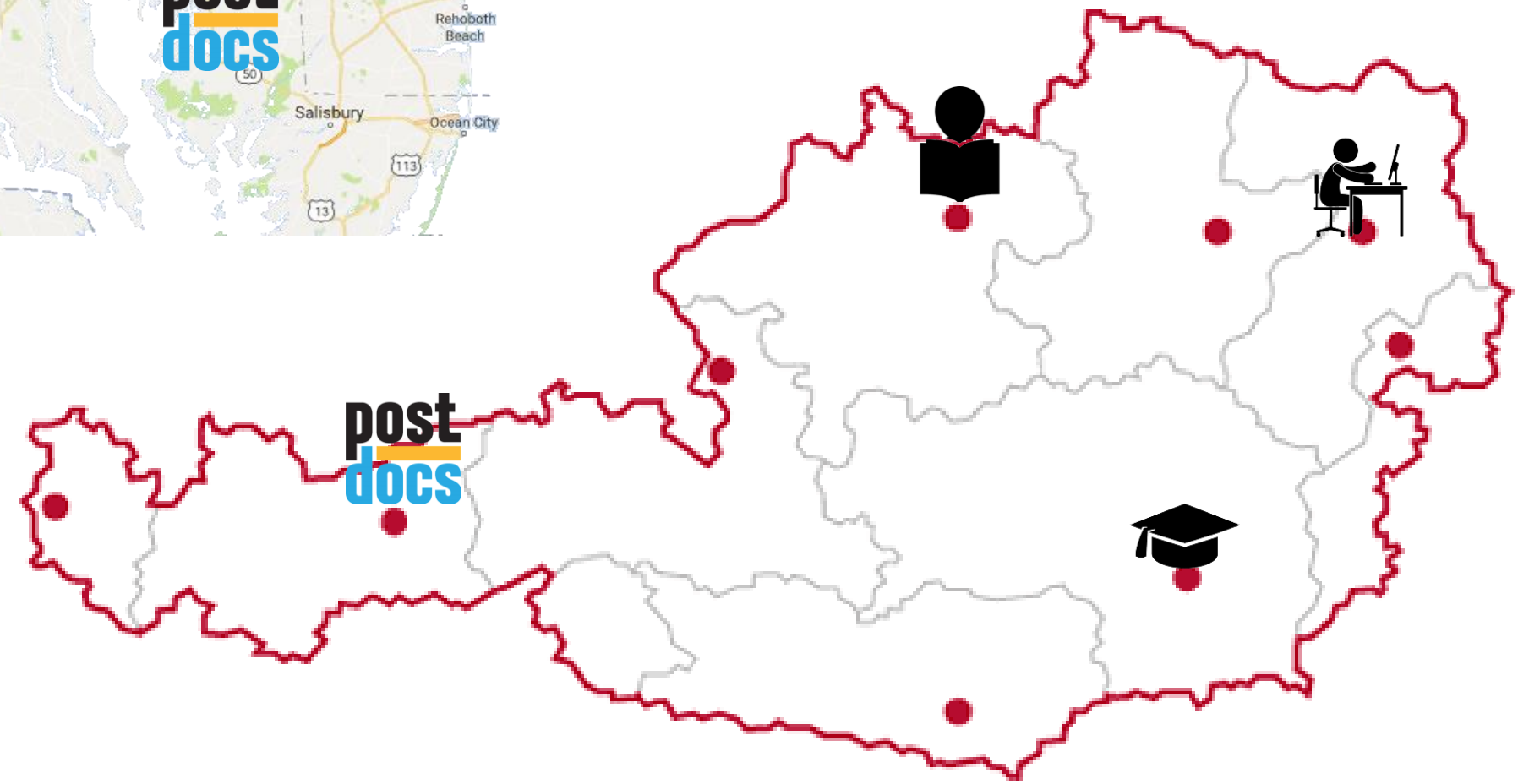
AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Tools for variant analysis of Next Generation Genome Sequencing data

Lecture

Stephan Pabinger
stephan.pabinger@ait.ac.at

My background



My background

Bioinformatics

- Software development
- Web tools
- Pipeline design

Working with sequencing data

- DNASeq
- RNASeq
- MethylationSeq

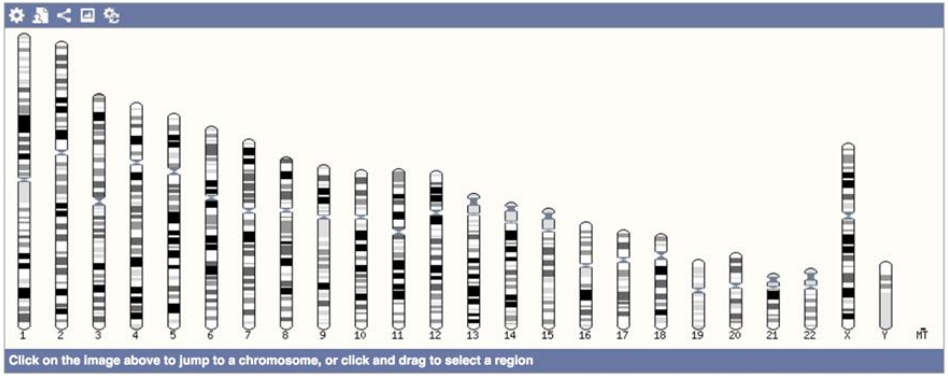
What to expect

- Finish with an understanding of major concepts and tools
- Know how to perform variant calling
- Ready to make informed choices about what kind of variant calling tools you may need
- Focus is on Variant Calling and Functional Annotation
  - Alignment refinement
  - Alignment metrics reports
  - Base quality score recalibration
  - Variant calling
  - Variant annotation and filtering

Terms

The human genome - basic stats

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



Summary

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

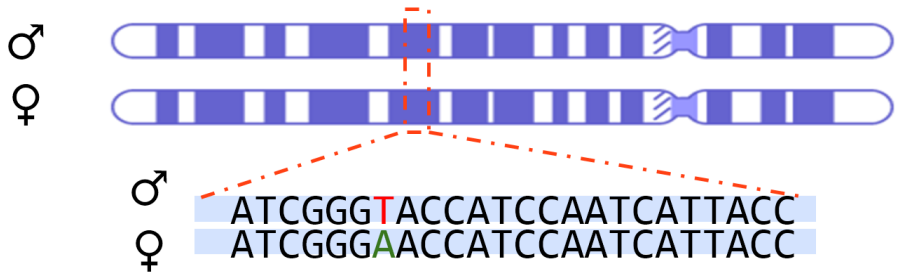
Gene counts (Primary assembly)

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

http://uswest.ensembl.org/Homo_sapiens/Location/Genome

Genetics Terms

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



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

Trait: A distinguishing feature, a genetically determined characteristic or condition

Genotype: Genetic makeup, distinguished from the physical appearance

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

Forms of genetic variations

Single nucleotide substitution
Replacement of one nucleotide with another

Tandem repeat
ATTCG ATTCG ATTCG

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

Deletions or insertions
Loss or addition of one or more nucleotides

Structural variations
Changes in chromosome number, segmental rearrangements and deletions

Genetics Terms

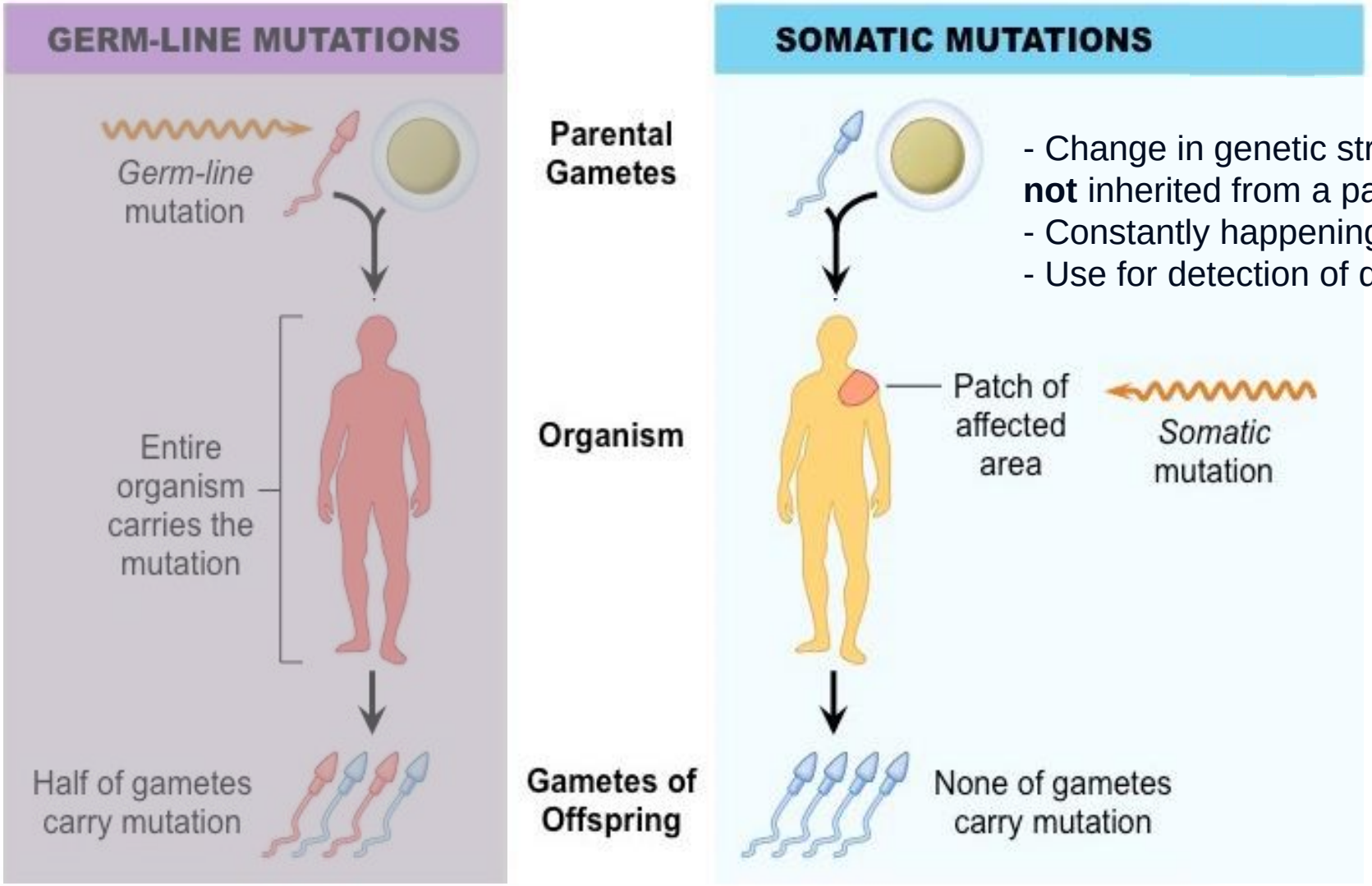
Polymorphism

- Variations in DNA sequence (substitutions, deletions, insertion, etc) that are present at a frequency greater than 1% in a population.
- Have a WEAK EFFECT or NO EFFECT at all.
- Ancient and COMMON

