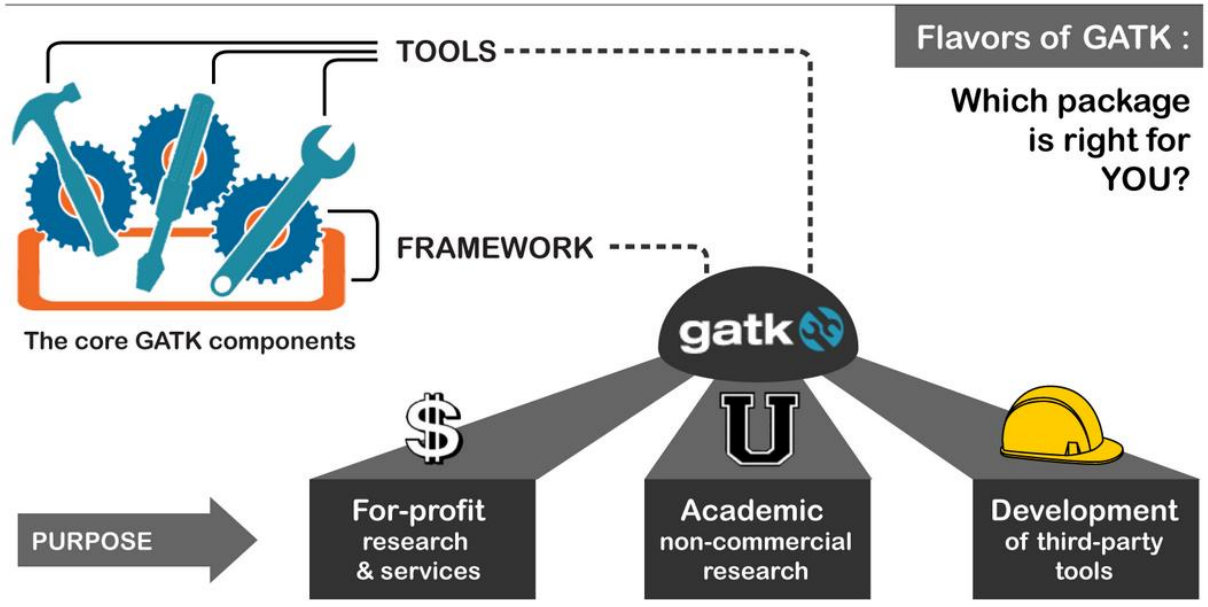
http://ged.msu.edu/angus/tutorials-2013/snp_tutorial.html

