AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Bioinformatics and Genome Analyses Course

Lecture 3 (Nov 14)

Stephan Pabinger
stephan.pabinger@ait.ac.at

http://pabinger.site44.com

Recap

Genome reference
Coordinate systems

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

FASTQ

Old format

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

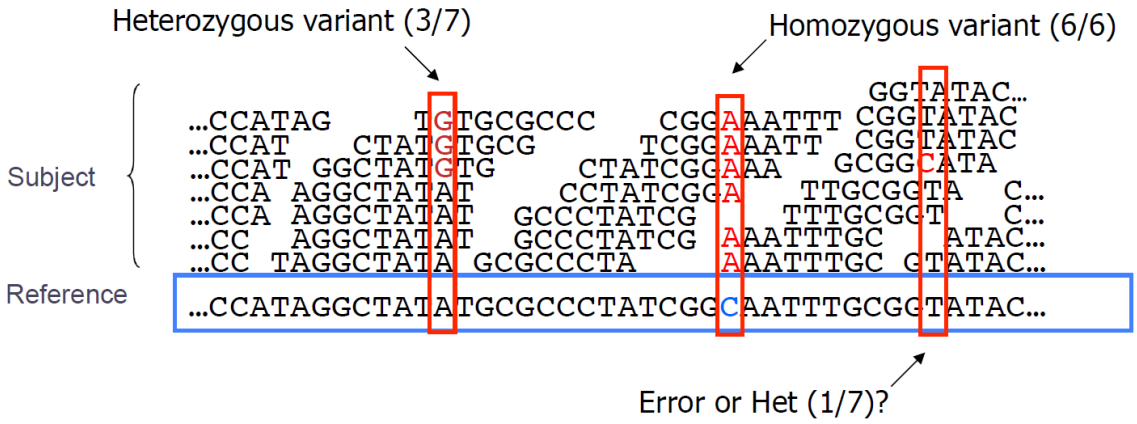
HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end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

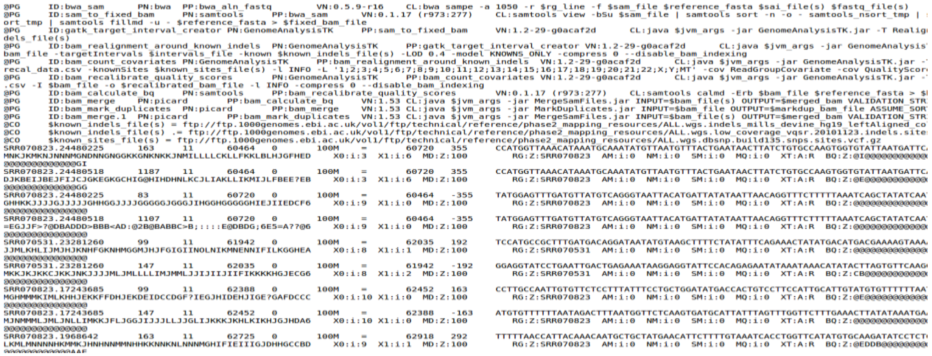
SAM
Variants
Variant calling



Recap

Python programs

- sam_examiner.py



- Vcf_examiner.py

```
##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
```

Check out the "introduction to Python links" on the Github page

Data storage

Guidelines for diagnostic next-generation sequencing

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4795226/

Eur J Hum Genet. 2016

- Should stick to the standard open file formats FASTQ, BAM, and VCF
- Full-log files have to be stored in addition to the analysis results
  - complete as possible
  - making the whole analysis from FASTQ data to the diagnostic report reproducible.
- No (international) consensus yet on what should be stored

ACMG clinical laboratory standards for next-generation sequencing

https://www.acmg.net/docs/acmg_lab_standards_next_generation_sequencing_sept2013.pdf

- Recommend that the laboratory consider a minimum of 2-year storage
- File type that would allow regeneration of the primary results as well as reanalysis
- Retention of the VCF, along with the final clinical test report interpreting the subset of clinically relevant variants

Low quantity DNA Sequencing

A robust targeted sequencing approach for low input and variable quality DNA from clinical samples

- npj Genomic Medicine, 2018
- osSeq, which permits the enrichment of genes and regions of interest and the identification of sequence variants from low amounts of damaged DNA
- https://www.nature.com/articles/s41525-017-0041-4

osSeq - Oligonucleotide Selective Sequencing

- https://emea.illumina.com/science/sequencing-method-explorer/kits-and-arrays/os-seq.html?langsel=/at/
- *Detection of ultra-rare mutations by next-generation sequencing*, PNAS, 2012
  - http://www.pnas.org/content/109/36/14508.long

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

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

Refined INDEL analysis

Examine sources of INDEL errors

- Experimental validation of INDELs called from 30x whole genome vs. 110x whole exome of the same sample
- Most of the errors due to short microsatellite errors introduced during exome capture, also misses most long INDELs
- Recommend WGS for INDEL analysis instead

	All INDELs	Valid	PPV	INDELs >5bp	Valid (>5bp)	PPV (>5bp)
Intersection	160	152	95.0%	18	18	100%
WGS	145	122	84.1%	33	25	75.8%
WES	161	91	56.5%	1	1	100%

PPV = TP/(TP-FP); WGS = WGS specific; WES = WES specific

Reducing INDEL calling errors in whole-genome and exome sequencing data
Fang, H, Wu, Y, Narzisi, G, O'Rawe, JA, Jimenez Barrón LT, Rosenbaum, J, Ronemus, M, Iossifov I, Schatz, MC §, Lyon, GL §
Genome Medicine (2014) 6:89. doi:10.1186/s13073-014-0089-z

Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, INDELs) within exome-capture data.

Features

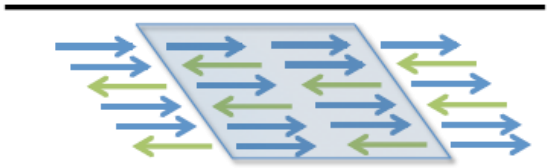
- Combine mapping and assembly
- Exhaustive search of haplotypes
- De novo mutations

Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi et al. (2014) *Nature Methods.* doi:10.1038/nmeth.3069

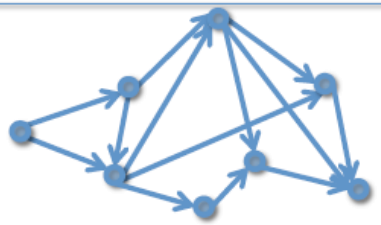
Scalpel Algorithm

AUSTRIAN INSTITUTE OF TECHNOLOGY

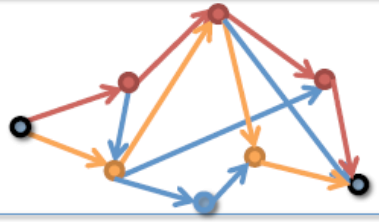
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



Find end-to-end haplotype paths spanning the region



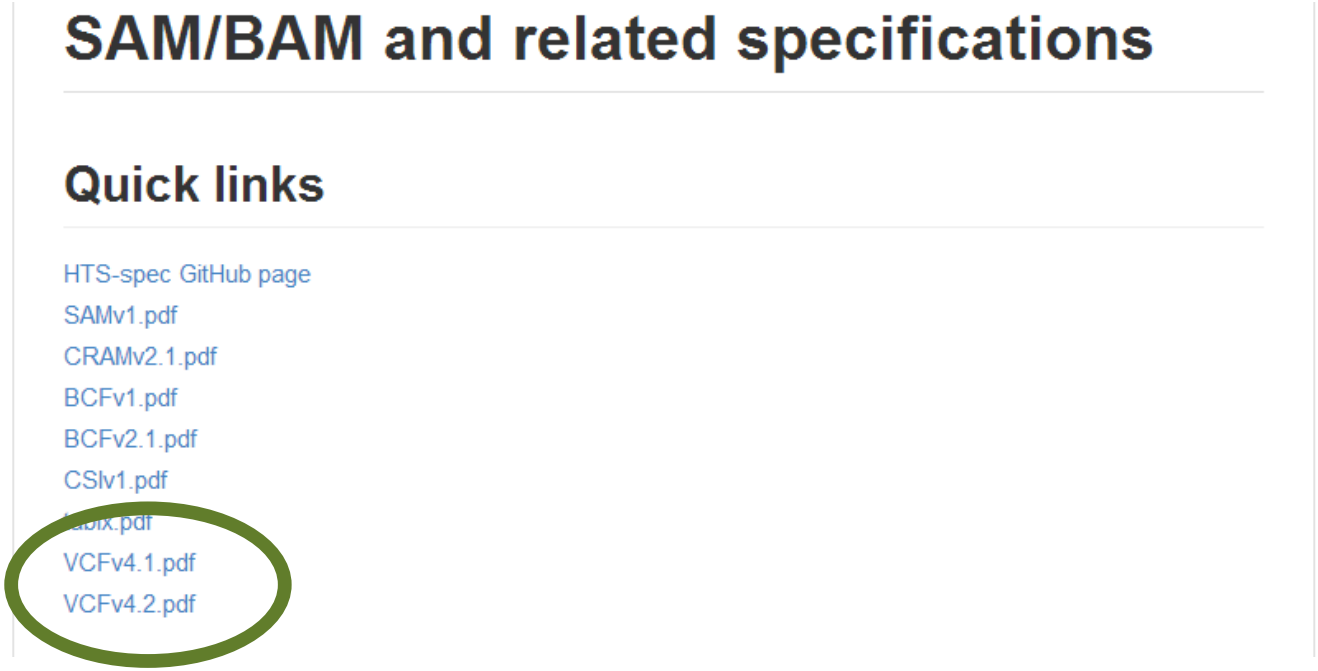
Align assembled sequences to reference to detect mutations

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

QUAL - Variant Calling Quality

Factors

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

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

...

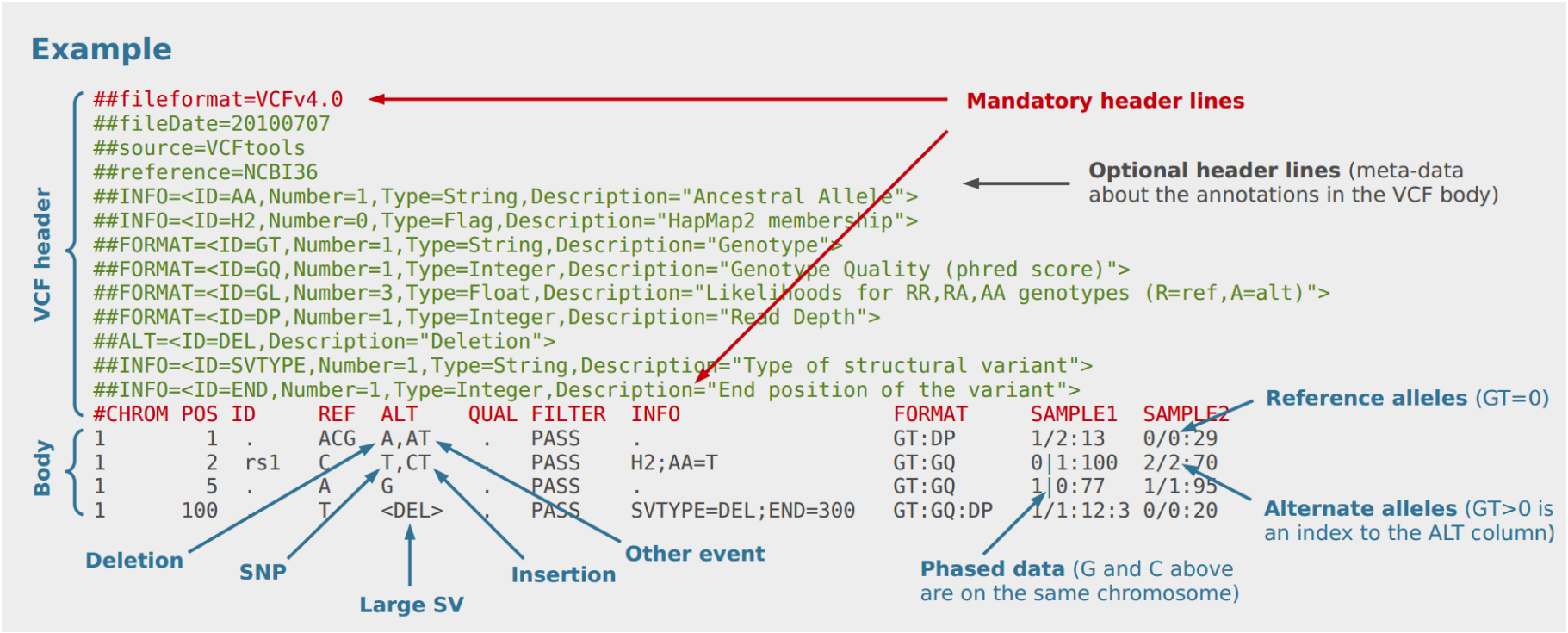
Genotype field

- 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?