Mutation

- Variations in DNA sequence (substitutions, deletions, etc) that are present at a frequency lower than 1% in a population.
- Can produce a gain of function and a loss of function.
- Recent and RARE.

Somatic mutations



- Change in genetic structure **not** inherited from a parent
- Constantly happening
- Use for detection of disease

Glossary and Definitions

Homozygote
An organism with two identical alleles

Heterozygote
An organism with two different alleles

Hemizygote
Having only one copy of a gene
→ males are hemizygous for most genes on the sex chromosomes

Dominant trait
A trait that shows in a heterozygote

Recessive trait
A trait that is hidden in a heterozygote

Coordinate systems

Coordinate system

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

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

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

AUSTRIAN INSTITUTE OF TECHNOLOGY

Genome reference

Genome reference

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

UCSC and Genome Reference Consortium (GRCh)

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

Newest version (hg38) release on Dez 24th 2013

Download (e.g.)

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

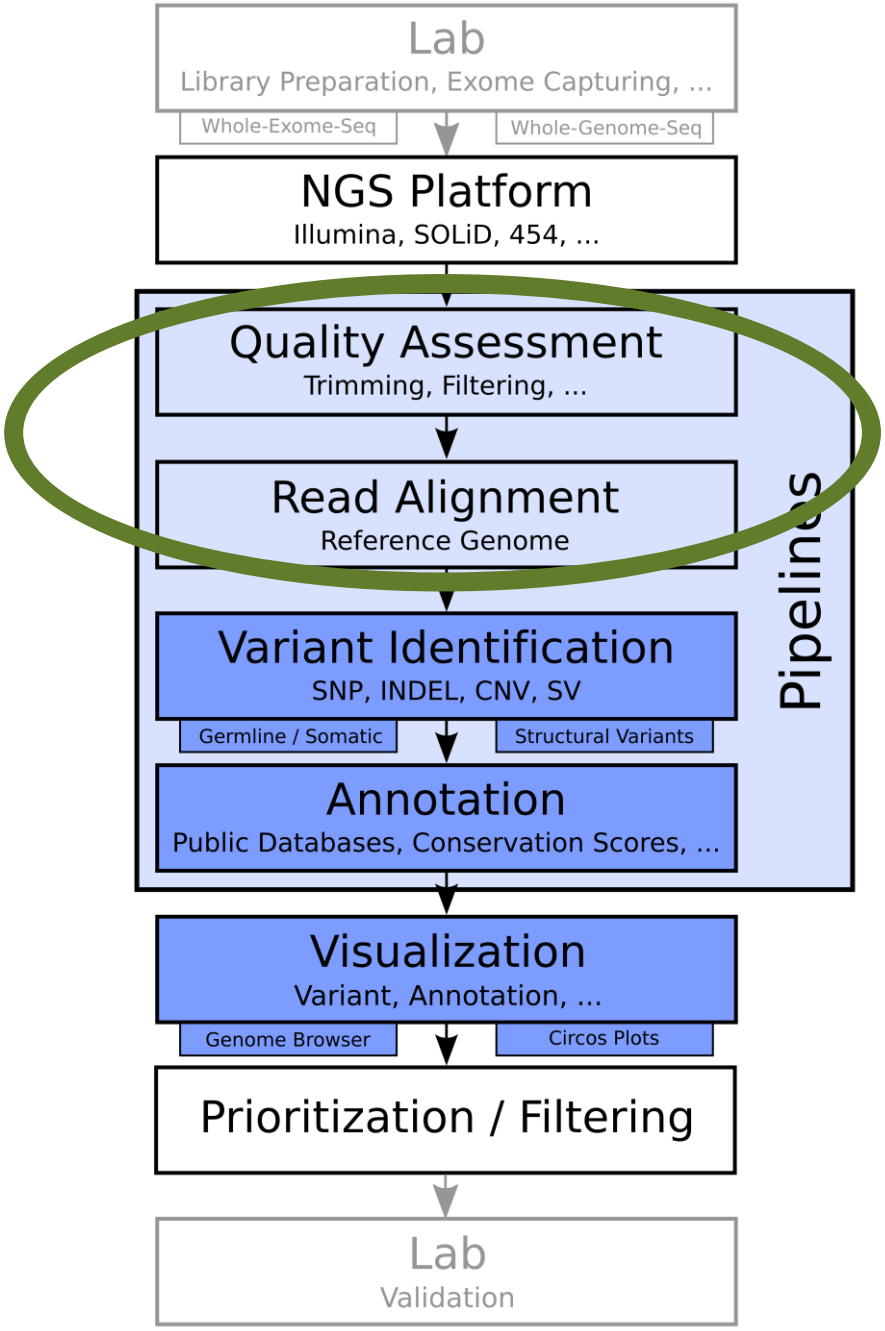
- Fai file (created by samtools faidx)
  contig, size, location, bases-per-line and for efficient random
- Dict file (created by Picard CreateSequenceDictionary)
  SAM style header describing the contents of the fasta file
- Indices from different mapping programs

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

Workflow overview


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

Informative NGS QC page

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

- PRINSEQ: http://prinseq.sourceforge.net/
  HTML output, trimming, filtering, contaminations

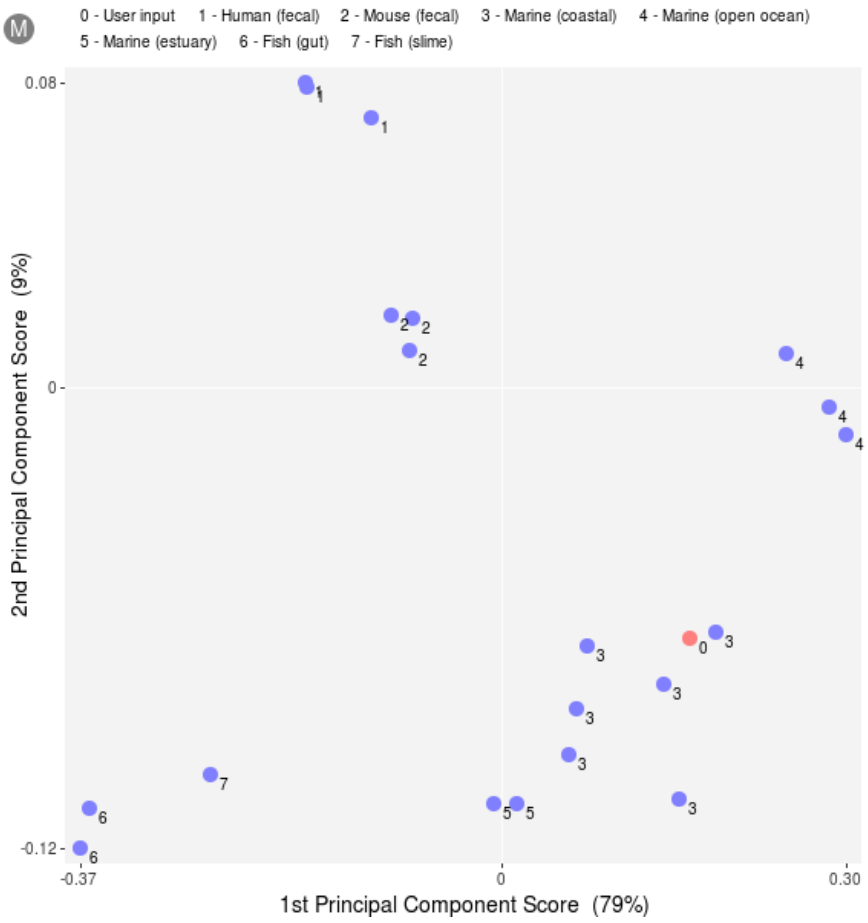
- Trimmomatic http://www.usadellab.org/cms/?page=trimmomatic
  Paird-end trimming

http://www.molecularecologist.com/2017/01/handling-microbial-contamination-in-ngs-data/#more-9579