GATK - Genome Analysis Toolkit

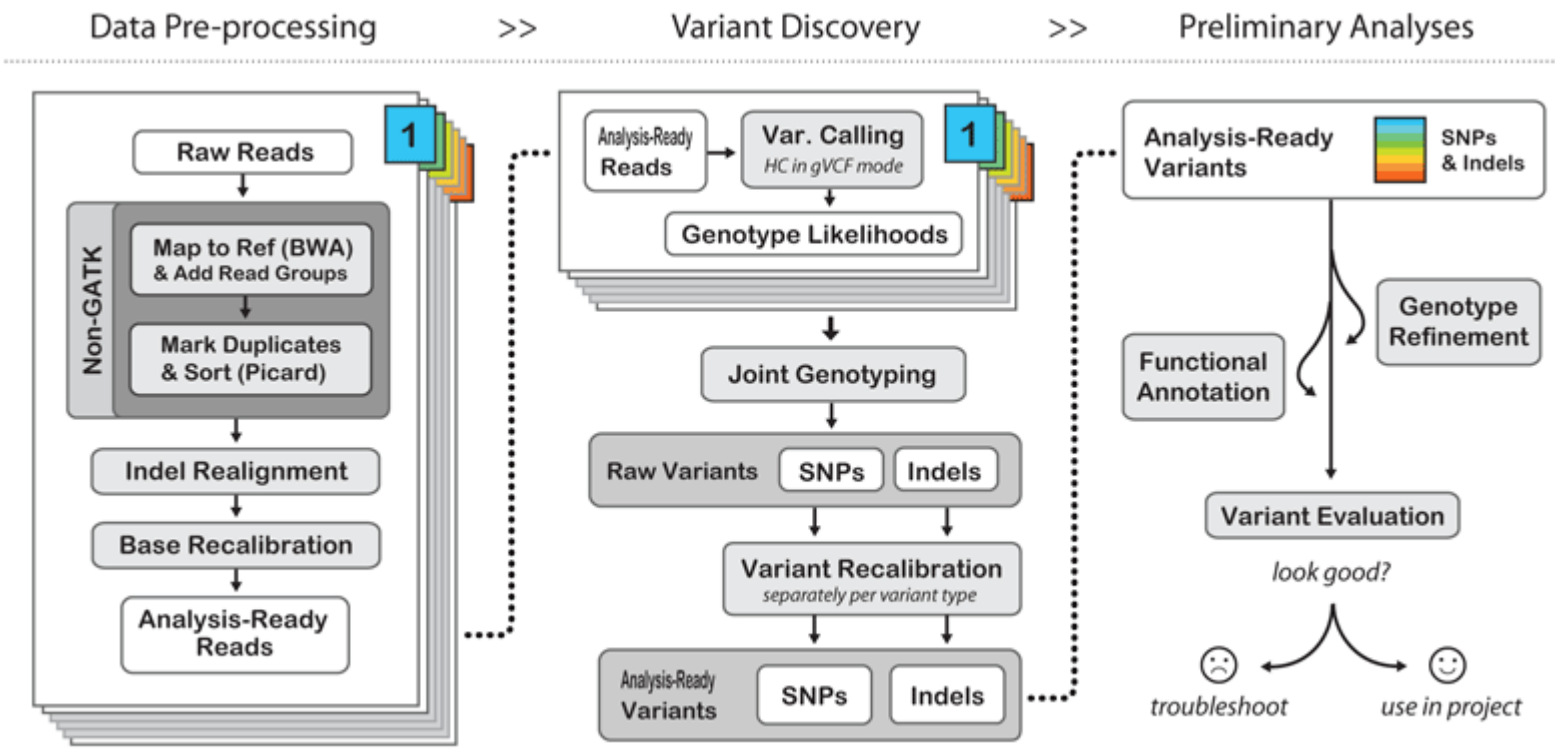
Variant calling pipeline

GATK

- JAVA, command line software
- Linux (Mac)
- Mixed closed/open-source model



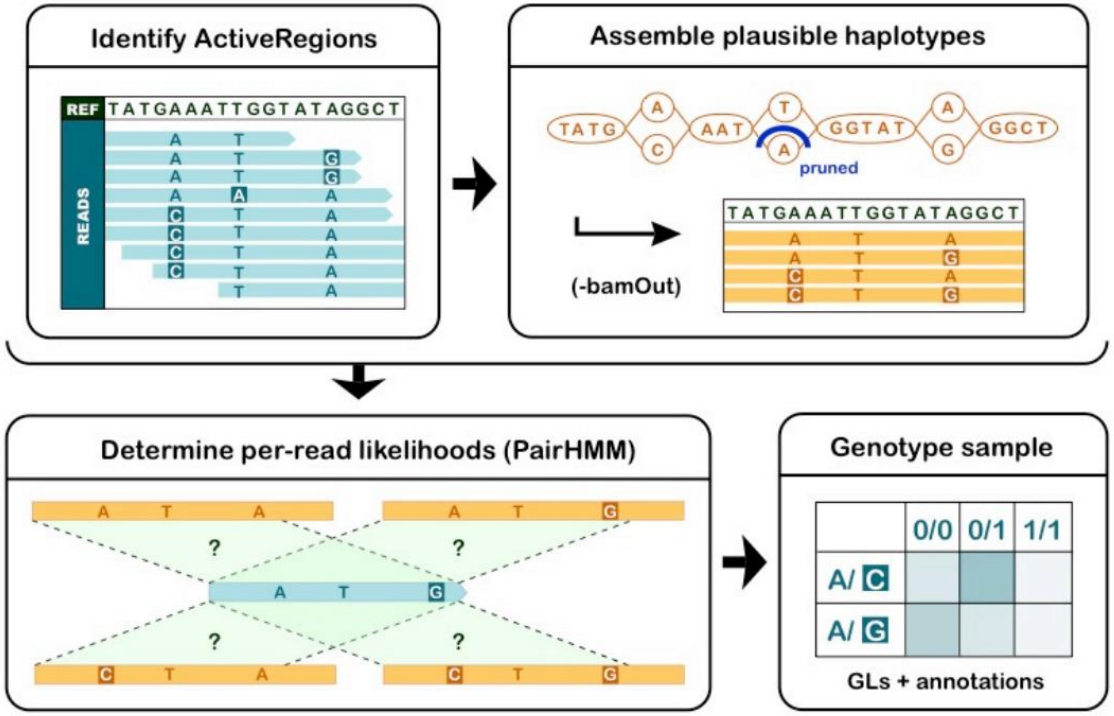
GATK


Data Pre-processing
>>
Variant Discovery
>>
Preliminary Analyses
Raw Reads
1
Non-GATK
Map to Ref (BWA)
& Add Read Groups
Mark Duplicates
& Sort (Picard)
Indel Realignment
Base Recalibration
Analysis-Ready
Reads
Analysis-Ready
Reads
Var. Calling
HC in gVCF mode
1
Genotype Likelihoods
Joint Genotyping
Raw Variants
SNPs
Indels
Variant Recalibration
separately per variant type
Analysis-Ready
Variants
SNPs
Indels
Analysis-Ready
Variants
SNPs
& Indels
Genotype
Refinement
Functional
Annotation
Variant Evaluation
look good?
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



Variant quality score recalibration

Purpose

- Assign a well-calibrated probability to each variant call
- Uses a list of **true variant sites** as input (HapMap, 1000Genomes, own set)

1 - Create recalibration file

- Takes the overlap of the training/truth resource sets and of your callset
- Models the distribution relative to specified annotations (depth, quality, read position, …) → group them into clusters
- Variants closer to cluster center → higher score than outliers

2 - Apply recalibration

- Use recalibration file to assign score
- Output field: VQSLOD

Detailed tutorials

AUSTRIAN INSTITUTE OF TECHNOLOGY

(howto) Recalibrate base quality scores = run BQSR

Comments (27) →

Objective

Recalibrate base quality scores in order to correct sequencing errors and other experimental artifacts.

Prerequisites

- TBD

Steps

1. Analyze patterns of covariation in the sequence dataset
2. Do a second pass to analyze covariation remaining after recalibration
3. Generate before/after plots
4. Apply the recalibration to your sequence data

1. Analyze patterns of covariation in the sequence dataset

Action

Run the following GATK command:

```
java -jar GenomeAnalysisTK.jar \
    -T BaseRecalibrator \
    -R reference.fa \
    -I realigned_reads.bam \
    -L 20 \
    -knownSites dbsnp.vcf \
    -knownSites gold_indels.vcf \
    -o recal_data.table
```

Expected Result

This creates a GATKReport file called `recal_data.grp` containing several tables. These tables contain the covariation data that will be used in a later step to recalibrate the base qualities of your sequence data.

It is imperative that you provide the program with a set of known sites, otherwise it will refuse to run. The known sites are used to build the covariation model and estimate empirical base qualities. For details on what to do if there are no known sites available for your organism of study, please see the online GATK documentation.

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 classfier 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

Somatic variants

Variant calling

Somatic calling – some tools

MuTec

- Statistical analysis to identifies sites carrying somatic mutations using Bayesian classifiers
- http://www.broadinstitute.org/cancer/cga/mutect

VarScan 2

- Heuristic method and a statistical test based on aligned reads supporting each allele
- http://varscan.sourceforge.net/

SomaticSniper

- Calculates the probability that the tumor and normal genotypes are different
- http://gmt.genome.wustl.edu/somatic-sniper/

Somatic calling - more tools ....

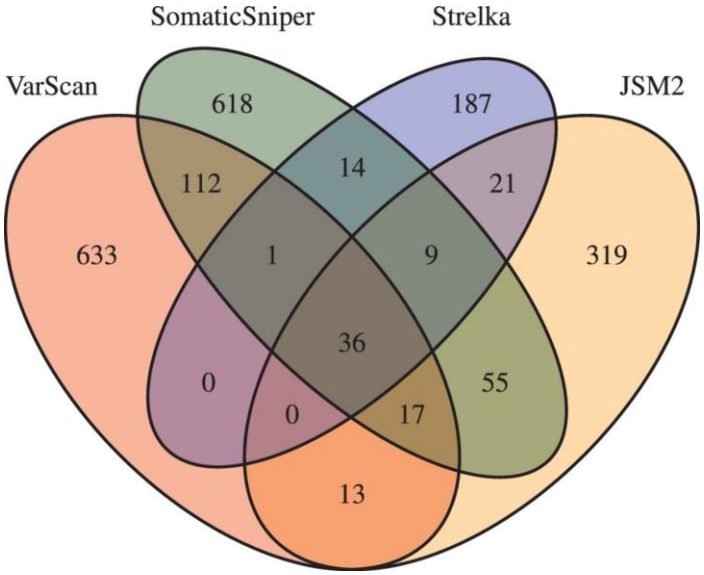
https://www.biostars.org/p/19104/

Here are a few more, a summary of the other answers, and updated link

- deepSNV (abstract) (paper)
- EBCall (abstract) (paper)
- GATK SomaticIndelDetector (note: only available after an annoyi
- Isaac variant caller (abstract) (paper)
- joint-snv-mix (abstract) (paper)
- LoFreq (abstract) (paper) (call on tumor & normal separately and
- MutationSeq (abstract) (paper)
- MutTect (abstract) (paper) (note: only available after an annoying
- QuadGT (for calling single-nucleotide variants in four sequenced the two parents)
- samtools mpileup - by piping BCF format output from this to bcftc
- Seurat (abstract) (paper)
- Shimmer (abstract) (paper)
- SolSNP (call on tumor & normal separately and then compare to
- SNVMix (abstract) (paper)
- SOAPsnv
- SomaticCall (manual)
- SomaticSniper (abstract) (paper)
- Strelka (abstract) (paper)

Roberts et al., Bioinformatics (2013)