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

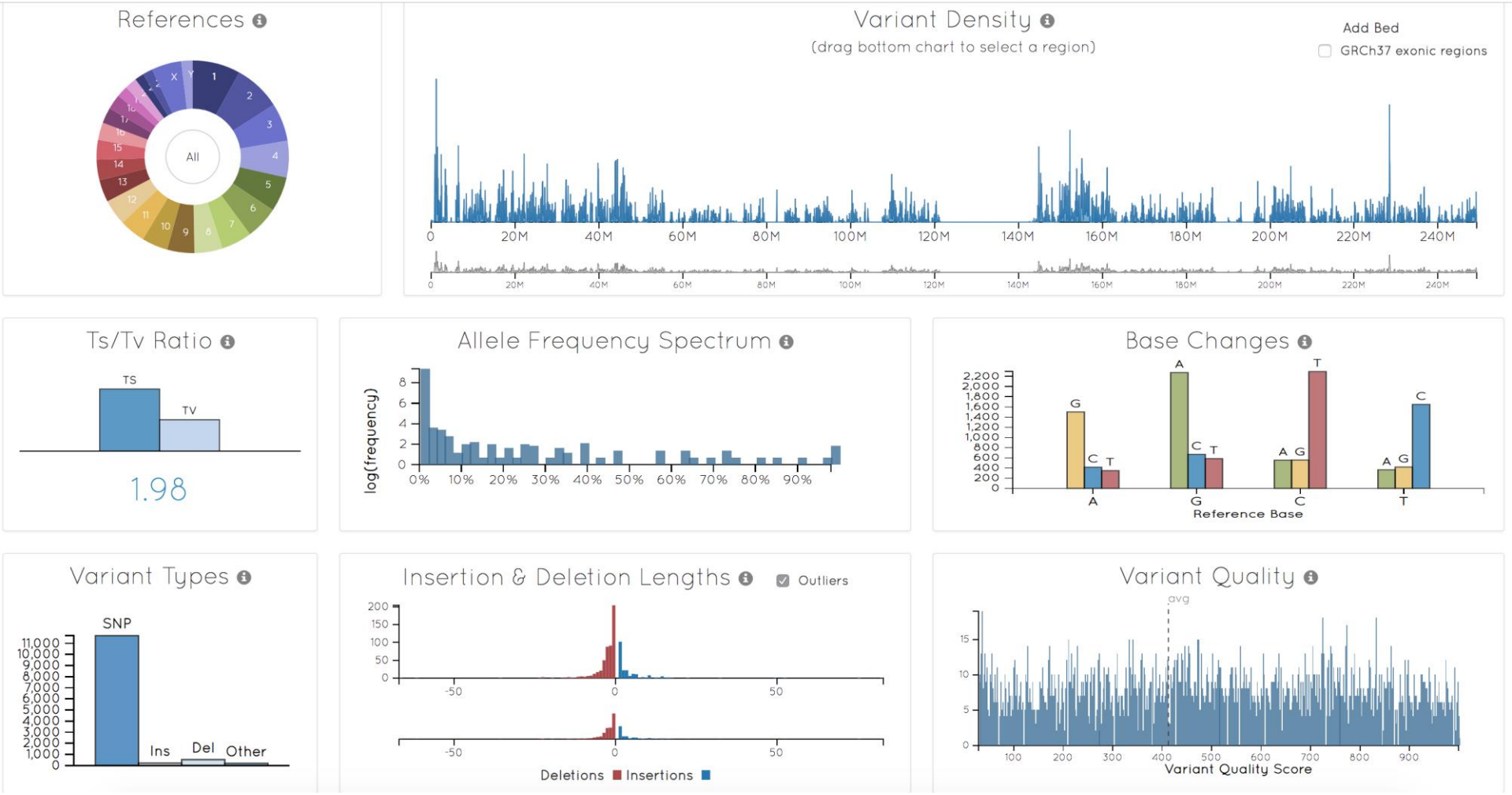
Insertion in VCF

#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

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

Realignment around INDELs

Why?

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

Realignment principles

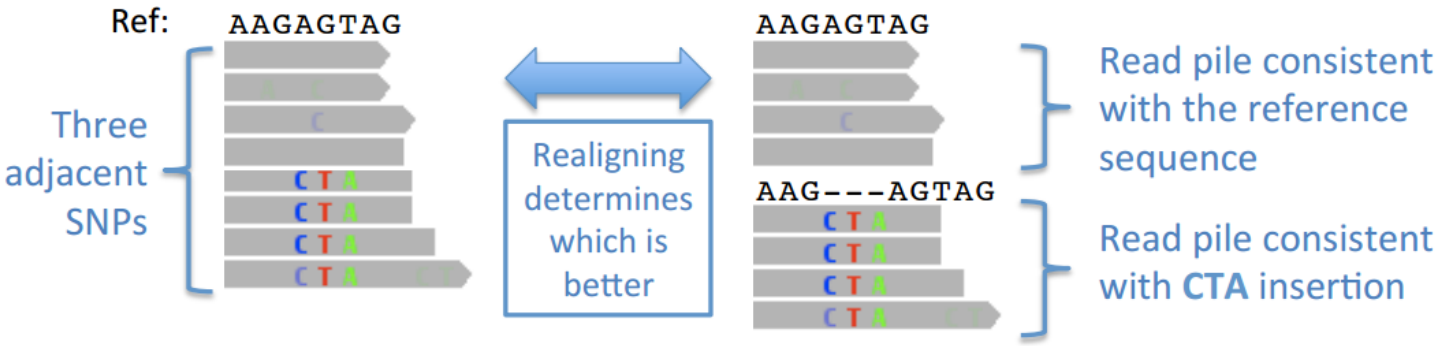
- Realign locally around INDELs→ GATK

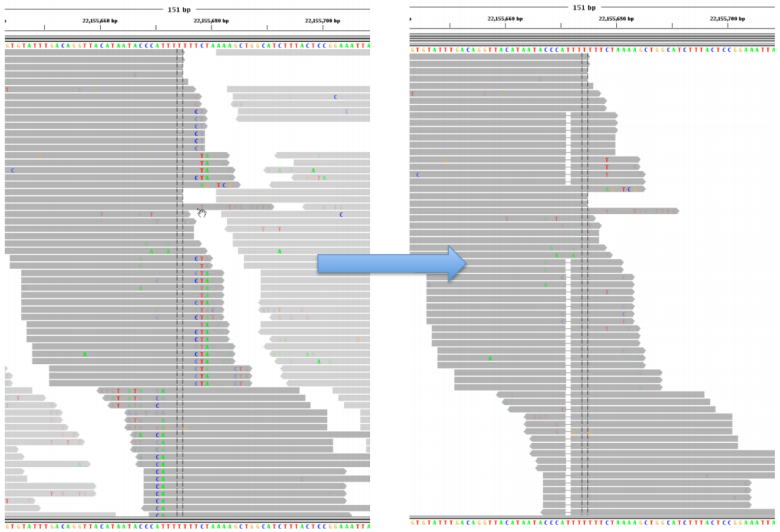
- Input set of known INDEL sites (dbSNP, 1000genomes)
- INDELs from alignments
- Presence of mismatches & softclips (BAM file)

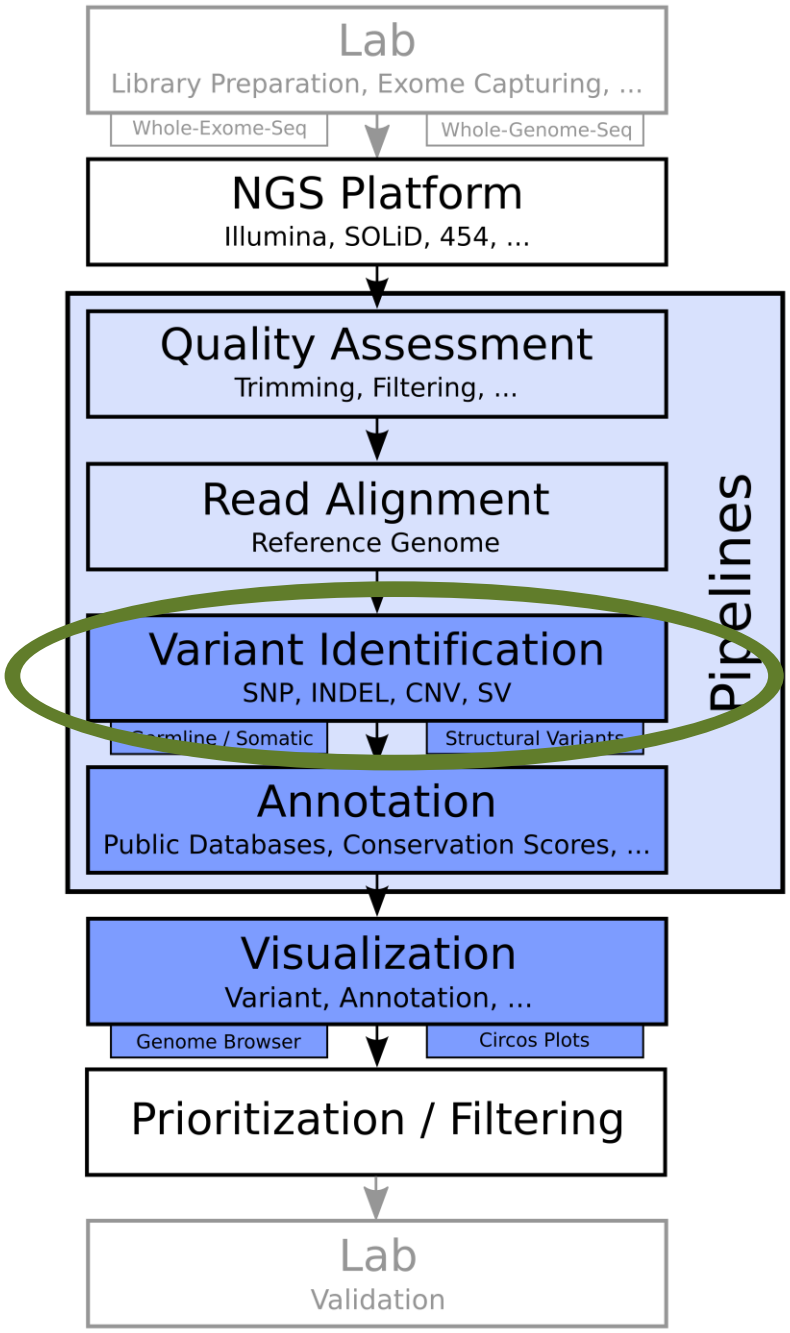
Realignment around INDELs

Realignment

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







Variant callers

SAMtools

Variant calling

SAMtools variant calling

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

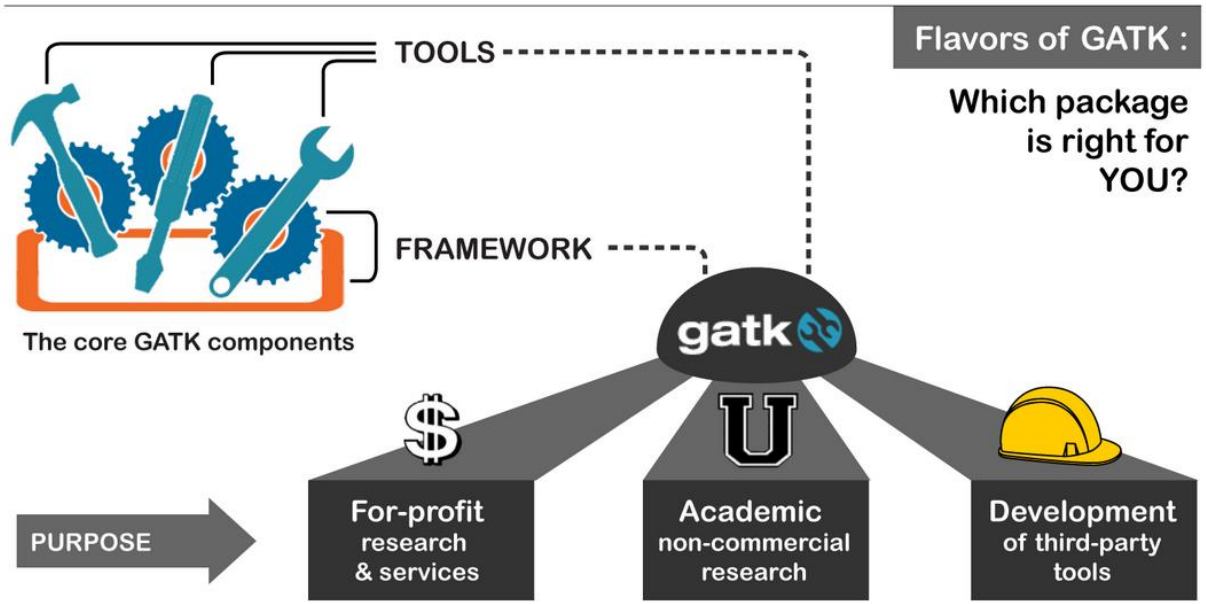
http://ged.msu.edu/angus/tutorials-2013/snp_tutorial.html