Check for contamination

PRINSEQ

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

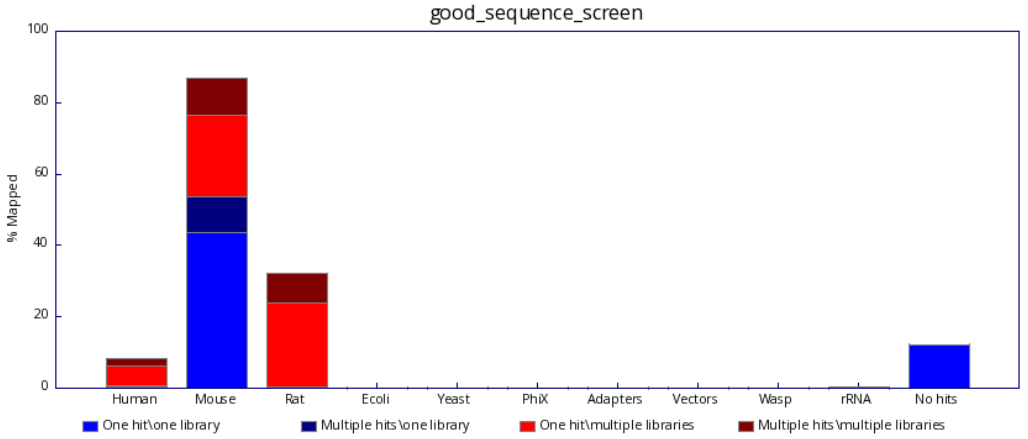


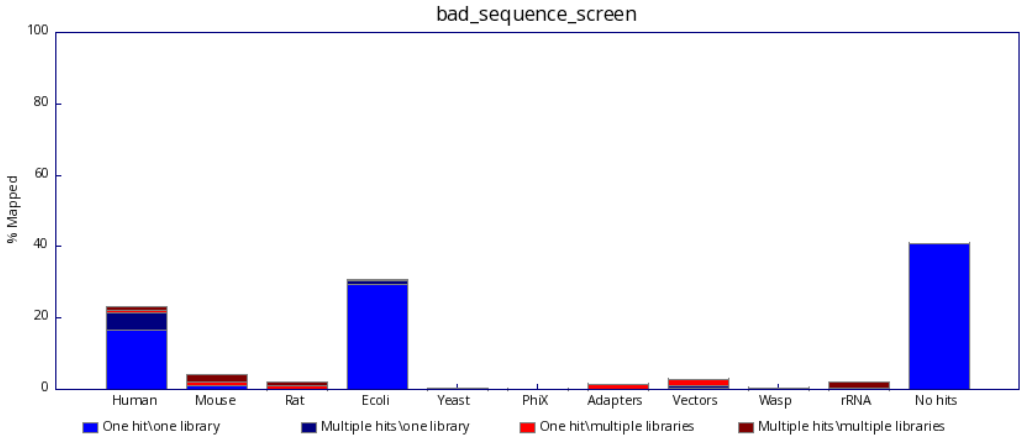
Check for contamination

FastQ Screen

Screen against a set of sequence databases:

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

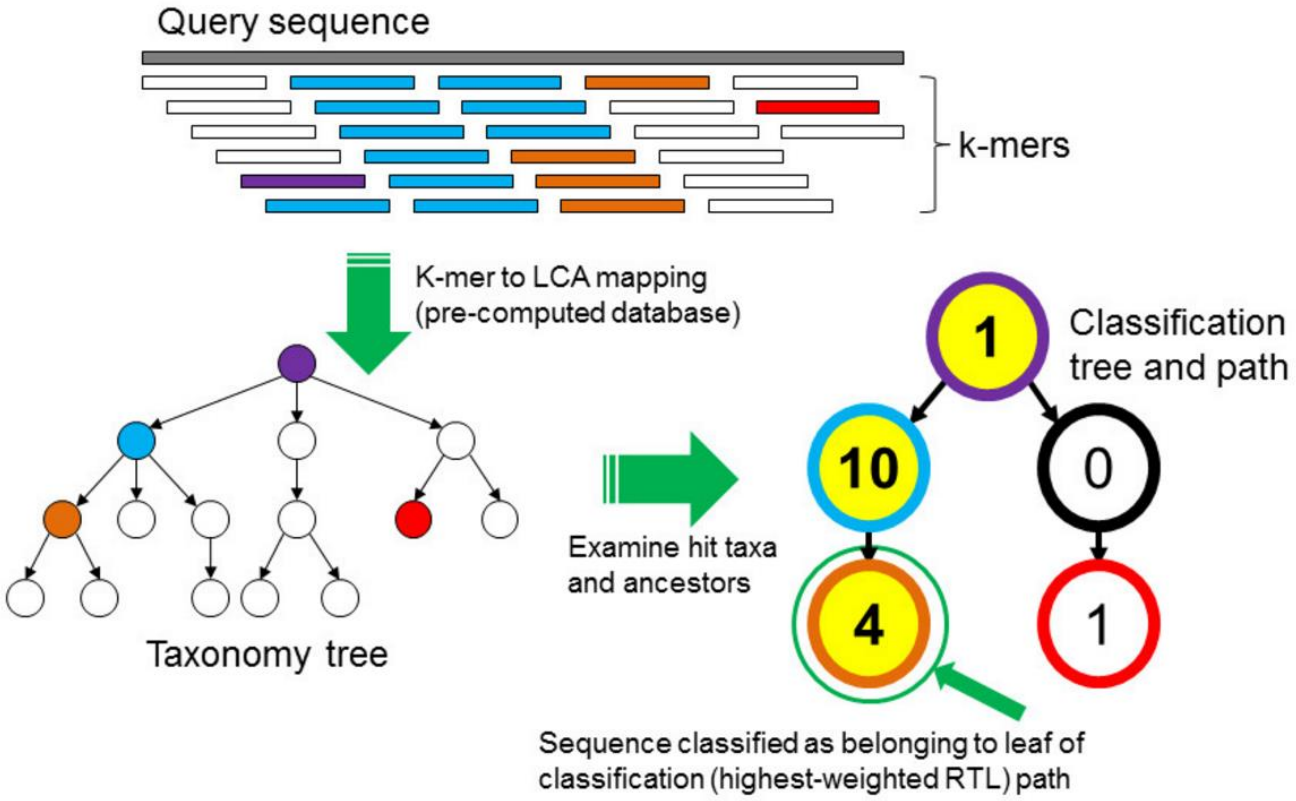




Check for contamination

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

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

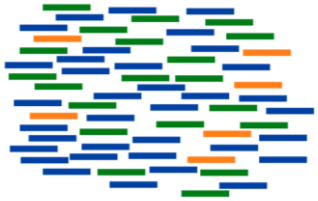


Mapping and QC

SAM - Sequence Alignment/Map Format

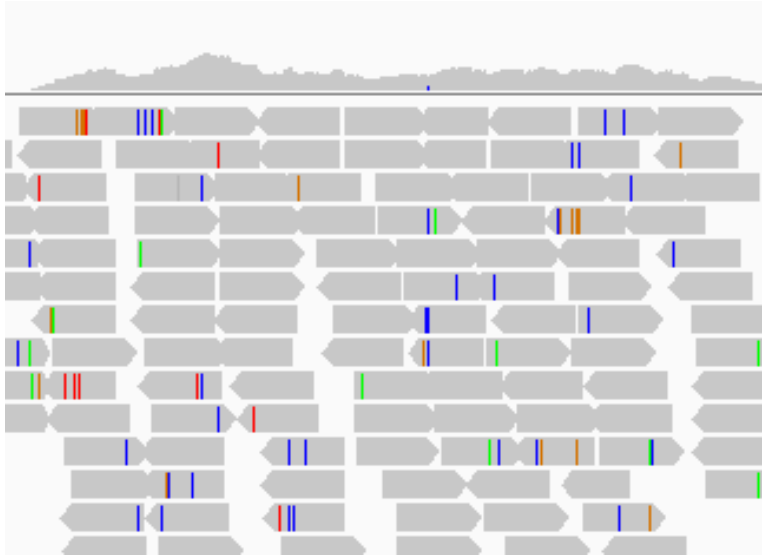
Mapping - Principle

Short sequence reads



Alignment

Reference



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

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

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

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

Flow:

Sequencing Reads → Alignment → SAM file

6 CIGAR

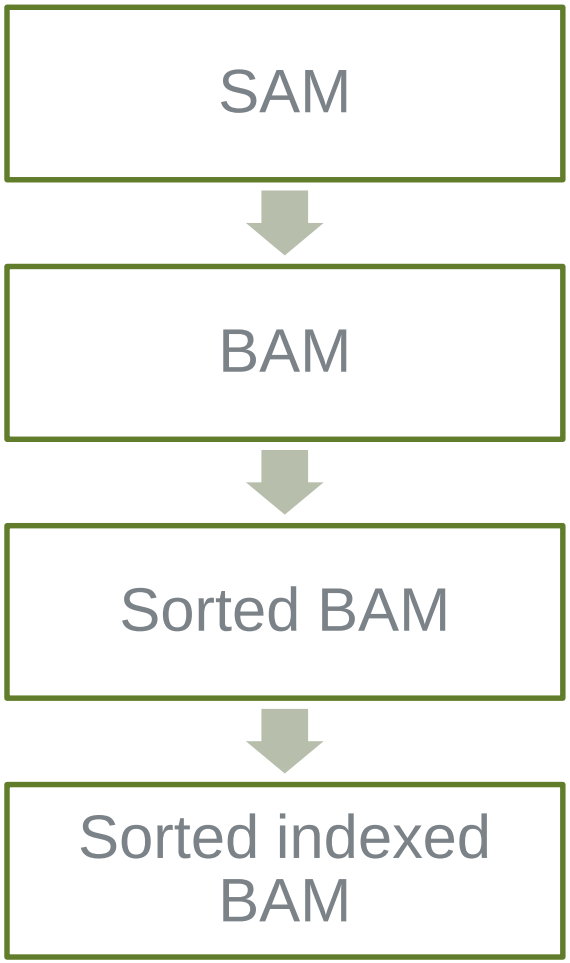
Used as a compact way to represent sequence alignment