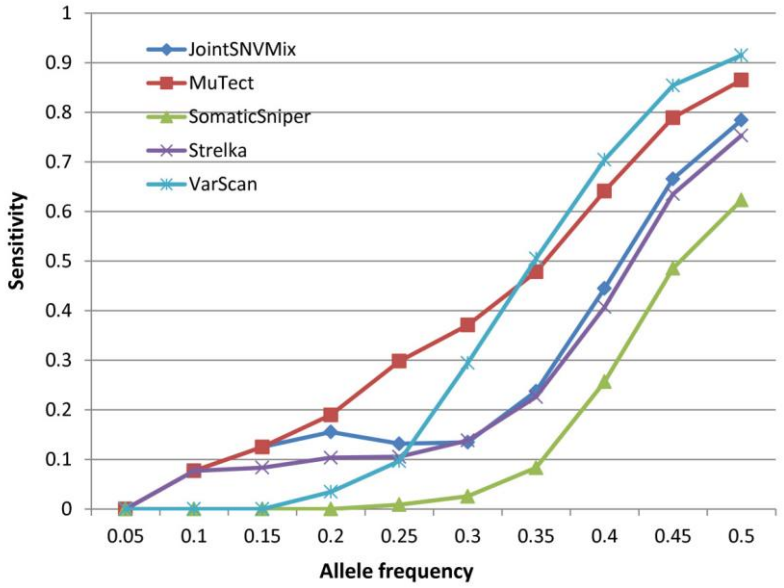
VarScan, SomaticSniper, JSM2 and Strelka revealed substantial differences as to the number and character of sites returned; the somatic probability scores assigned to the same sites; their susceptibility to various sources of noise; and their sensitivities to low-allelic-fraction candidates



Roberts et al. **A comparative analysis of algorithms for somatic SNV detection in cancer**
Bioinformatics (2013) 29 (18): 2223-2230

Wang et al., Genome Medicine (2013)

Tools (EBCall, JointSNVMix, MuTect, SomaticSniper, Strelka, and VarScan 2) have significant room for improvement, especially in the discrimination of low coverage/allelic-frequency sSNVs and sSNVs with alternate alleles in normal samples.

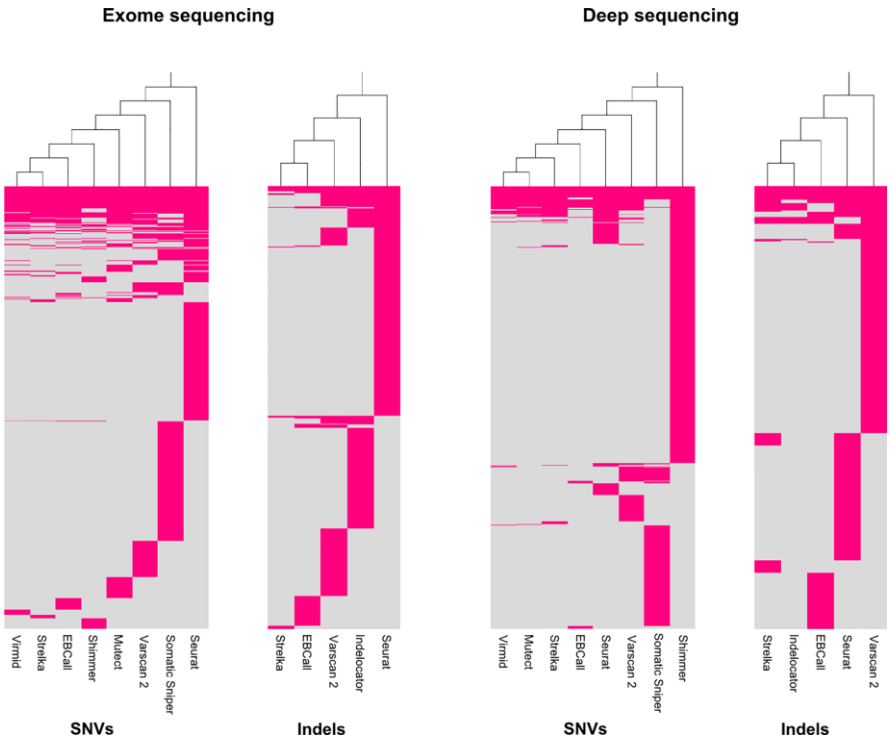


Wang et al. **Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers**
Genome Medicine (2013), **5**:91

Krøigård et al, PLOS One 2016

Evaluation of Nine Somatic Variant Callers

- Major differences among the nine studied somatic variant callers
- EBCall, Mutect, Strelka and Virmid all perform well in our study
- Sequencing depth had markedly diverse impact on individual callers

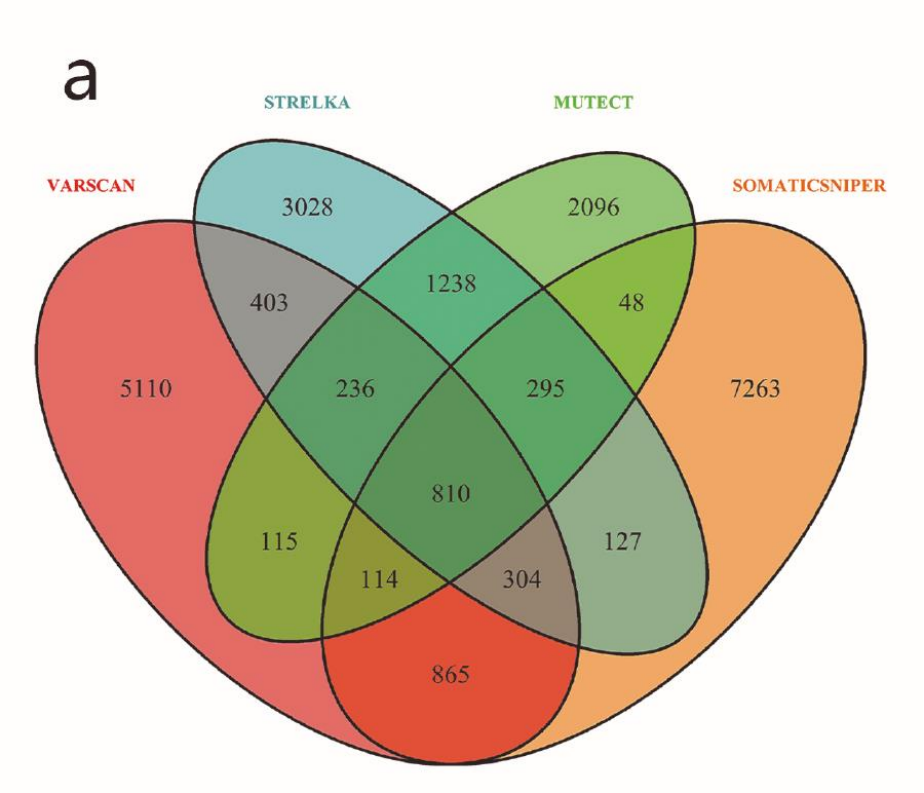


Hierarchical cluster analysis of mutations called by the somatic variant callers in exome and deep sequencing data in left and right panel, respectively. Each red line represents a called somatic mutation

Unpublished

- Mutect & Strelka performed best
- Different results based on coverage
  - Higher coverage → more TP, but also more FP

- Filtering based on germline information



Cake & SomaticSeq

Cake

- Integrates 5 somatic variant callers (Samtools mpileup, Varscan 2, Bambino, SomaticSniper, CaVeMan)
- Outputs **high-confidence set of somatic alteration**
- Tradeoff --- specificity vs. sensitivity