GATK - Genome Analysis Toolkit

Variant calling pipeline

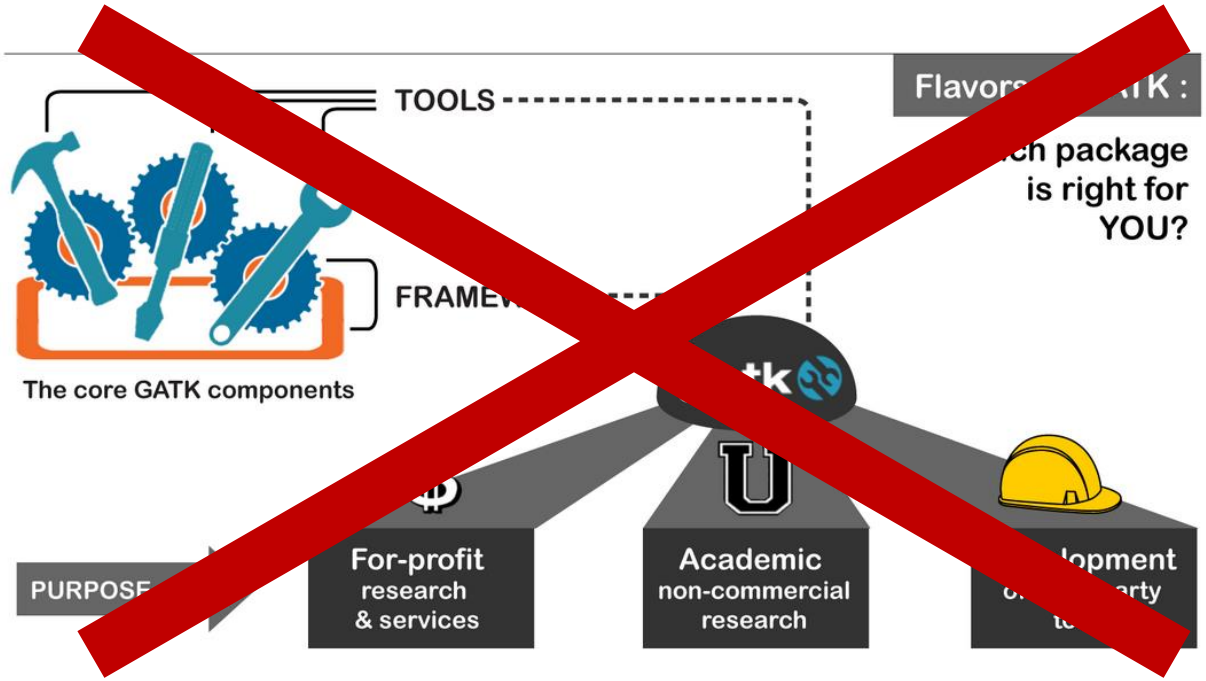
GATK

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



GATK

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

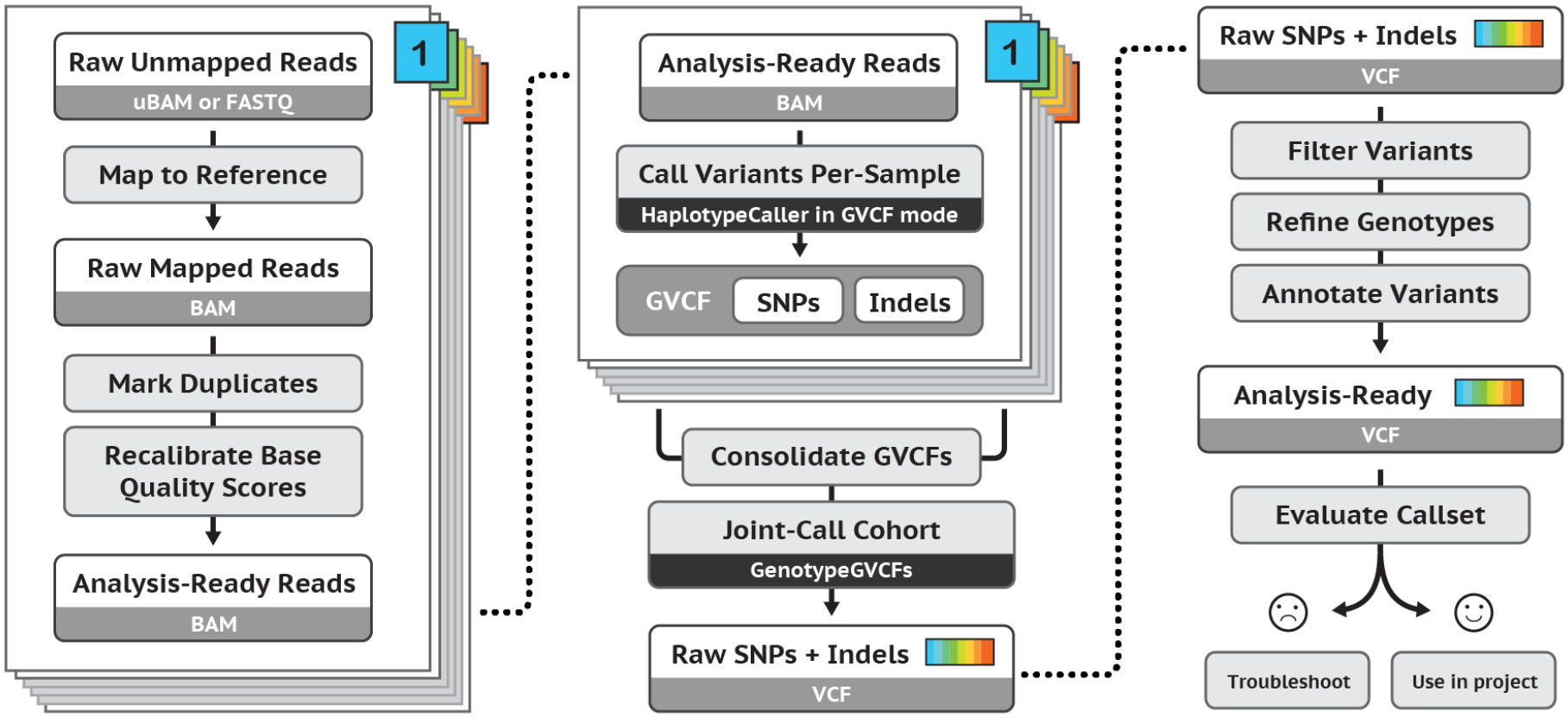


GATK

- 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		

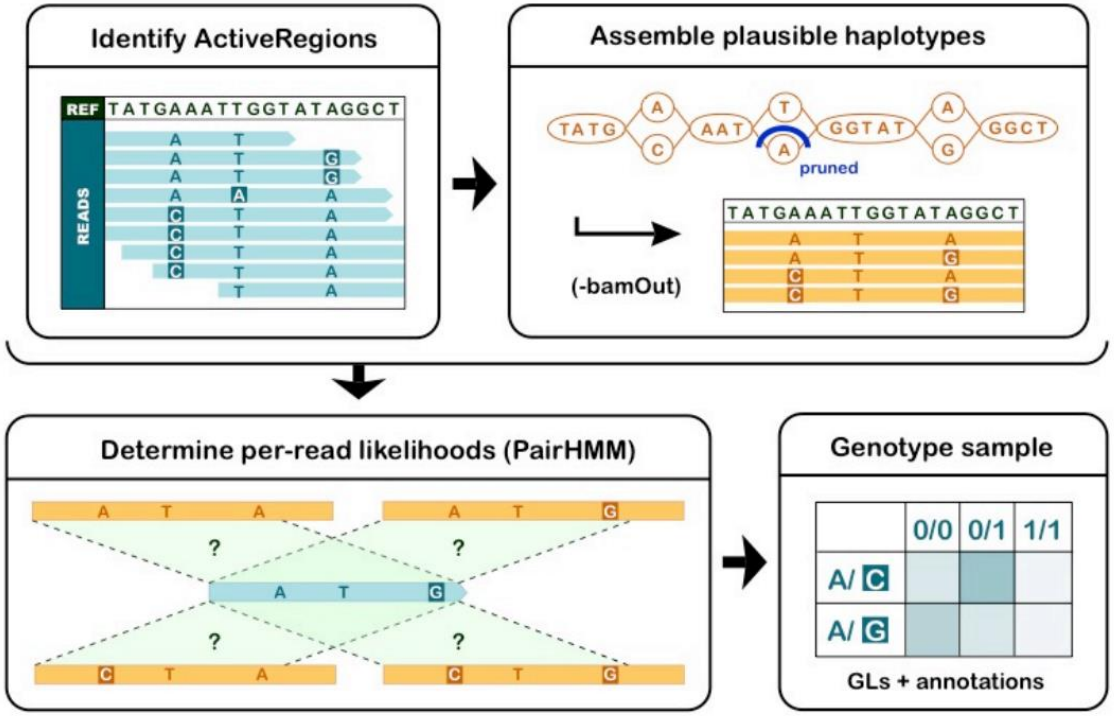
GATK


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

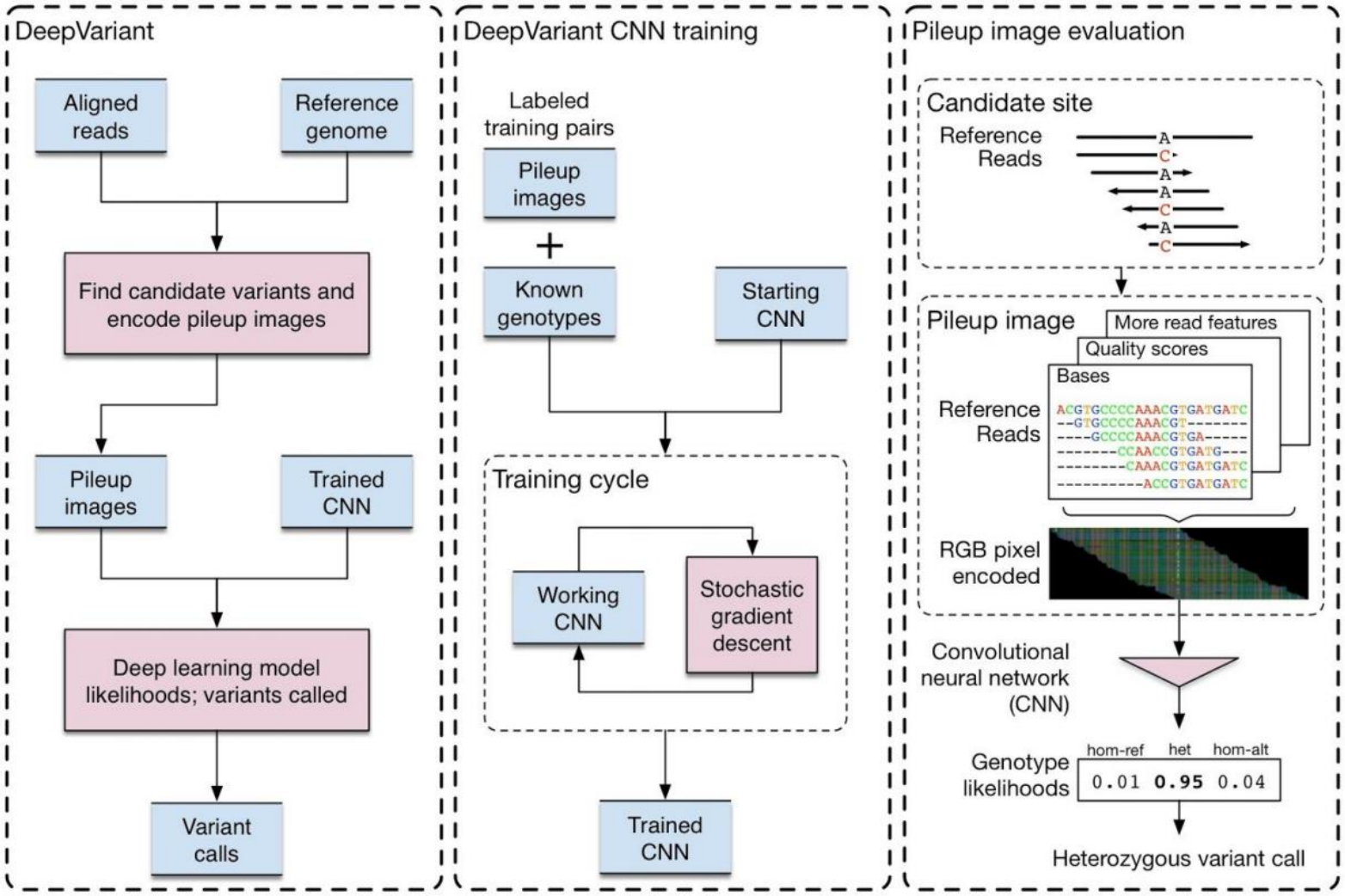
Haplotype Caller

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

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

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

GATK - 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 call-set
- 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