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

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

uBAM

- unmapped BAM → compress FASTQ files

CRAM

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

Quality check of alignment

Based on SAM/BAM files

Detect biases in the sequencing and/or mapping

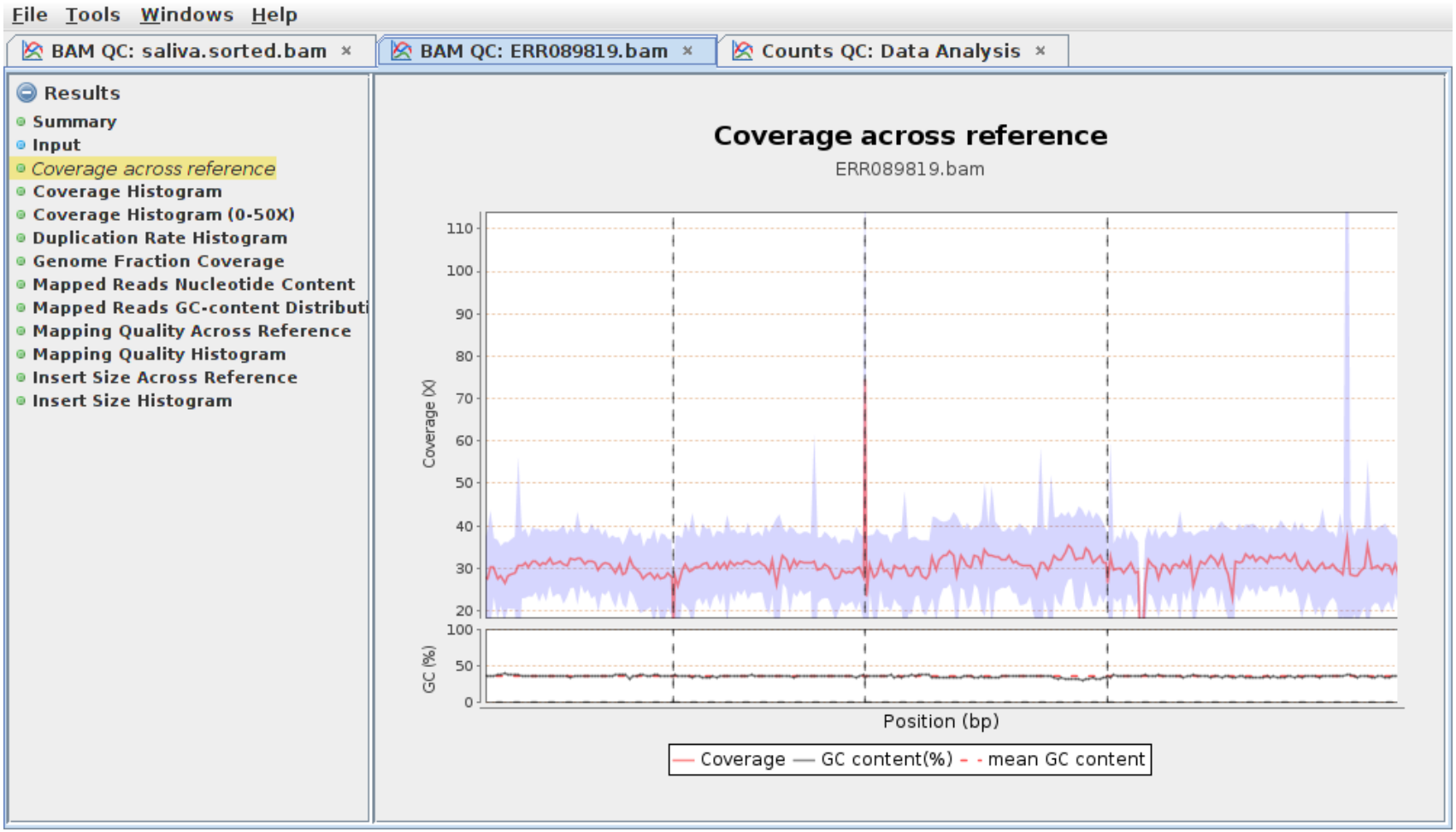
Metrics

- 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/

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

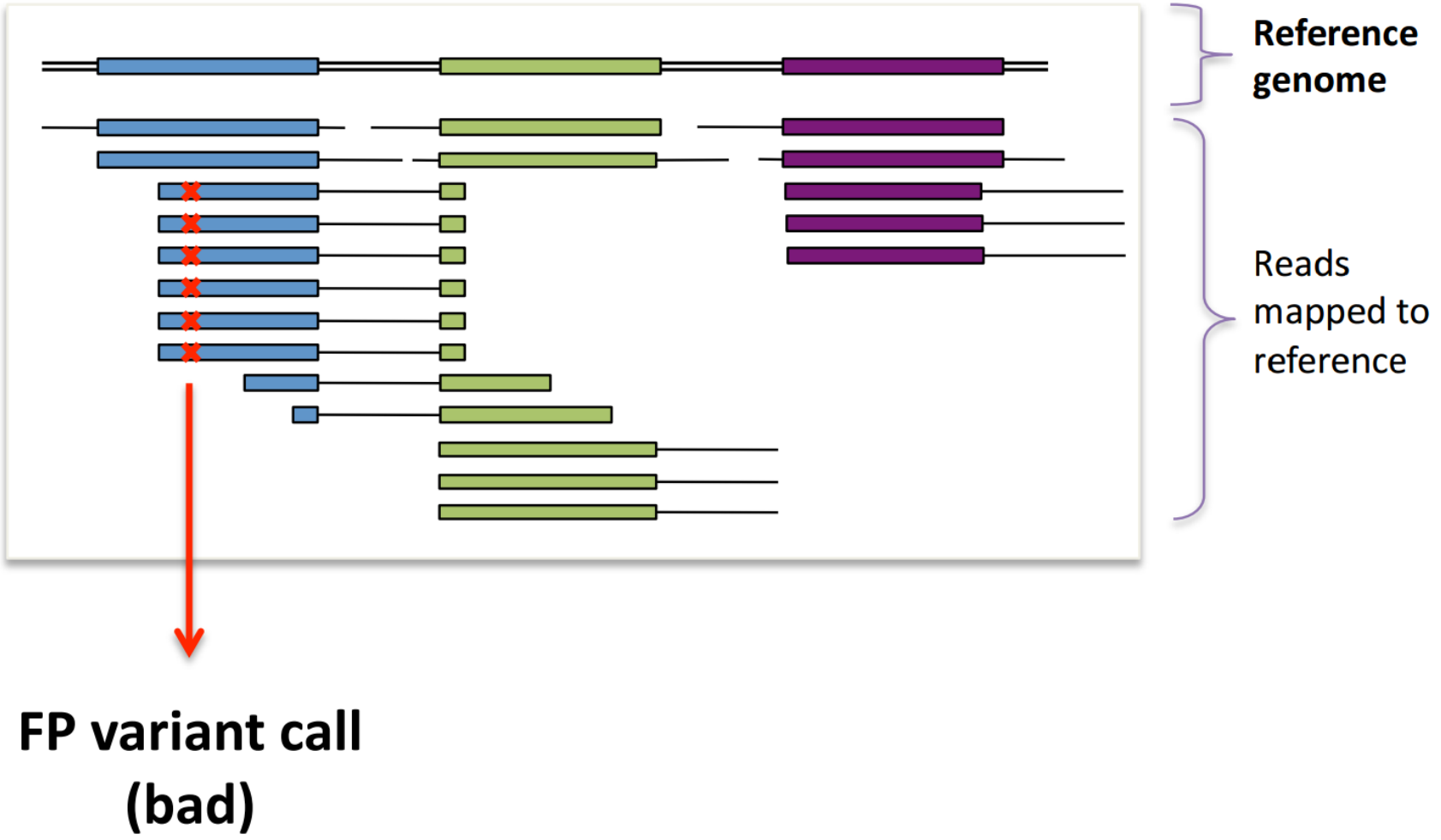
Read duplicates

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

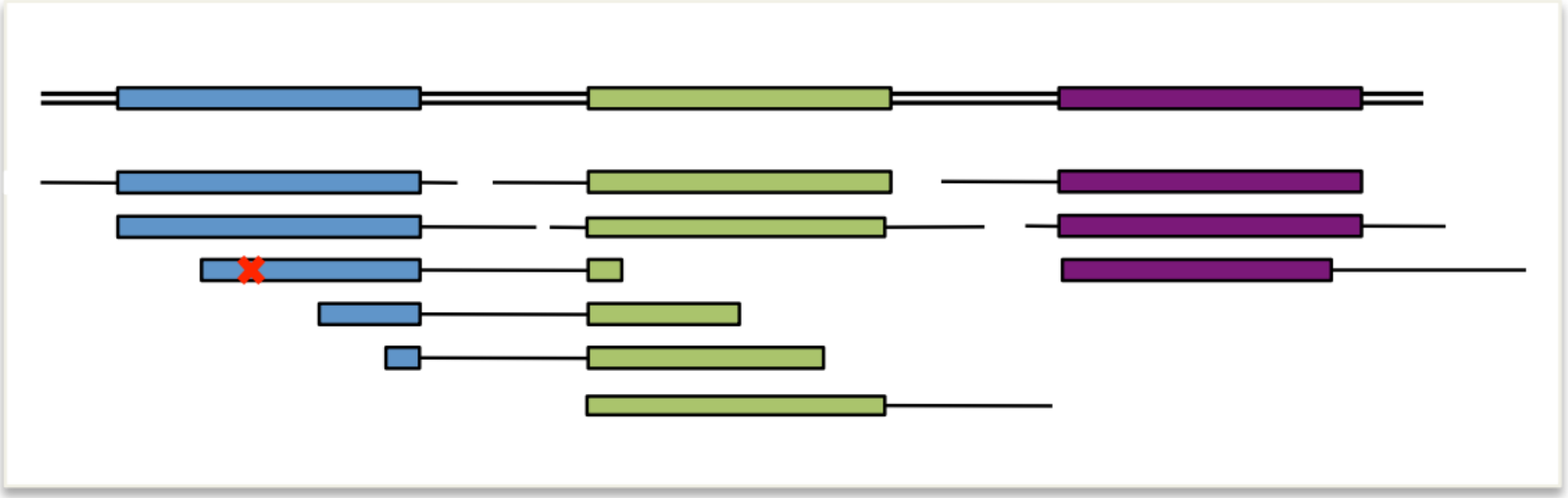


http://www.broadinstitute.org/gatk/events/2038/GATKwh0-BP-1-Map_and_Dedup.pdf

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

Working with SAM/BAM files

More tools

Other tools

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

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

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