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3740632/

SomaticSeq

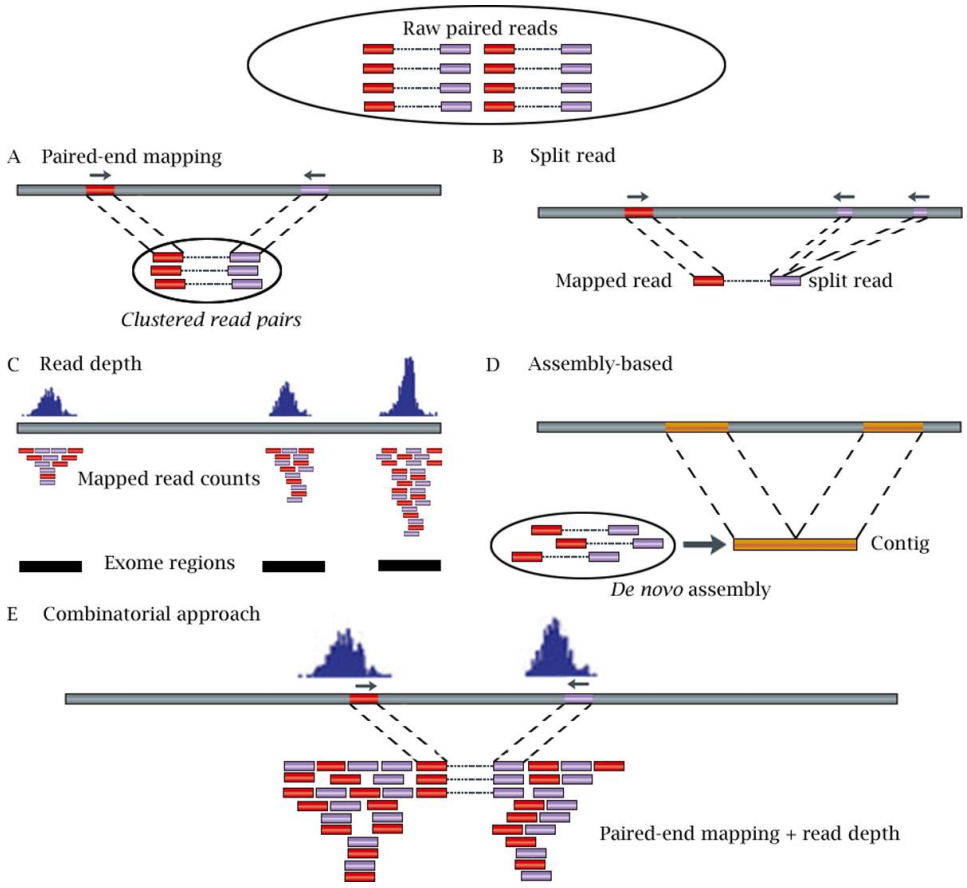
- Integrates 5 somatic variant callers (MuTect, SomaticSniper, VarScan2, JointSNVMix2, and VarDict)
- Achieves better overall accuracy than any individual tool incorporated

http://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0758-2

Structural Variations (SV)

Variant calling

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.

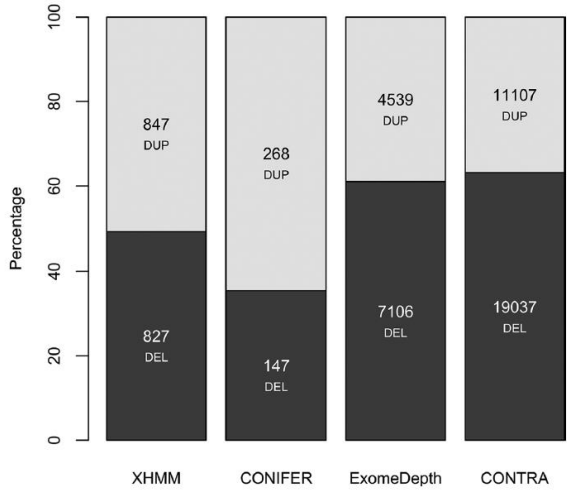
D. Assembly (AS)-based approach detects CNVs by **mapping contigs** to the reference genome.

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

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

CNV detection

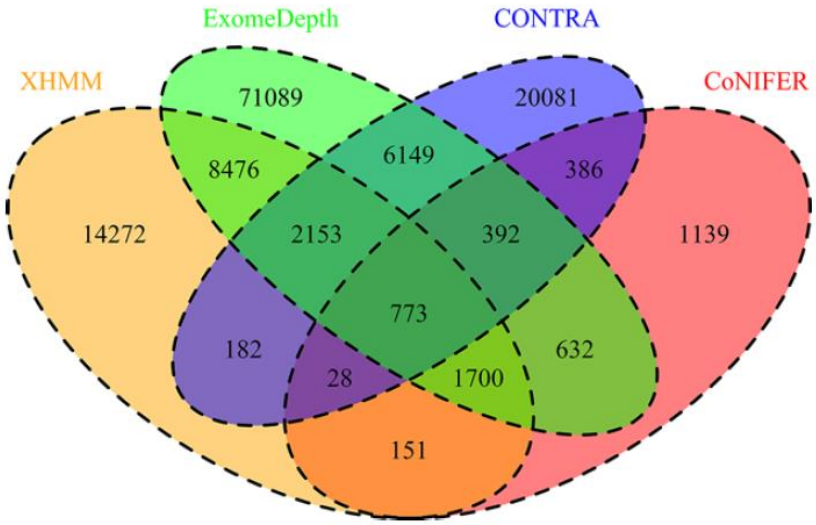
- XHMM
- CoNIFER
- ExomeDepth
- CONTRA



33 individual WES data (cumulatively)
Deletion and duplication CNVs

CNV detection

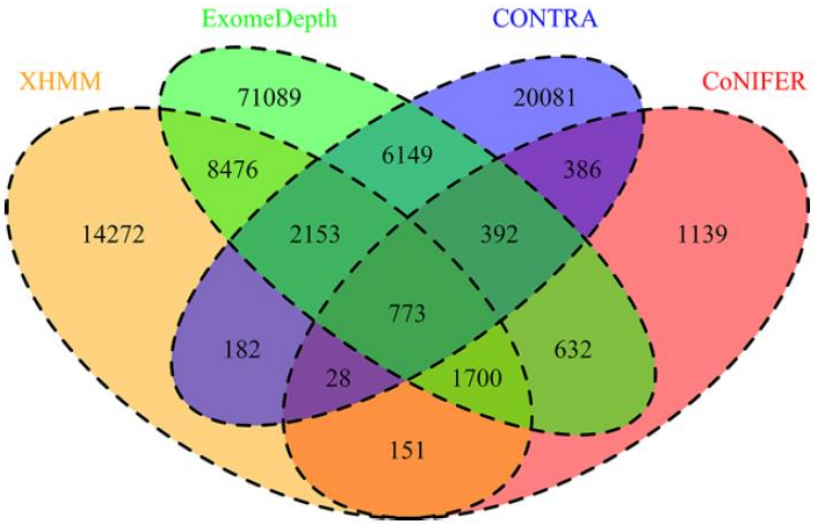
- XHMM
- CoNIFER
- ExomeDepth
- CONTRA



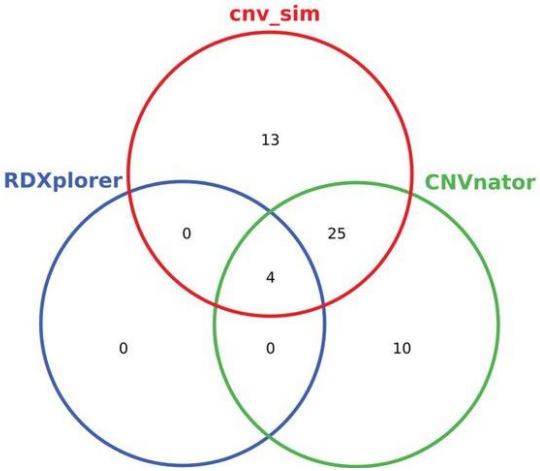
Overlap of CNVs at exon level (33 WES)