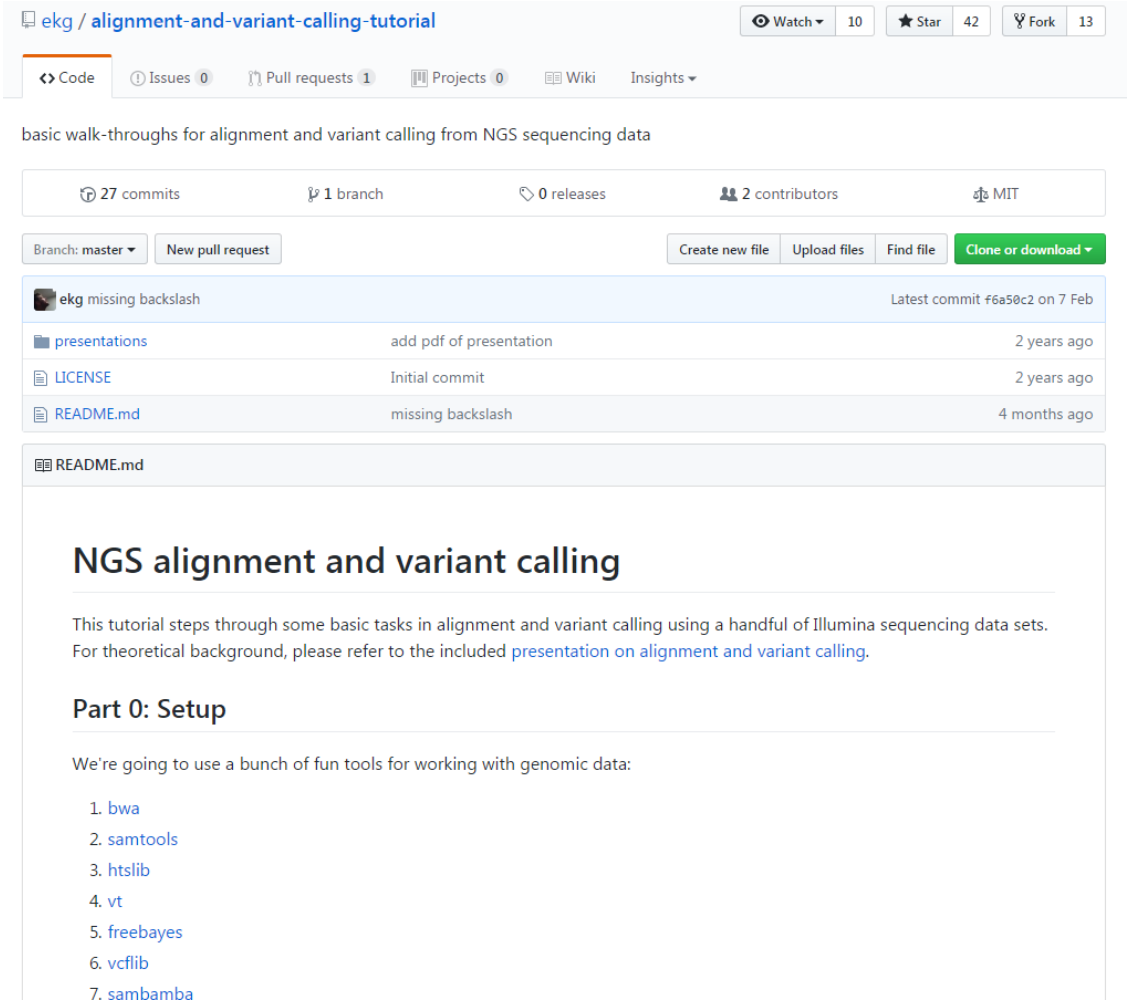
Validation

- 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

Useful information on how-to perform variant calling

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



ekg / alignment-and-variant-calling-tutorial

Watch 10 | Star 42 | Fork 13

Code | Issues 0 | Pull requests 1 | Projects 0 | Wiki | Insights

basic walk-throughs for alignment and variant calling from NGS sequencing data

27 commits | 1 branch | 0 releases | 2 contributors | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

ekg missing backslash — Latest commit f6a50c2 on 7 Feb

presentations	add pdf of presentation	2 years ago
LICENSE	Initial commit	2 years ago
README.md	missing backslash	4 months ago

README.md

NGS alignment and variant calling

This tutorial steps through some basic tasks in alignment and variant calling using a handful of Illumina sequencing data sets. For theoretical background, please refer to the included presentation on alignment and variant calling.

Part 0: Setup

We're going to use a bunch of fun tools for working with genomic data:

1. bwa
2. samtools
3. htslib
4. vt
5. freebayes
6. vcflib
7. sambamba

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?

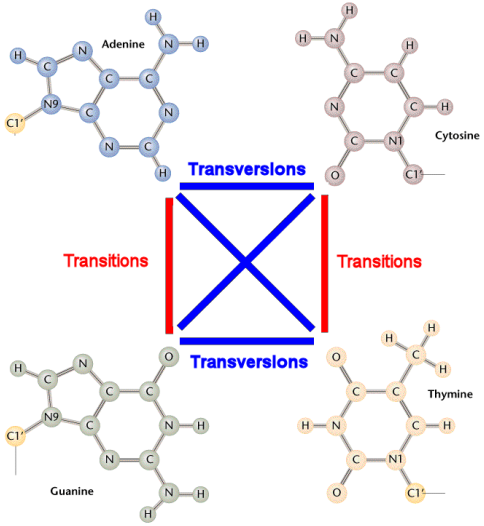
Somatic variant calling

- Find acquired mutations

Ti/Tv ratio

Transition (Ti)

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



Transversion (Tv) purine <-> pyrimidine

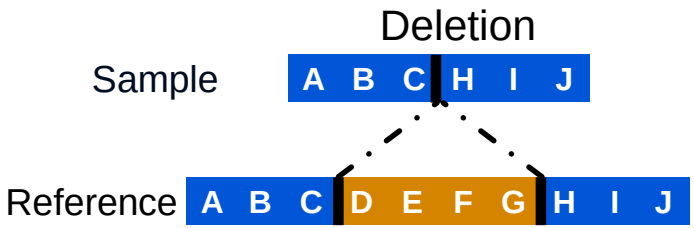
A <->C, A <->T, G <->C, G <->T

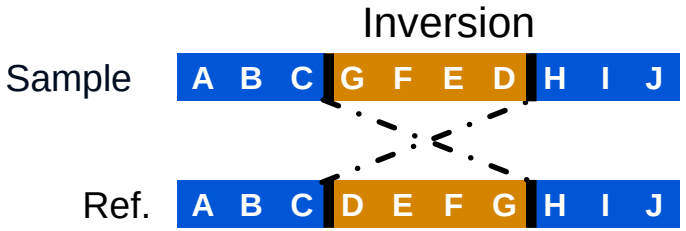
Transition is more frequent than transversion

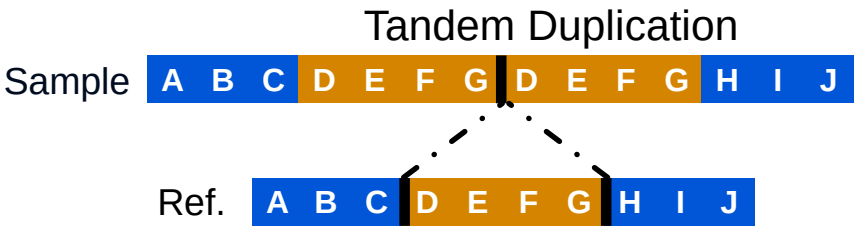
- Ti/Tv ~ 2.0 - 2.1 for genome wide
- Ti/Tv ~ 3.0 - 3.3 for exonic variations
- Ti/Tv = 2/4 = 0.5 = 2/4 =0.5 for random, uniform sequencing - sequencing error

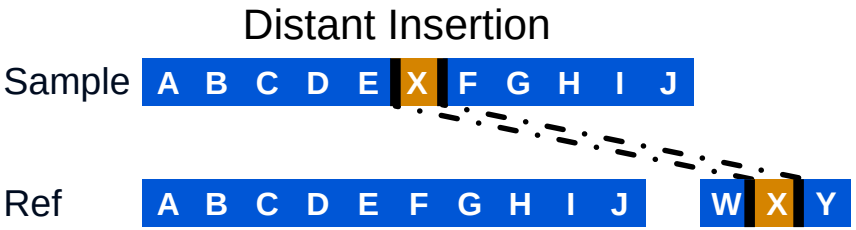
Structural variation calling

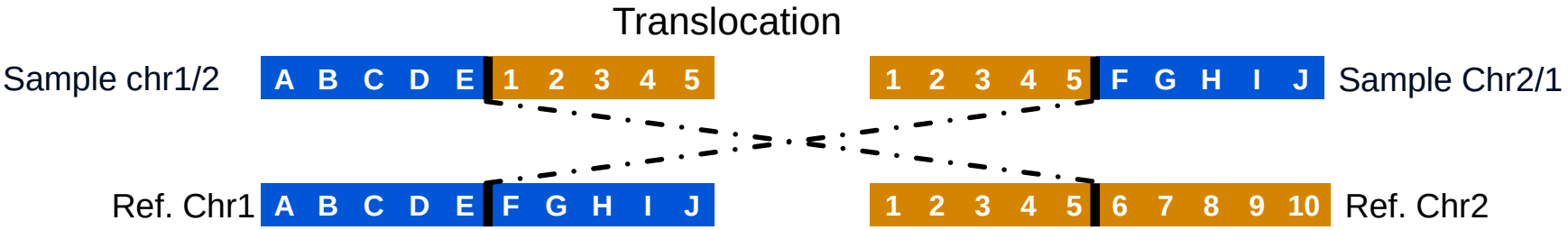
Structural variations











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 and human disease phenotypes

Table 2 Examples of copy number variations (CNVs) and conveyed genomic disorders[a]

Phenotype	OMIM	Locus	CNV
Mendelian (autosomal dominant)[b]			
Williams-Beuren syndrome	194050	7q11.23	del
7q11.23 duplication syndrome	609757	7q11.23	dup
Spinocerebellar ataxia type 20	608687	11q12	dup
Smith-Magenis syndrome	182290	17p11.2/*RAI1*	del
Potocki-Lupski syndrome	610883	17p11.2	dup
HNPP	162500	17p12/*PMP22*	del
CMT1A	118220	17p12/*PMP22*	dup
Miller-Dieker lissencephaly syndrome	247200	17p13.3/*LIS1*	del
Mental retardation	601545	17p13.3/*LIS1*	dup
DGS/VCFS	188400/192430	22q11.2/*TBX1*	del
Microduplication 22q11.2	608363	22q11.2	dup
Adult-onset leukodystrophy	169500	*LMNB1*	dup
Mendelian (autosomal recessive)			
Familial juvenile nephronophthisis	256100	2q13/*NPHP1*	del
Gaucher disease	230800	1q21/*GBA*	del
Pituitary dwarfism	262400	17q24/*GH1*	del
Spinal muscular atrophy	253300	5q13/*SMN1*	del
beta-thalassemia	141900	11p15/*beta-globin*	del
alpha-thalassemia	141750	16p13.3/*HBA*	del

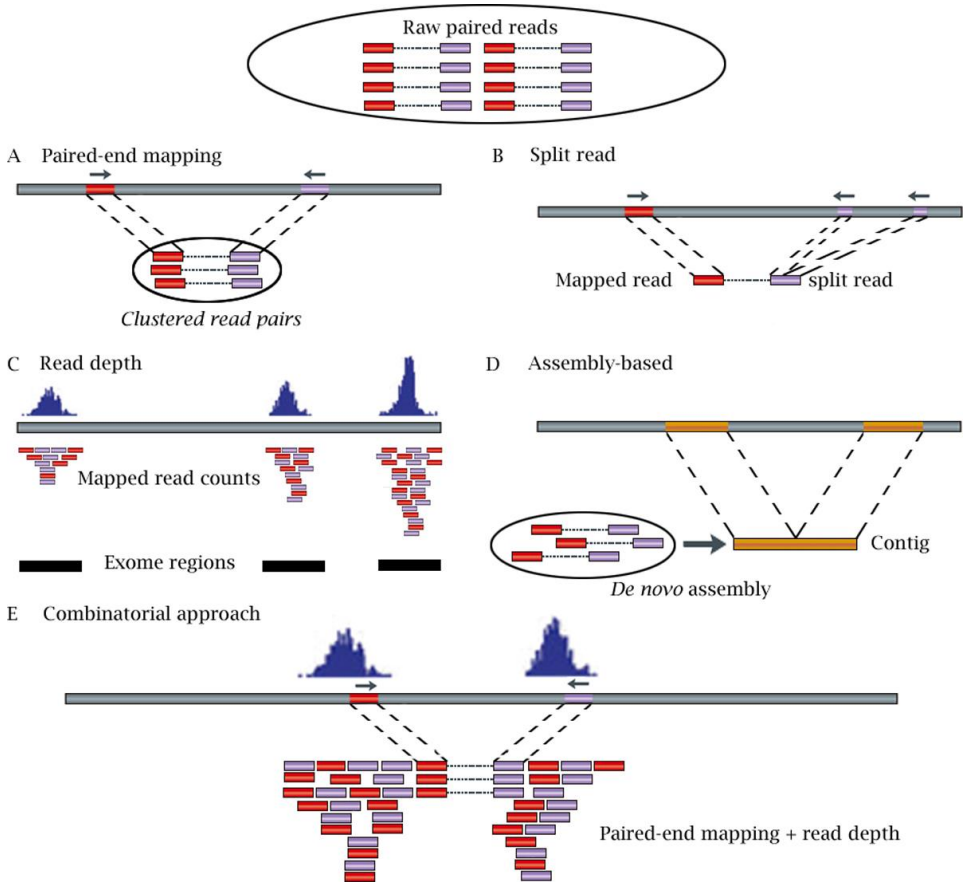
Zhang et al, 2009

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

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.