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

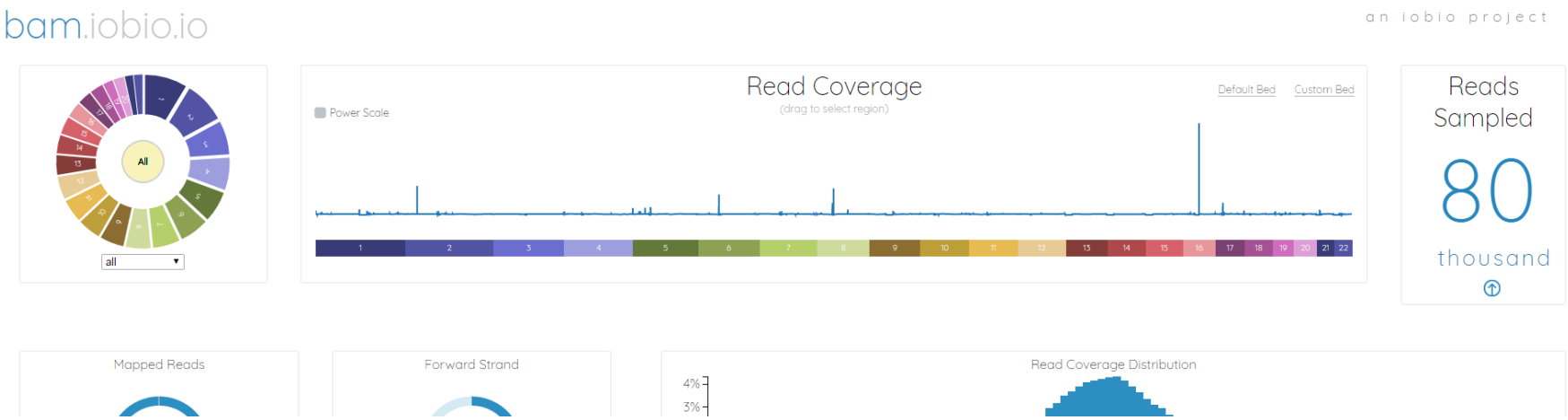
PICARD & iobio

PICARD

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

bam.iobio.io

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



Genetic variations

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

INDEL

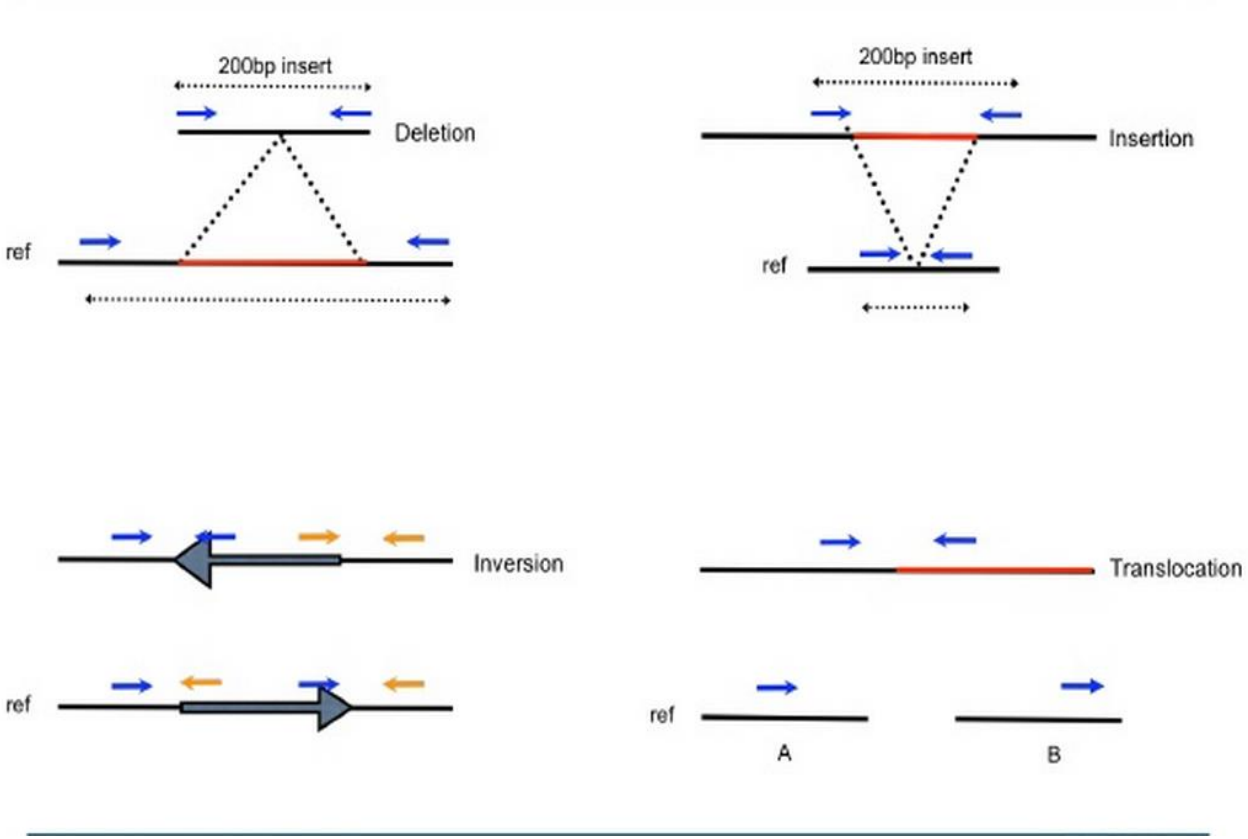
- Insertion / deletion of bases
- Coding regions of the genome - produce a frameshift mutation (unless multiple of 3)
- There are approximately 190-280 frameshifting INDELs in each person.
  A map of human genome variation from population-scale sequencing". Nature 467 (7319)

Types of genetic variations

Structural variations (SV)

- Variation in structure of an organism's chromosome

- Insertions
  Deletions
  CNV
  Inversions
  Translocations

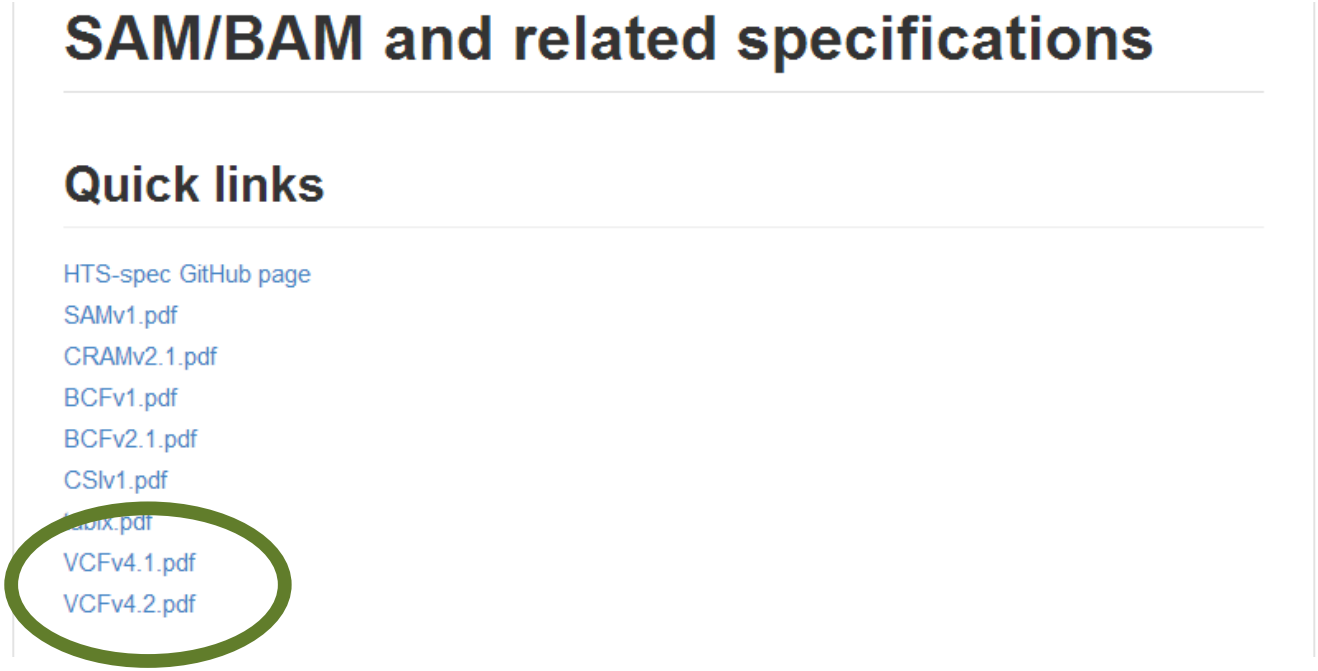


Variant Call Format VCF

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

(taken from Thomas Keane)