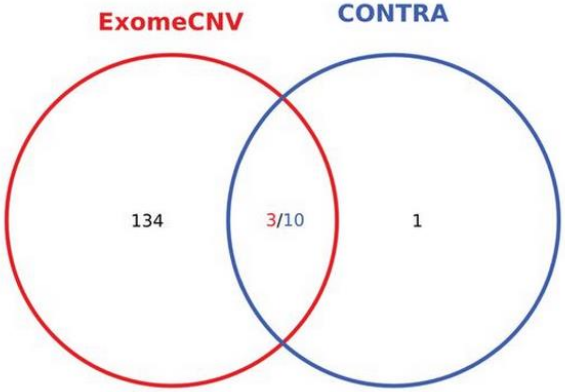
CNV detection

- XHMM
- CoNIFER
- ExomeDepth
- CONTRA



- CONTRA
- ExomeCNV
- RDXplorer





- CNVnator

CNV detection – more tools

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

- 37 CNV tools
  6 PEM, 4 SR, 26 RD, 3 AS, 9 combinatorial approaches

Tool	URL
SegSeq[a]	http://www.broad.mit.e
CNV-seq[a]	http://tiger.dbs.nus.edu
RDXplorer[b]	http:/
BIC-seq[a]	http:/
CNAseg[a]	http:/
cn.MOPS[b]	http:/
JointSLMh	http:/

Tool	URL	Language	Input
Control-FREEC[a]	http://bioinfo-out.curie.fr/projects/freec/ webcite	C++	SAM/BAM/pileup/E formats
CoNIFER[b]	http://conifer.sf.net webcite/	Python	BAM

Method	URL	L
		+
NovelSeq	http://compbio.cs.sfu.ca/strvar.htm webcite	C
HYDRA	http://code.google.com/p/hydra-sv webcite/	hon, P
CNVer	http://compbio.cs.toronto.edu/CNVer webcite/	a, P
GASVPro	http://code.google.com/p/gasv webcite/	hon, R, C
Genome STRiP	http://www.broadinstitute.org/software/genomestrip/genome-strip webcite	J
SVDetect	http://svdetect.sourceforge.net webcite/	P

Input
BAM
BAM/pileup
SAM/BAM
Sorted BED files
SAM/pileup
N/A

CNV detection

AUSTRIAN INSTITUTE OF TECHNOLOGY

Summary

- Many tools
- Result from different tools vary
- Check if tool is optimized for certain data

Structural variations

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

SVDetect

- Large deletions and insertions, inversions, intra- and inter-chromosomal rearrangements

SVMerge

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

SV – more tools

Abel *et al. Cancer Genetics* 2013 Pages 432–440

Review article

Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches

Haley J. Abel[a], Eric J. Duncavage[b,]

Show more

http://dx.doi.org/10.1016/j.cancergen.2013.11.002

Get rights and content

Next generation sequencing (NGS), or massiv
methods in which numerous sequencing reac
amounts of sequencing data for a small fraction

Table 1.
Software tools for evaluation of structural variation in NGS data

	Comment	Download link
Translocations and Inversions		
Discordant paired end		
BreakDancer	Fast, simple to run	http://breakdancer.sourceforge.net
Hydra	Considers multiple mappings of discordant pairs	https://code.google.com/p/hydra-sv/
VariationHunter	Considers multiple mappings of discordant pairs	http://variationhunter.sourceforge.net/Home
PEMer	Simulates structural	http://sv.gersteinlab.org/pemer/introduction.html

Variant filtering

Variant filtering

The biggest problem is large numbers of FPs and FNs:

- 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**

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 balance between strands, alternate carrying reads should be as well

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

Variant filtering – (some) tools

AUSTRIAN INSTITUTE OF TECHNOLOGY

vcfutils (with samtools)

```
./vcfutils.pl varFilter -Q 20 -d 10 -D 200 <file.vcf>
```

VCFtools

```
./vcftools --vcf <file.vcf> --min-meanDP
```

GATK pipeline

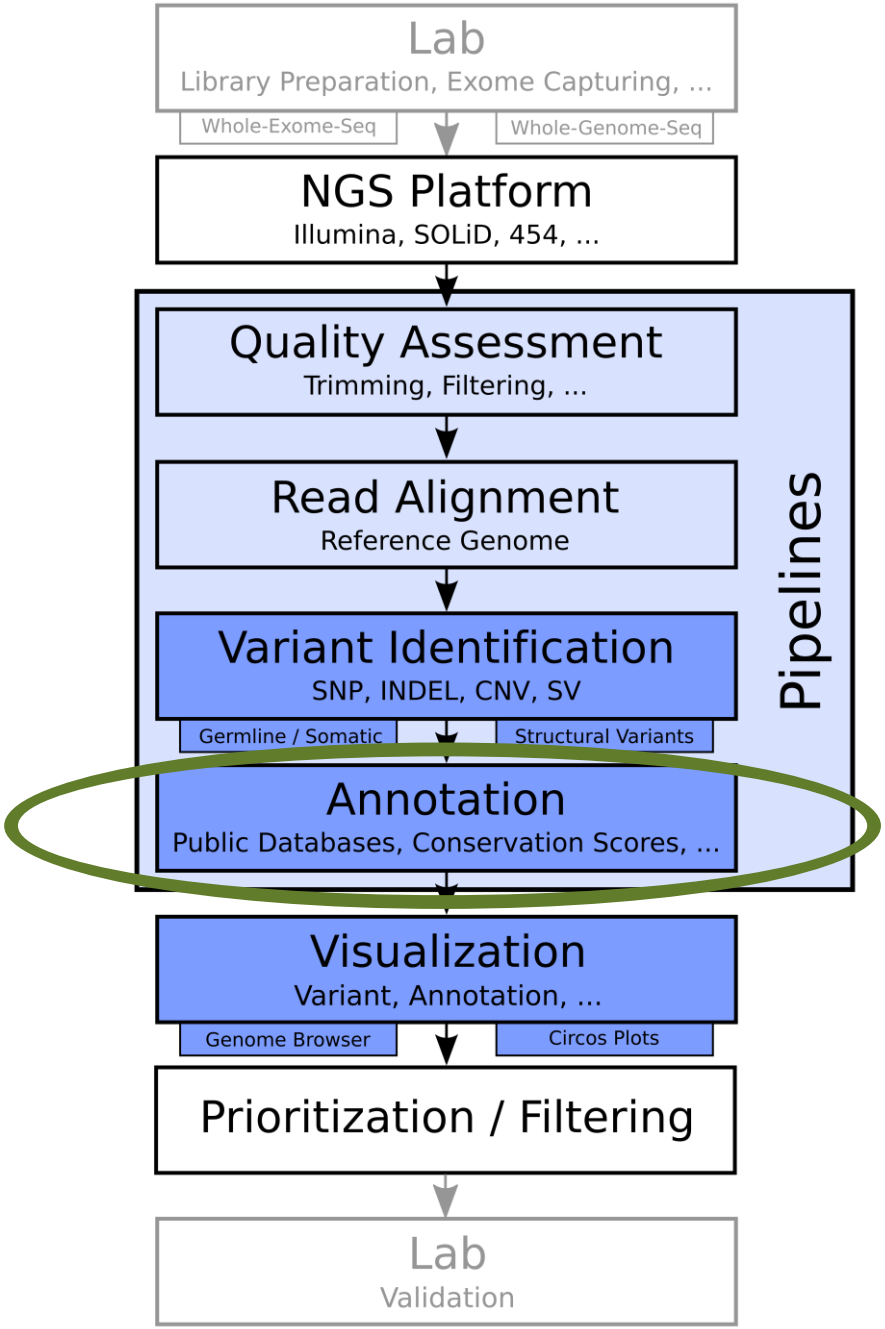
```
java -jar GenomeAnalysisTK.jar -T VariantFiltration -R
<reference.fa> -V <file.vcf>  --filterExpression "QD < 2.0 || MQ <
40.0 || MappingQualityRankSum < -12.5 --filterName "my_snp_filter"
-o <filtered_snps.vcf>
```