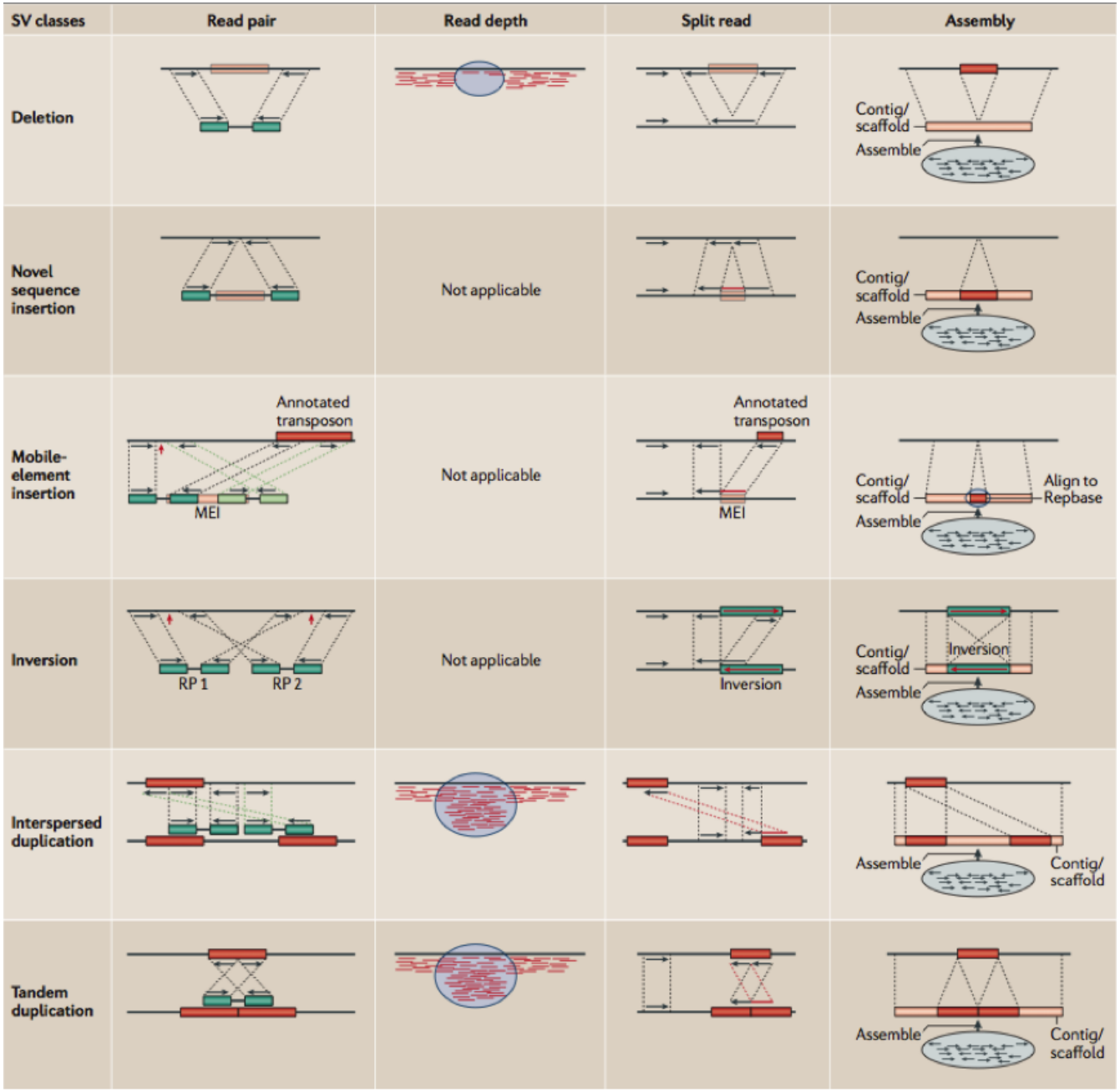
D. Assembly (AS)-based approach detects CNVs by **mapping contigs** to the reference genome.

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 variation - what can we detect



Can et. al.,

Genome structural variation discovery and genotyping, Nature Rev. 2011

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

SVDetect

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

SVMerge

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

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

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

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
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most paired-end mapping studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another

Long Read Technologies

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

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

PACBIO®

Hard to align

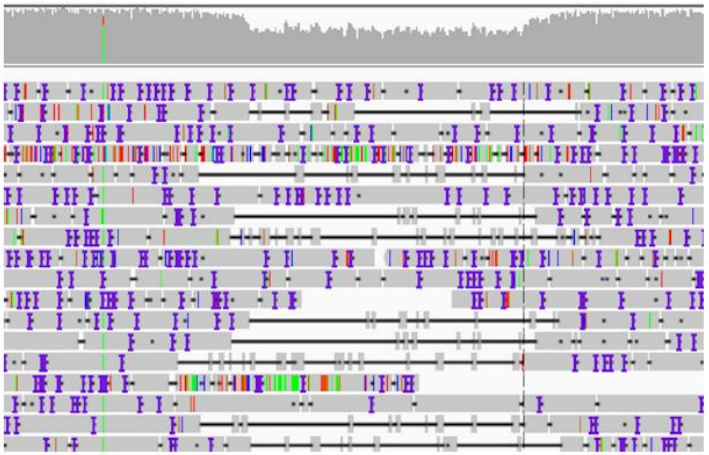
Human genome: 1kb Inversion

Improving long read alignment

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

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

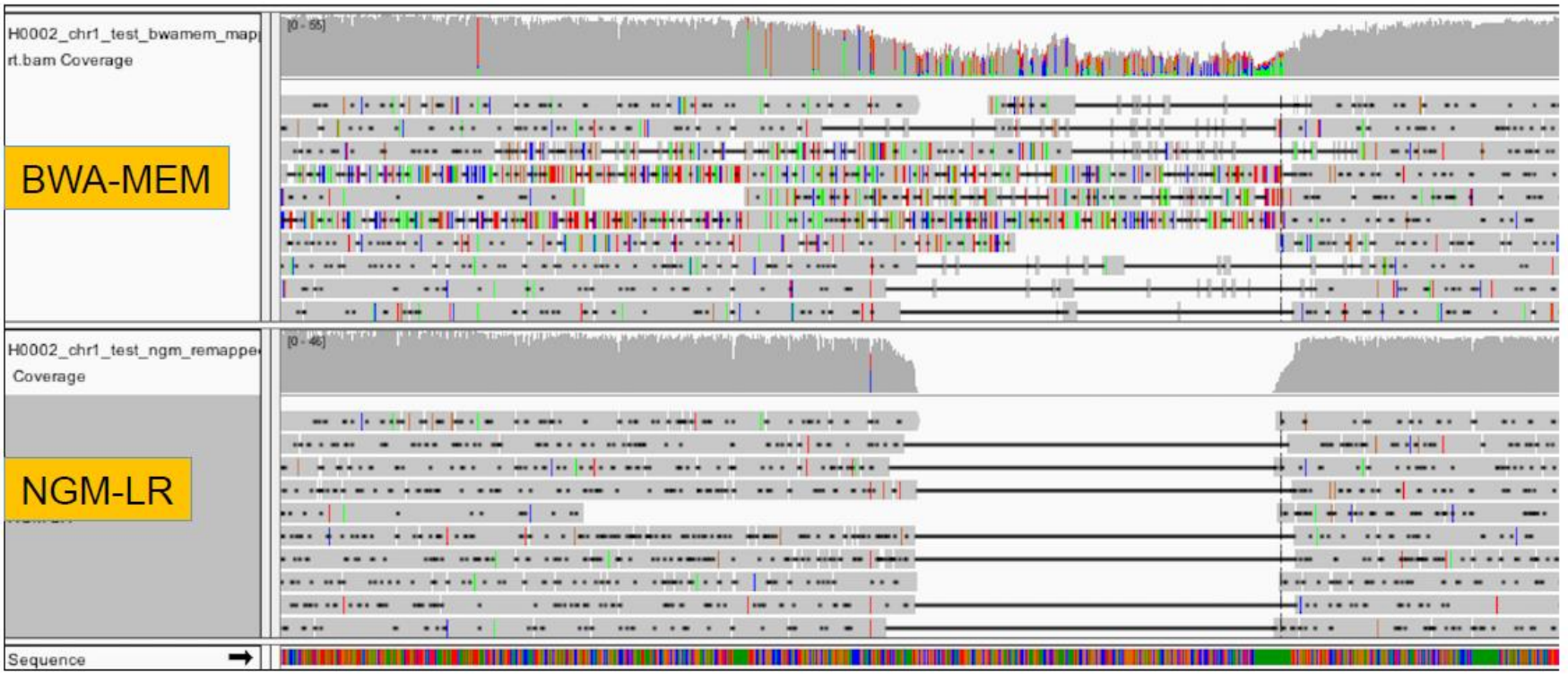
2. Improve alignment:
   - Insertions
   - Deletions



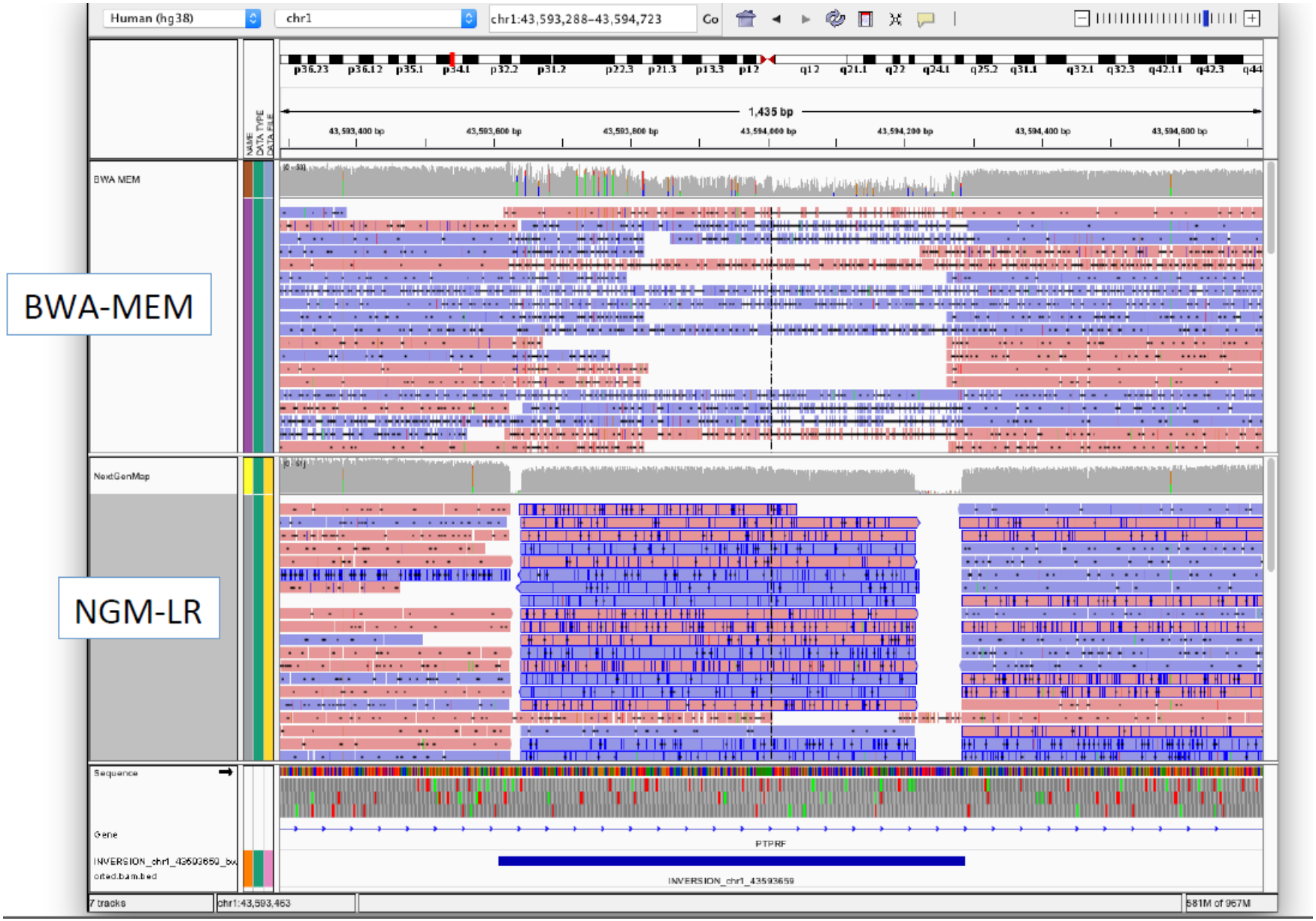
Structural variations with 3rd gen sequencing