AUSTRIAN INSTITUTE OF TECHNOLOGY

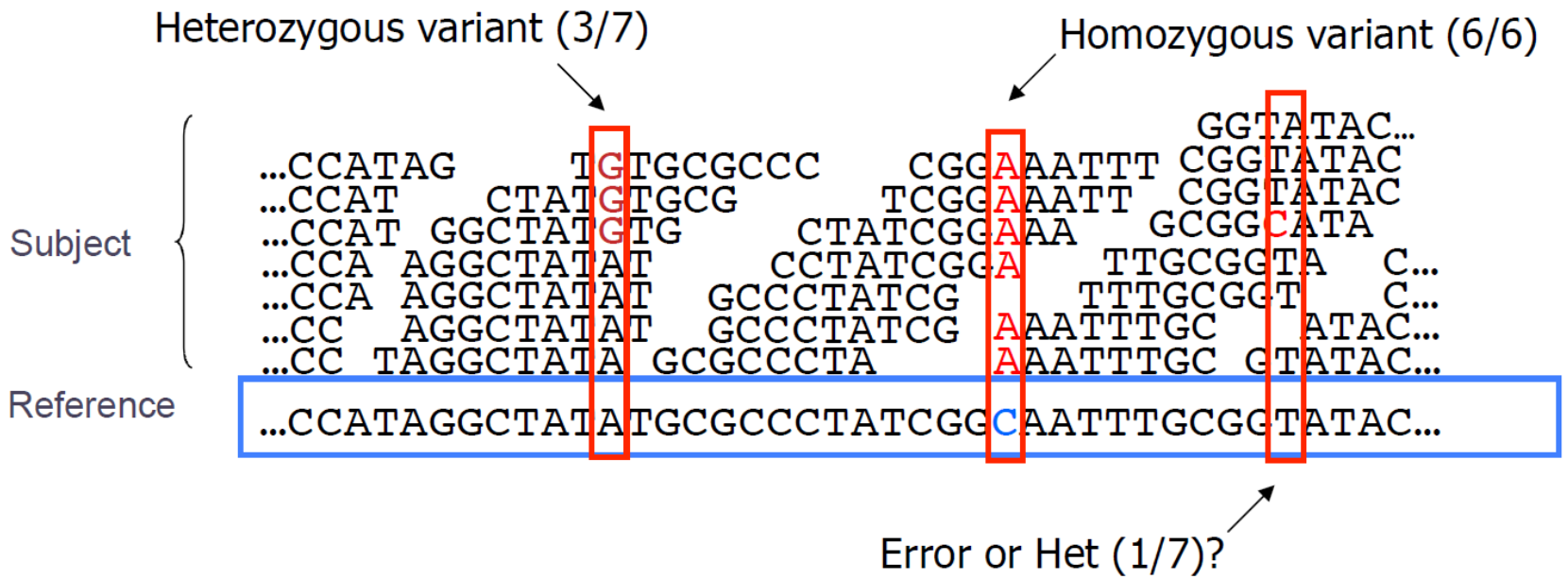
```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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?

AUSTRIAN INSTITUTE OF TECHNOLOGY

Variant Calling

Genotyping theory



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

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

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)

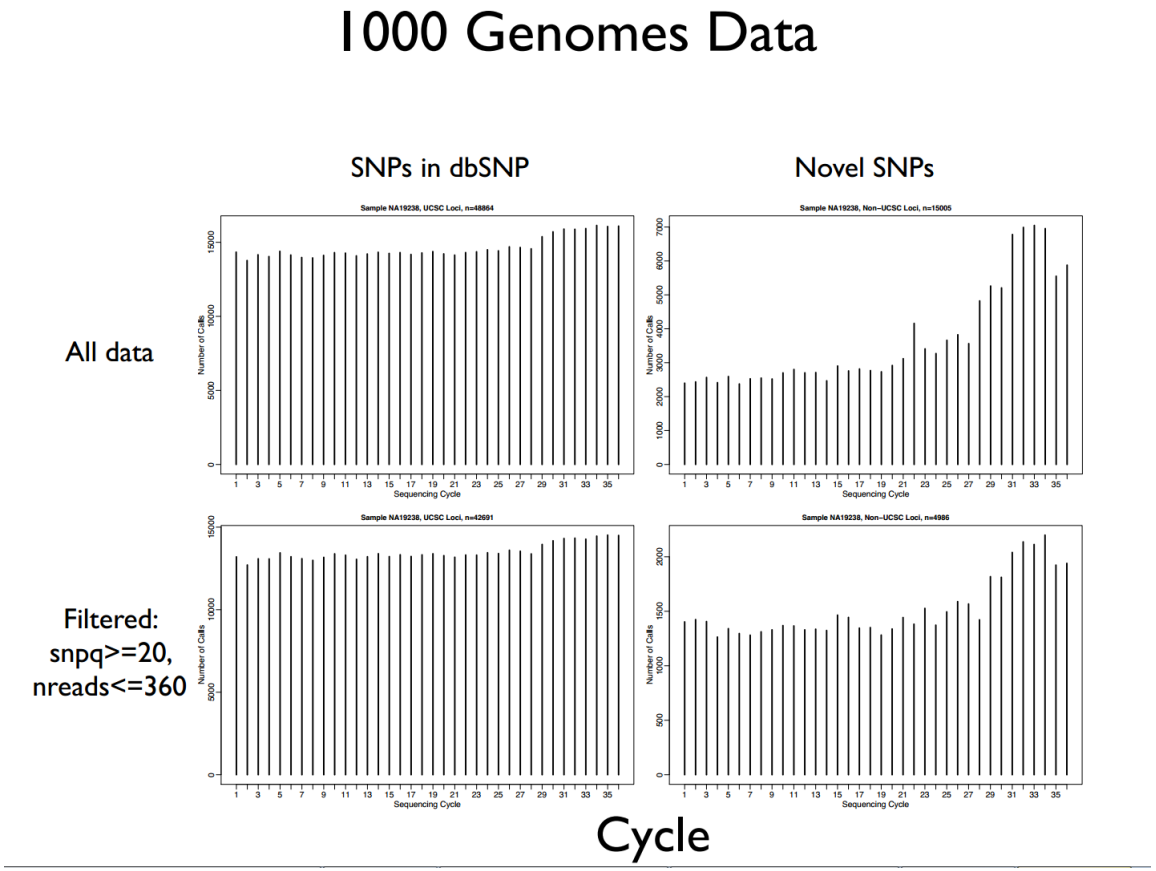
P(data|genotype): likelihood for observed (called) allele type

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

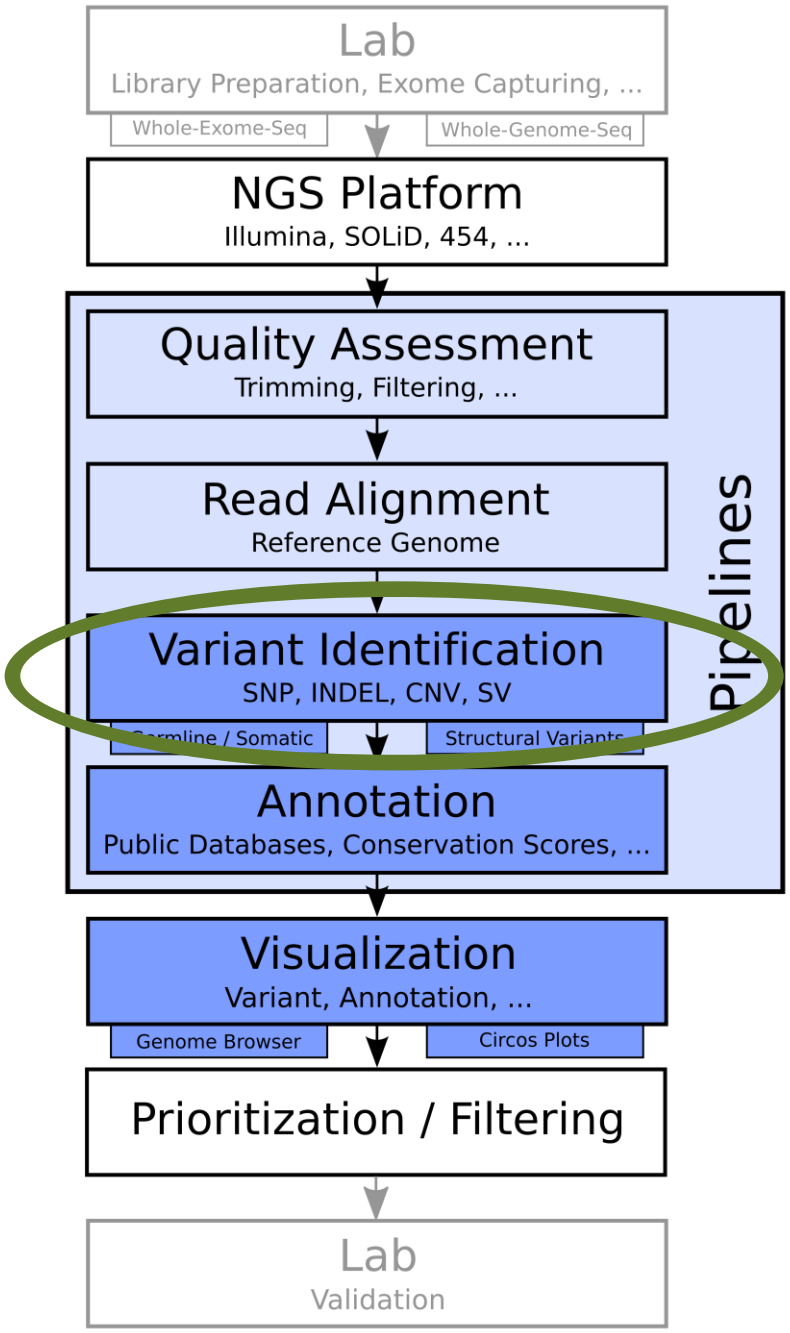
- Error rate increases as cycle numbers increase
- Error rate depends on substitution type (Ti/Tv)
- Error rate depends on local sequence environment
- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Proximity to INDEL

Difficulties

- Base calling gets more difficult the longer the read gets



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



Variant callers

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

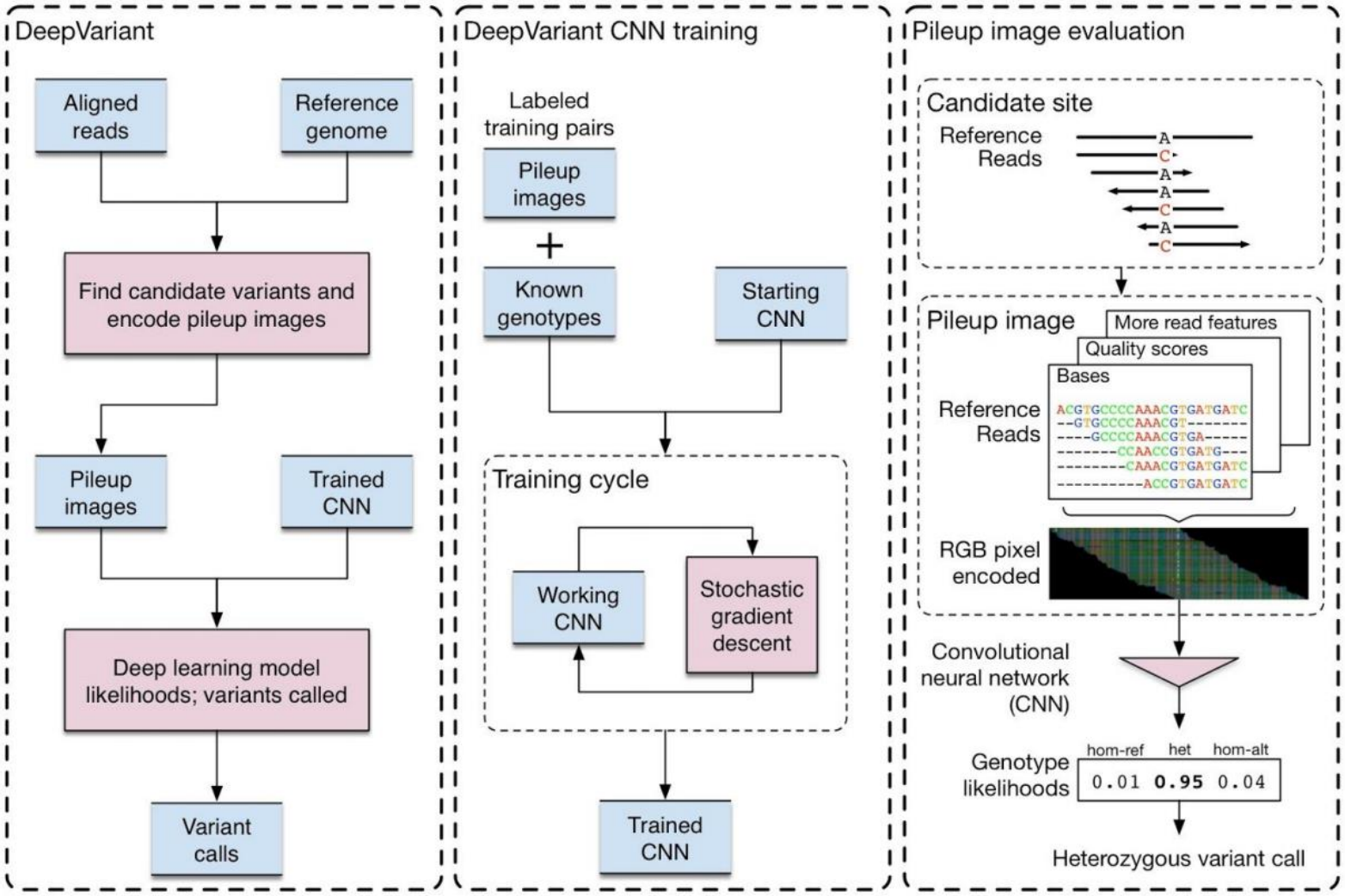