vcffilter (vcflib)

```
vcffilter -f "DP > 10 & MQ > 30 & QD > 20" file.vcf
```

AUSTRIAN INSTITUTE OF TECHNOLOGY

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

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

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

Sequence variant analyzer (SVA)

- Java based, GUI
- visualize variants

snpEff

- Integrated within Galaxy and GATK.
- SNPs and INDELs

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

Variant annotation is not yet a solved problem

- Differences between REFSEQ and ENSEMBL transcript set
  - More variants with annotations in interesting categories when using ENSEMBL transcripts
- Choice of annotation software can have a substantial effect
- Differences particularly large in annotation categories of most interest
  - putative loss-of-function
  - nonsynonymous variants

McCarthy et al. **Choice of transcripts and software has a large effect on variant annotation** *Genome Medicine* 2014, **6**:26

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/

Other widely used file formats

GFF / GTF / BED

GFF file format

GFF3 – Generic Feature Format

http://www.sequenceontology.org/gff3.shtml

- Tab separated with 9 columns
- Supports hierarchy levels (Parent attribute)
- Online validator available

Used for describing

- genes
- features of DNA
- protein sequences
- …

GFF columns

- Seqid (usually chromosome)
- Source (source of data)
- Type (usually term from seq. ontology)
- Start
- End
- Score (floating point number)
- Strand (+ - .)
- Phase (reading frame for coding sequences)
- Attributes (separated by ";") – some with predefined meaning: ID, Name, Parent, Gap …

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

http://www.ensembl.org/info/website/upload/gff.html

GTF

General Transfer Format (GTF)

- Based on GFF
- Feature types: "CDS", "start_codon", "stop_codon". Optional: "5UTR", "3UTR", "inter", "inter_CNS", "intron_CNS" "exon".
- Mandatory attributes
  - *gene_id* - unique identifier for the genomic locus of the transcript. If empty, no gene is associated with this feature.
  - *transcript_id* - unique identifier for the predicted transcript.

```
381 Twinscan  CDS          380   401   .   +   0  gene_id "001"; transcript_id "001.1";
381 Twinscan  CDS          501   650   .   +   2  gene_id "001"; transcript_id "001.1";
381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1";
381 Twinscan  stop_codon   708   710   .   +   0  gene_id "001"; transcript_id "001.1";
```

http://mblab.wustl.edu/GTF22.html

BED

- Tab separated - 3 required and 9 optional columns
- Flexible way to define the data lines
- Order of the optional fields is binding

Required

- chrom (name of the chromosome, sequence id)
- chromStart (starting position on the chromosome)
- chromEnd (end position of the chromosome, note this base is not included!)

Used for

- Annotation tracks
- Interval files (for variant calling)
- ...

Visualization

Visualization

Genome browsers - most widely-used tools

Read

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

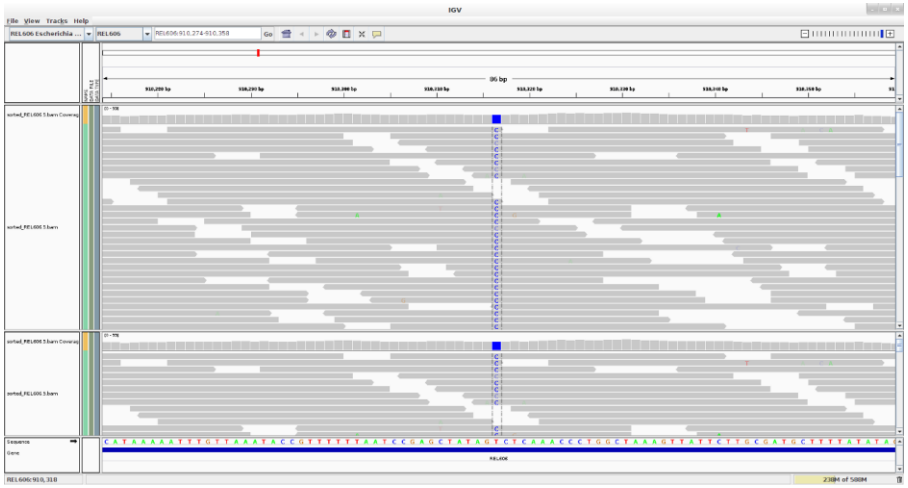
Able to

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

Genome Browsers

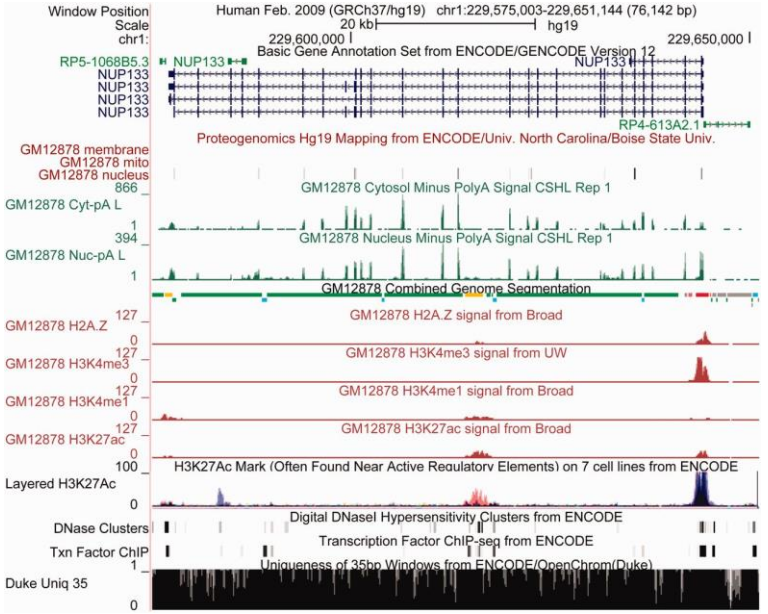
IGV (Integrative Genomics Viewer)