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

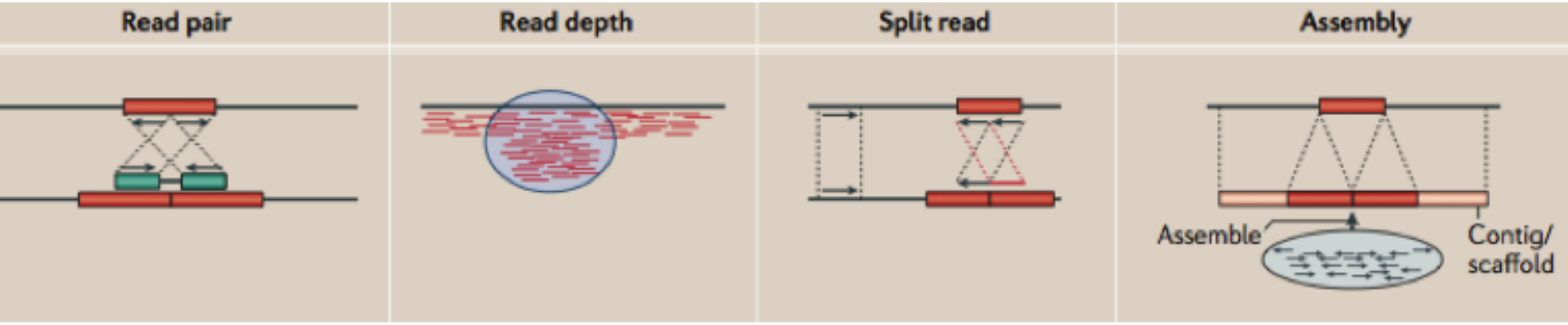


NGM-LR complex SV



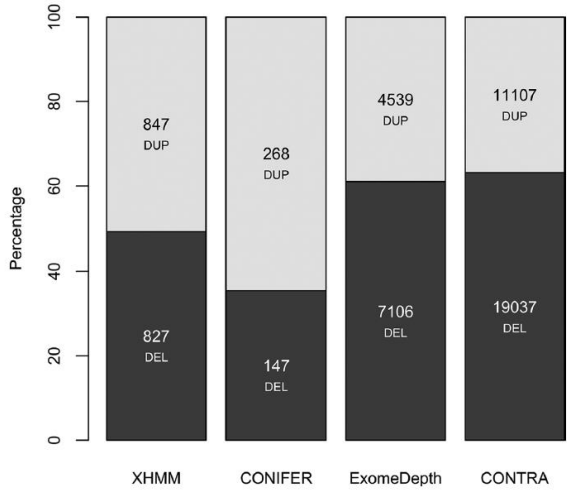
CNV detection

CNVs – how can we detect them



CNV detection

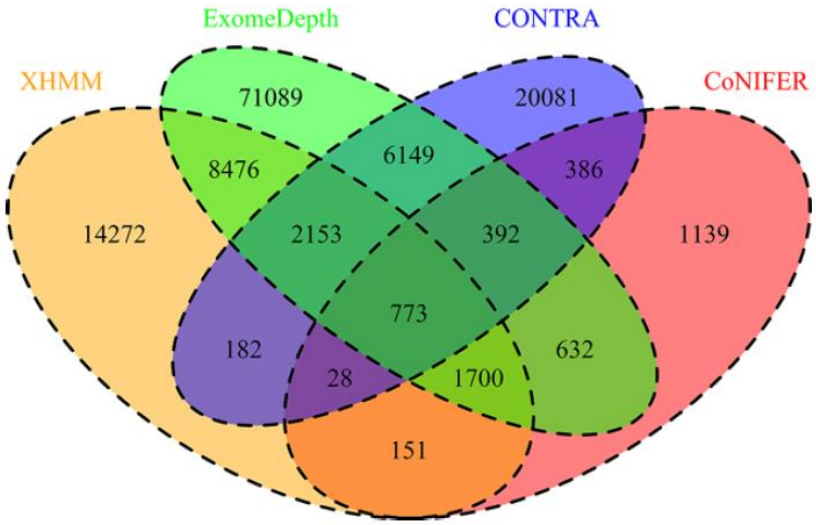
- XHMM
- CoNIFER
- ExomeDepth
- CONTRA



33 individual WES data (cumulatively)
Deletion and duplication CNVs

CNV detection

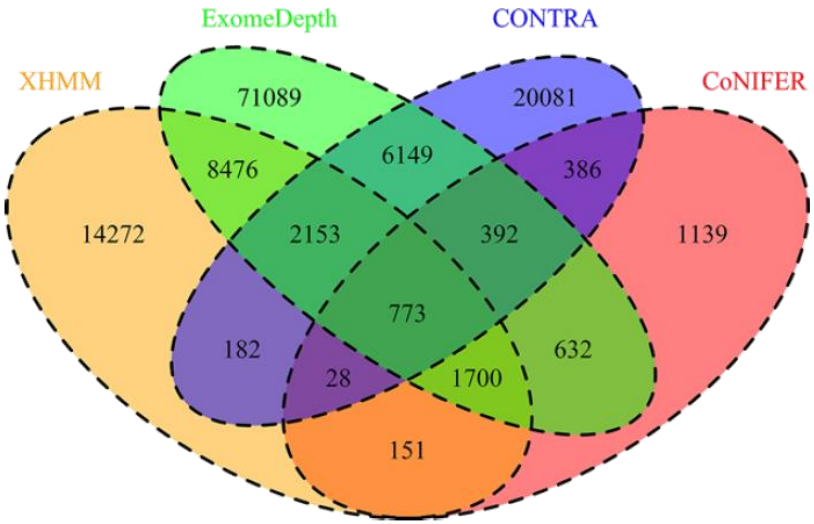
- XHMM
- CoNIFER
- ExomeDepth
- CONTRA



Overlap of CNVs at exon level (33 WES)

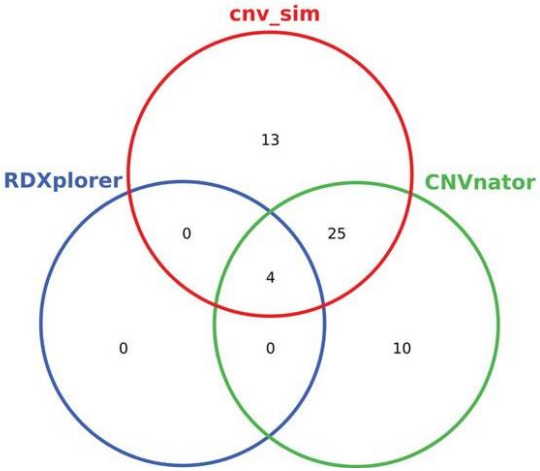
CNV detection

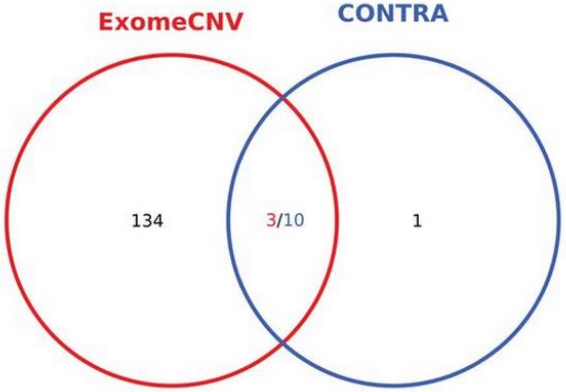
- XHMM
- CoNIFER
- ExomeDepth
- CONTRA



- CONTRA
- ExomeCNV

- CNVnator
- RDXplorer
- Simulated





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

Structural variations - databases

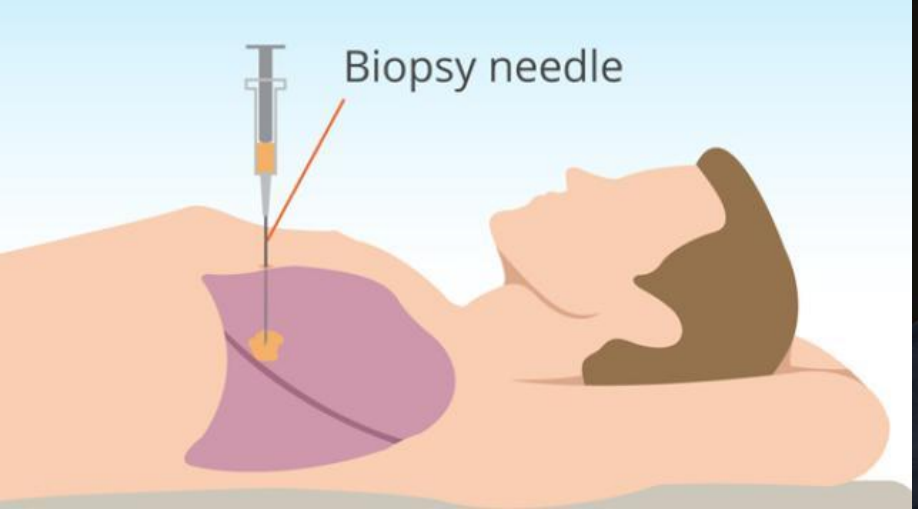
dbVar

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

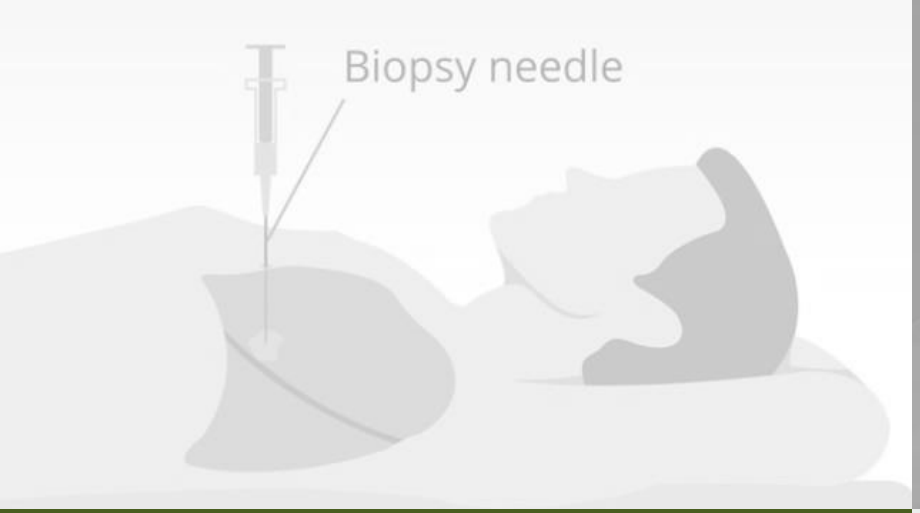
Database of Genomic Variants archive

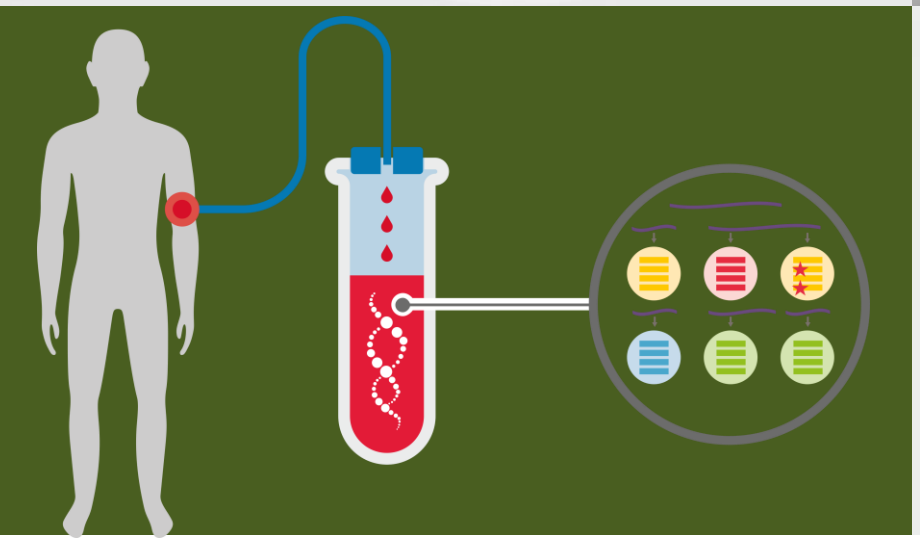
- Repository that provides archiving, accessioning and distribution
- Available in all species
- http://www.ebi.ac.uk/dgva