http://ged.msu.edu/angus/tutorials-2013/snp_tutorial.html

GATK - Genome Analysis Toolkit

Variant calling pipeline

Will be used in the practical session

Deep Variant



Creating a universal SNP and small indel variant caller with deep neural networks
Poplin et al. (2016) bioRxiv. doi: https://doi.org/10.1101/092890

Ensemble variant calling

Method

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

Validation

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

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

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%

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.

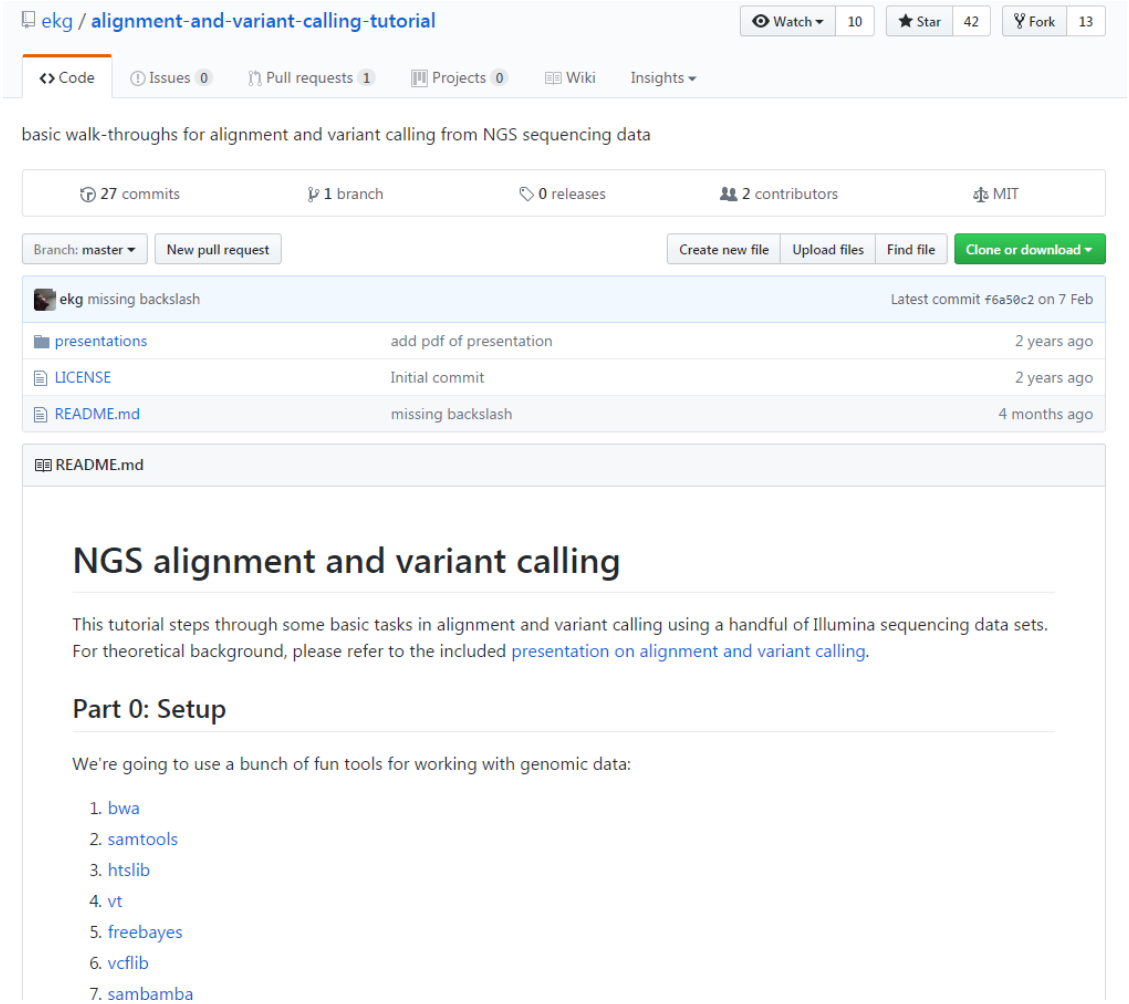
Features

- 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

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

Structural variation calling

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

Why is structural variation relevant / important?

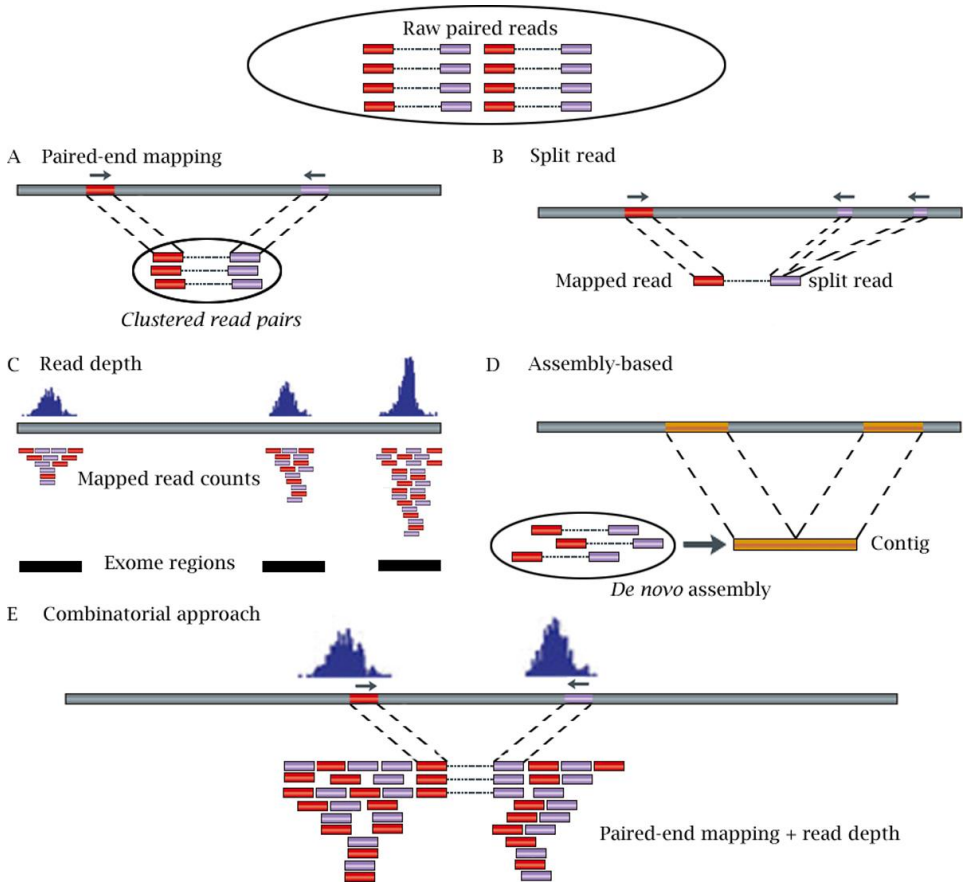
They are common and **affect a large fraction of the genome**

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

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