- Widely used viewer
- Java based – standalone tool
- Easy and fast to view own data



UCSC

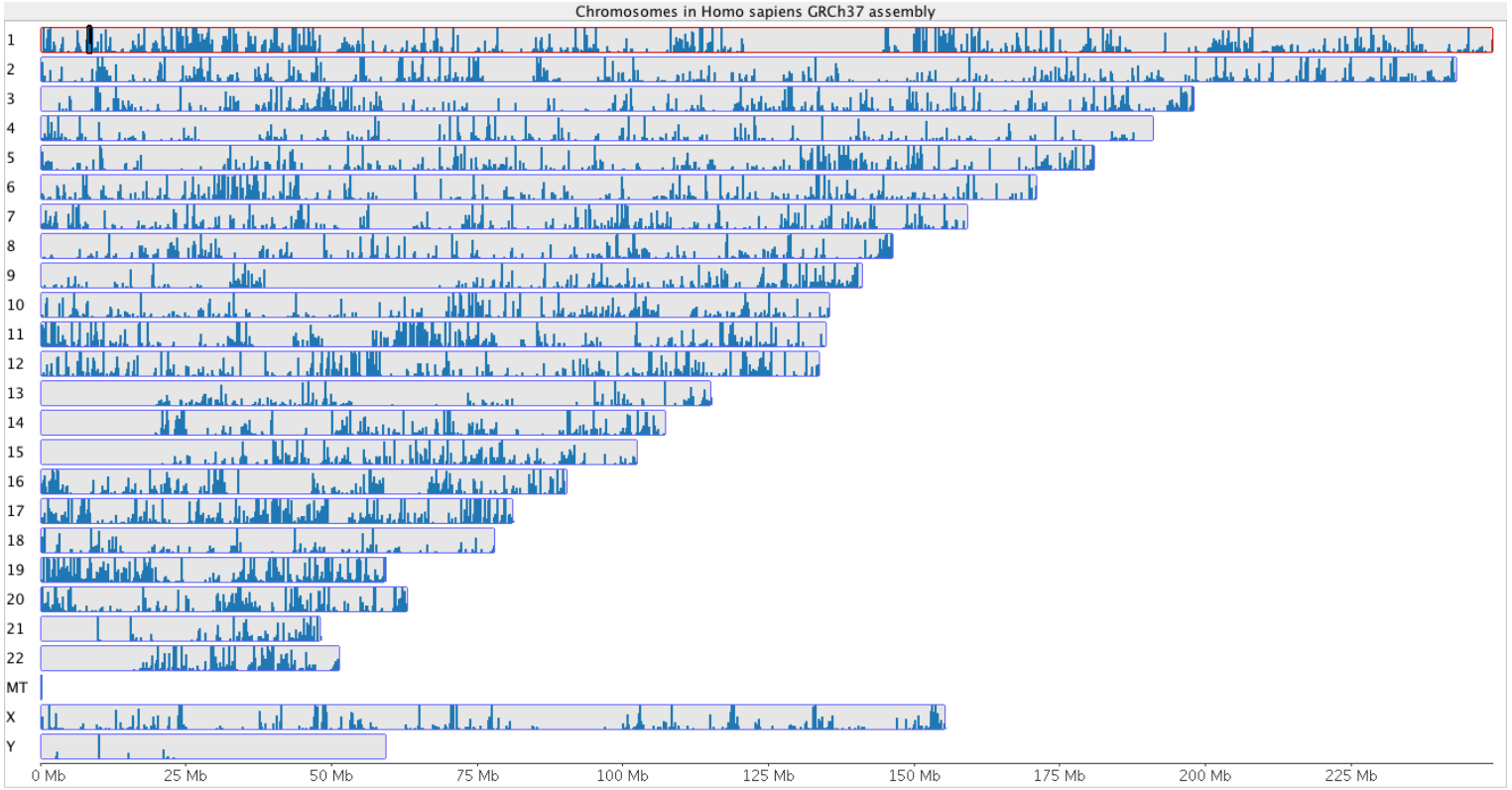
- Web based tool
- Offers many different annotation tracks
- Needs some configuration to display own data



Coverage visualization

Coverage histogram for chromosomes

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



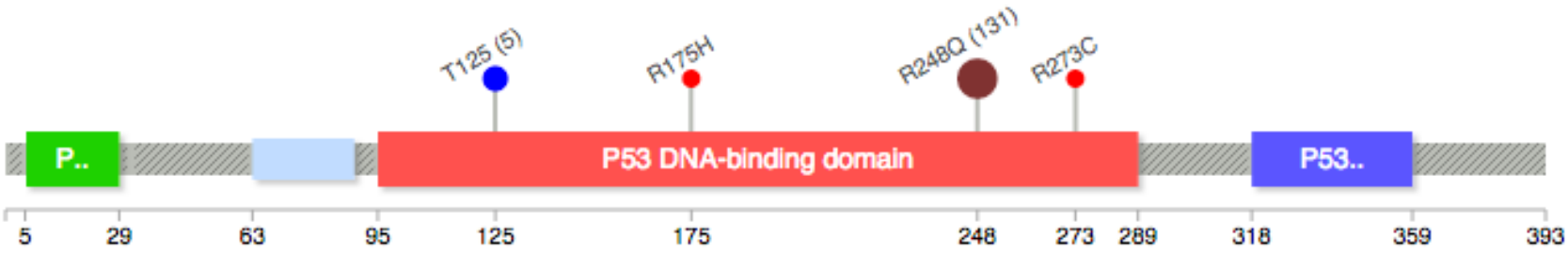
http://seqanswers.com/forums/attachment.php?attachmentid=2118&d=1364889859