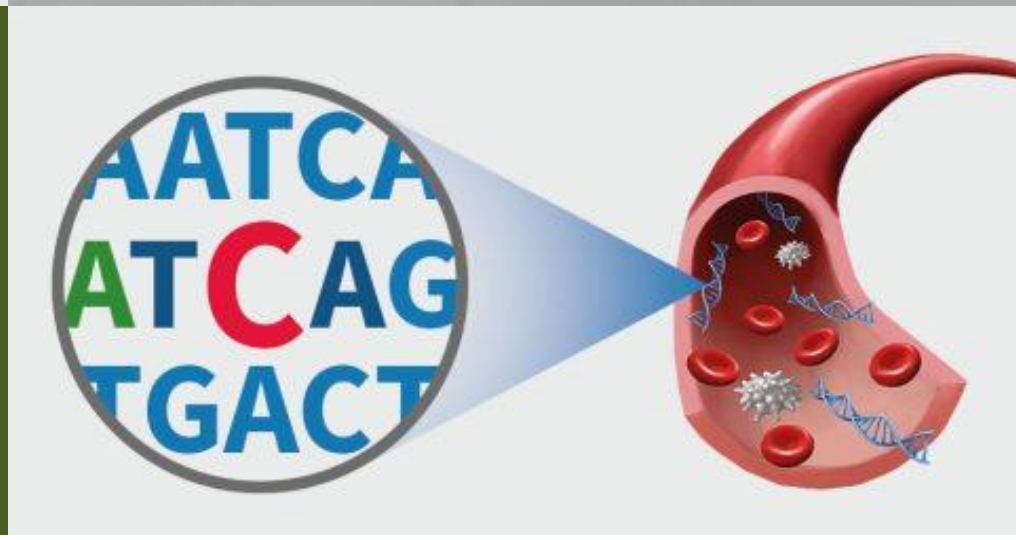
Liquid biopsies – CNV detection



http://medifitbiologicals.com/wp-content/uploads/2015/11/BIOPSY-2.jpg
https://fthmb.tqn.com/WQWRqsQfxwQKOCgZXALeLoYcTow=/768x0/filters:no_upscale()/about/iStock_000001373663_Large-56a5c6145f9b58b7d0de6b32.jpg

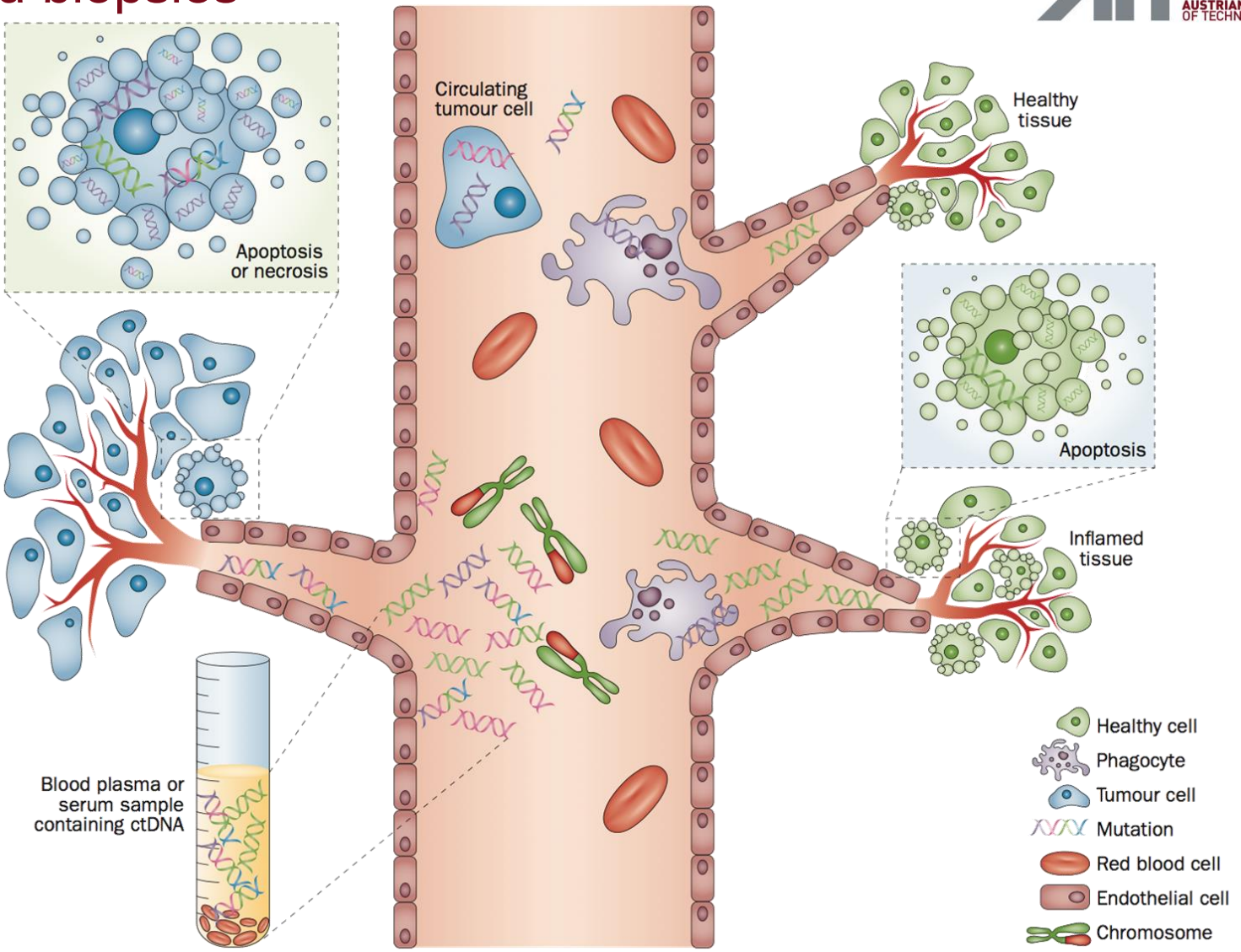






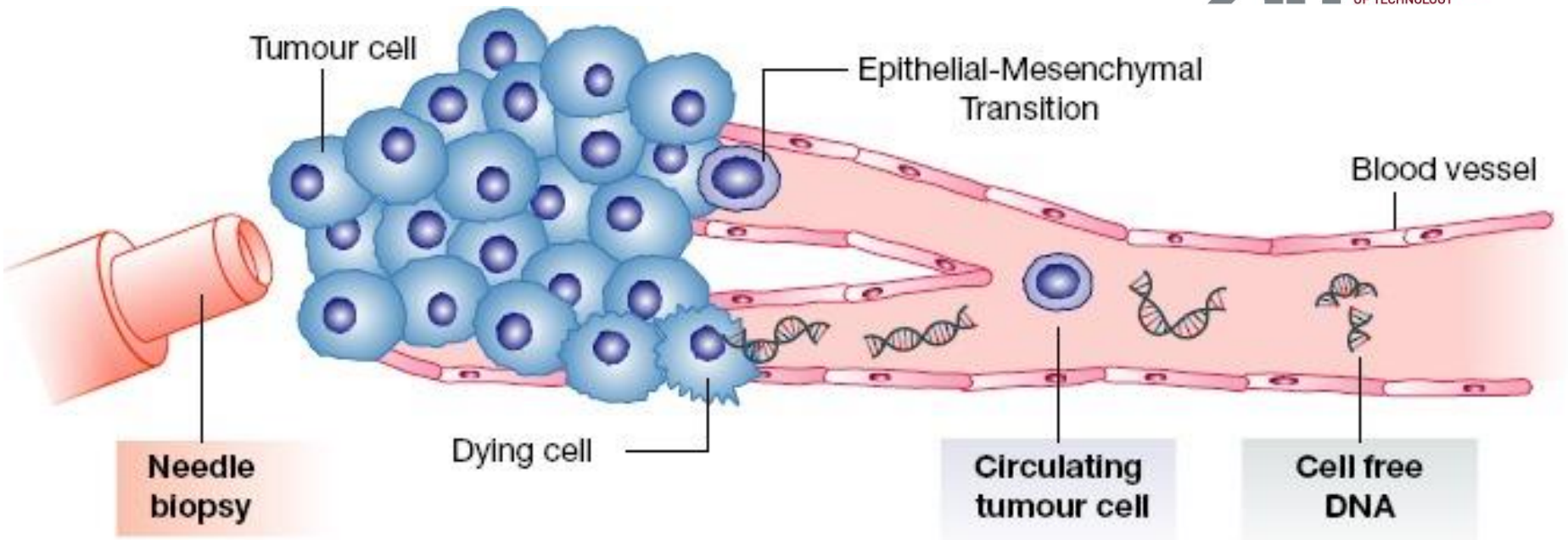
http://medifitbiologicals.com/wp-content/uploads/2015/11/BIOPSY-2.jpg
https://fthmb.tqn.com/WQWRqsQfxwQKOCgZXALeLoYcTow=/768x0/filters:no_upscale()/about/iStock_000001373663_Large-56a5c6145f9b58b7d0de6b32.jpg
http://www.gene-quantification.de/wyatt-gleave-liquid-biopsy-2105.jpg

Liquid biopsies



Crowley E, et al. (2013) Liquid biopsy: monitoring cancer genetics in the blood. *Nat Rev Clin Onc* 10: 472–484.

Liquid biopsies



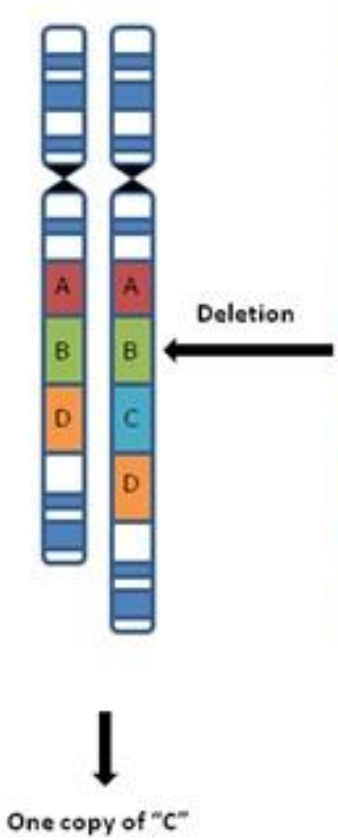
	Biopsy	CTC	cfDNA
Invasive	+	–	–
All patients eligible	–	+	+
Instrumentation required	+	+	–
Biomarker applicability	–	++	+++

WGA = Whole-genome amplification

Use of liquid biopsies to monitor disease progression in a sarcoma patient: a case report. BMC Cancer
Targeting the adaptive molecular landscape of castration-resistant prostate cancer. EMBO Mol Med. 2015

Copy Number Variation - CNV

Sections of the genome are repeated and the number of repeats in the genome varies between individuals

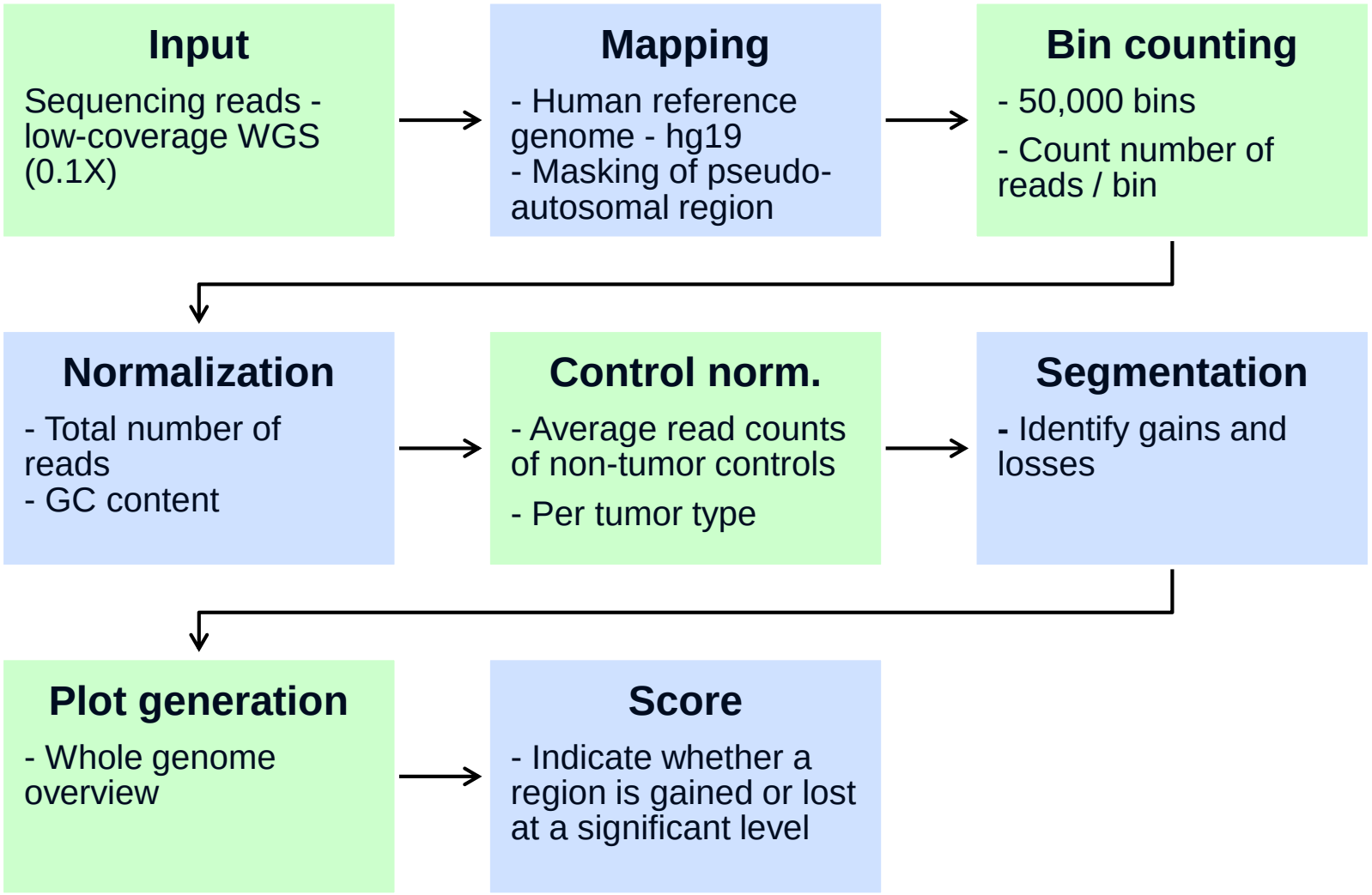


CNVs associated with complex diseases

DEFB4, DEFB103, DEFB104	Crohn's disease	Amplification[a,b]	Low copy number
CCL3L1	HIV/AIDS	Amplification[a,b]	Low copy number
C4	Lupus	Amplification[a,b]	Low copy number
FCGR3B	Glomerulonephritis in Lupus patients	Amplification[a,b]	Low copy number
FCGR3B	Lupus	Amplification[a,b]	Low copy number
IRGM	Crohn's disease	Upstream deletion	Deletion
CFHR1, CFHR3	Age-related macular degeneration	Deletion[a]	No deletion
CYP2D6	Reduced drug metabolism	Deletion[a]	Deletion
RHD	Rh-negative blood group	Deletion[a]	Deletion
OPN1LW, OPN1MW	Color blindness	Deletion	Deletion
LPA	Coronary heart disease	Amplification[a,b]	Low copy number
SMN2	Severity of spinal muscular atrophy	Amplification[a,b]	Low copy number
AZFc region	Spermatogenetic failure	Deletion[a]	Deletion
CYP21A2	Congenital adrenal hyperplasia	Amplification[a,b]	Two copies/chrom
UGT2B17	Osteoporosis	Deletion[b]	No deletion
UGT2B17	Graft-versus-host disease	Deletion[b]	Deletion
LCE3B, LCE3C	Psoriasis	Deletion	Deletion
NEGR1	Obesity	Upstream deletion	Deletion
NBPF23	Neuroblastoma	Deletion[a,b]	Deletion
TSPAN8	Type 2 diabetes	Amplification[a,b]	Low copy number

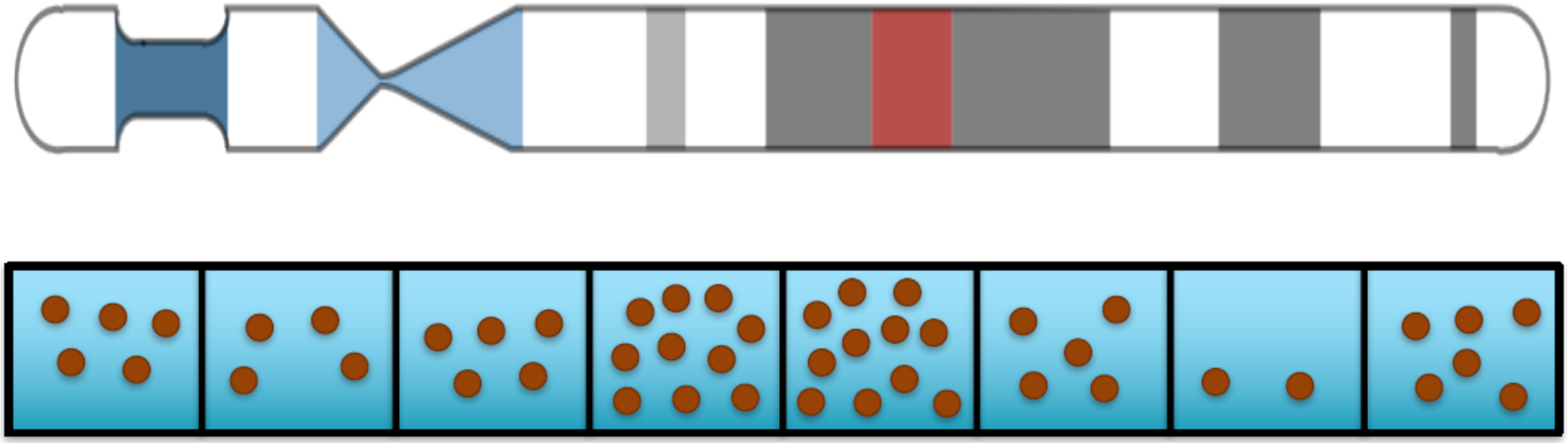
http://readingroom.mindspec.org/wp-content/genetics_CNV.jpg
Human Copy Number Variation and Complex Genetic Disease. Annual Review of Genetics. 2011

CNV Analysis



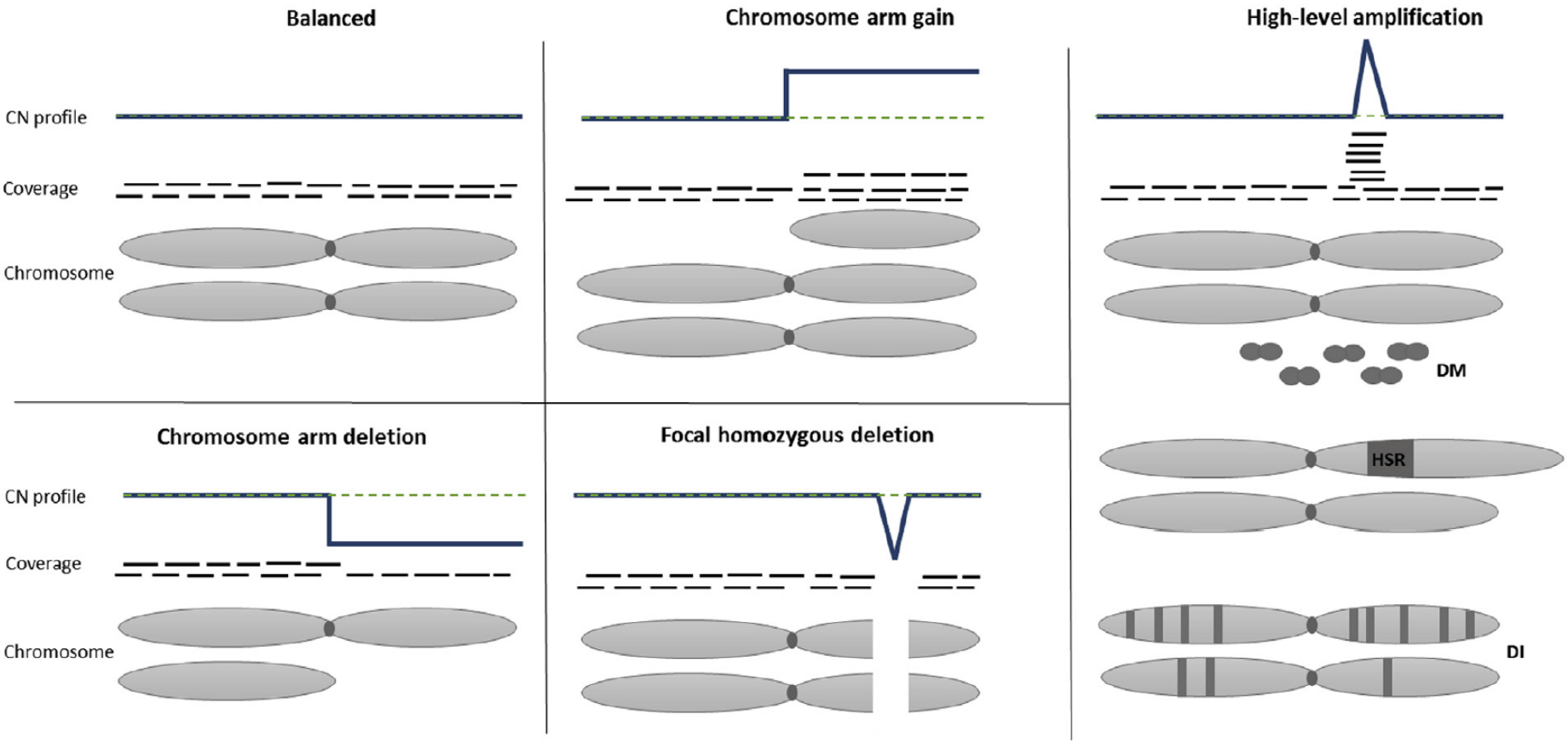
Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. Nat com
Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. Genome Medicine

CNV Analysis - Overview



- Divide genome into bins
- Map reads
- Count reads
- Normalization
- Segmentation

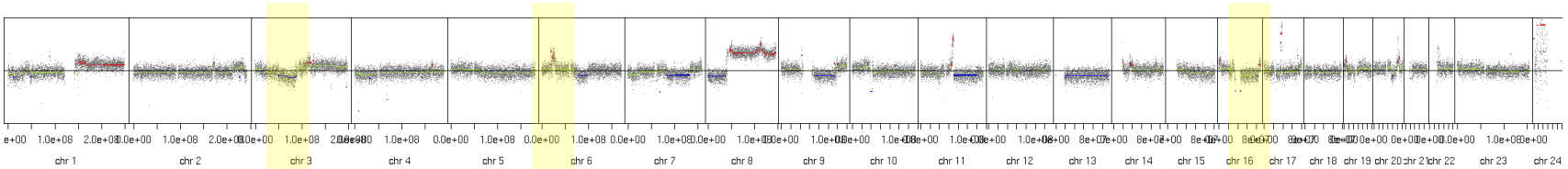
What can we detect



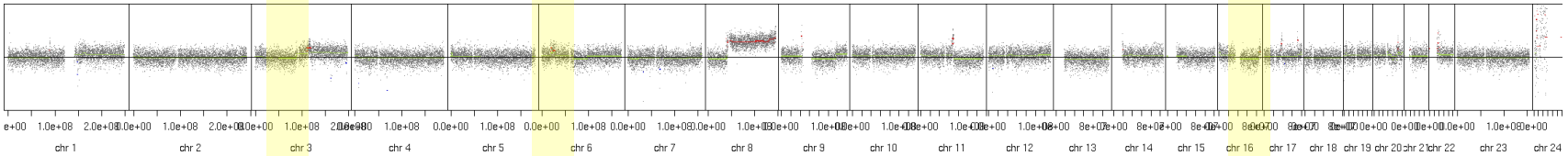
Pipeline result

Breast cancer patient (F)

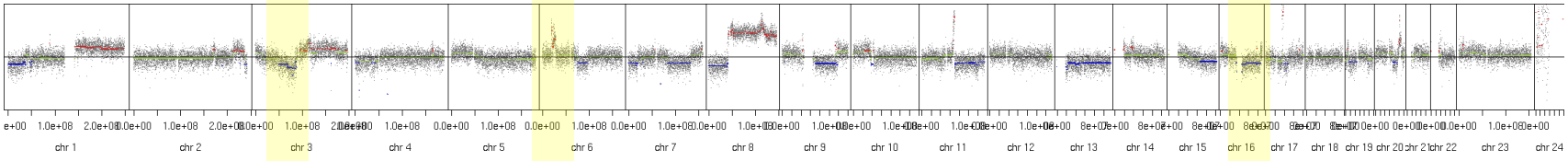
Timepoint 1



Timepoint 2



Timepoint 3



Practicals – Day 3