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
```



Analysis pipelines and workflow systems

Pipelines

Bcbio pipeline

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

Crossbow

- Bowtie & SoapSNP
- Runs in the cloud using Hadoop cluster

HiPipe

- DNA and RNA analysis
- BWA, GATK

ngs_backbone

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

Pipelines

SeqWare

- Analysis on Grid and Cloud
- Workflow deployment and management system

Firehose

- Used at the Broad institute
- Focus on automation
- Java based – web frontend

Galaxy

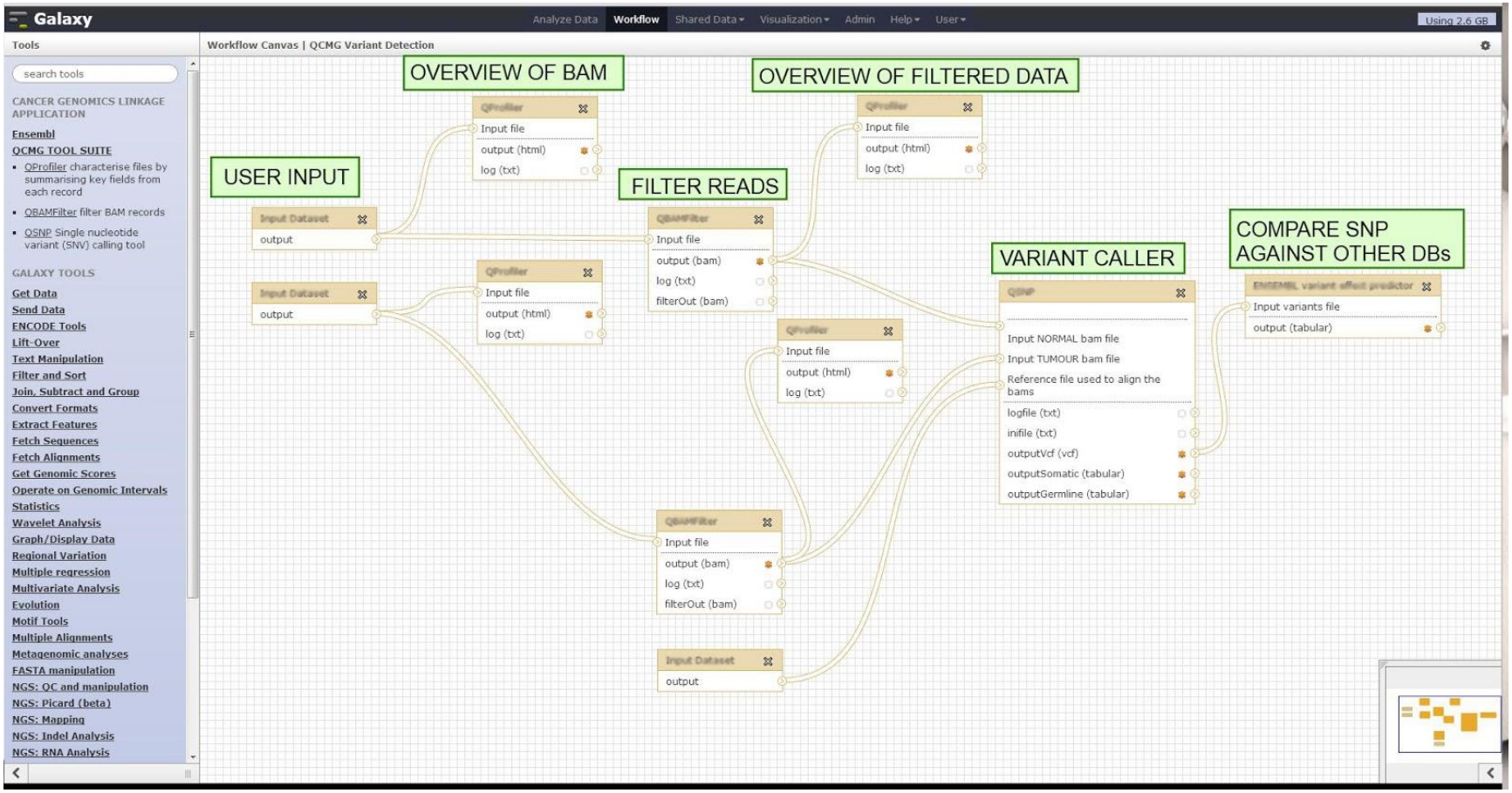
"Galaxy is an open, web-based platform for data intensive biomedical research"

- 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

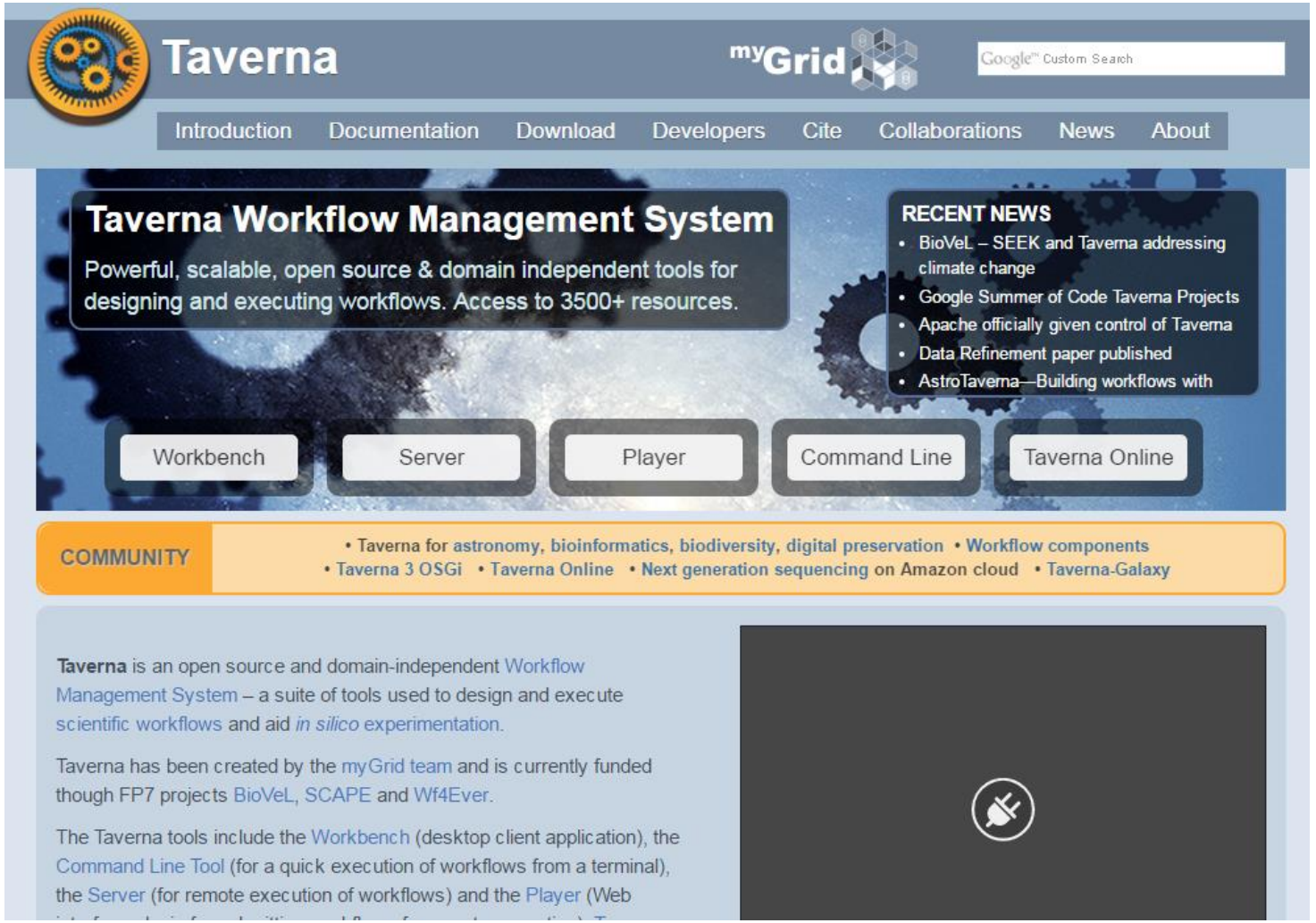
Galaxy



„The Cancer Genomics Linkage Application”

Taverna

- Workflow management system



Misc – useful information

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

Courses

- http://www.ebi.ac.uk/training/online/outline_print/6635/all
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- http://www.personal.psu.edu/iua1/courses/2013-BMMB-597D.html
- http://www.coursera.org

Practical 2

https://github.com/tadKeys/BioinformaticsAndGenomeAnalyses2016