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

SV/CNV detection



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

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

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

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

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

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)
- Difficult to install

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.

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
- 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

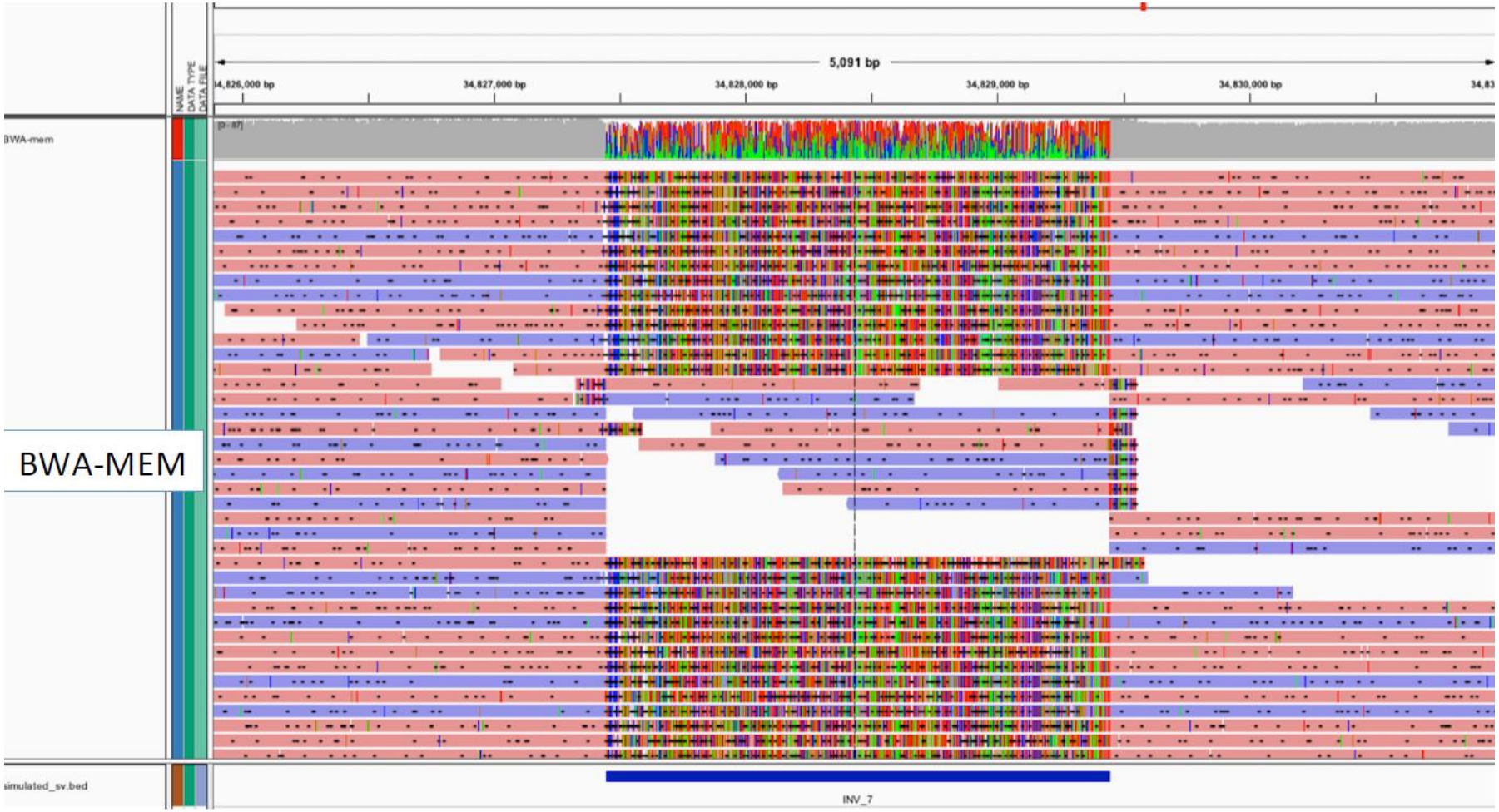
Long Read Technologies

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

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

PACBIO®

Hard to align



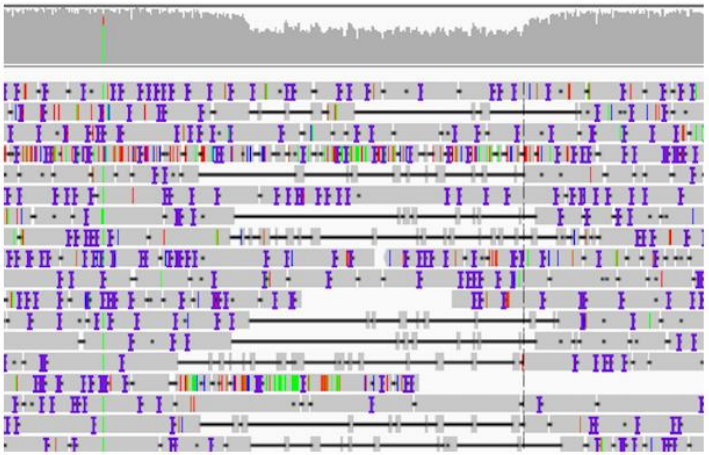
Human genome: 1kb Inversion

Improving long read alignment

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

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

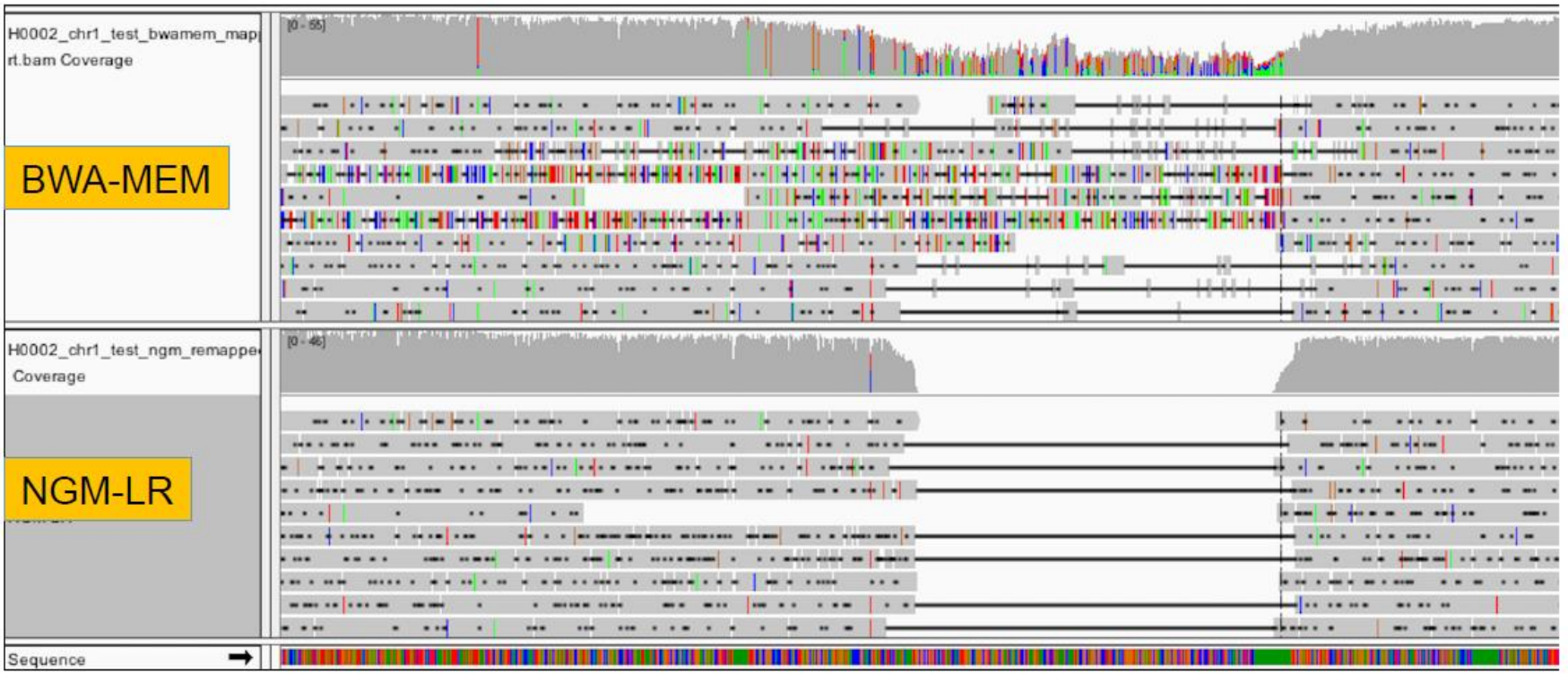
2. Improve alignment:
   - Insertions
   - Deletions



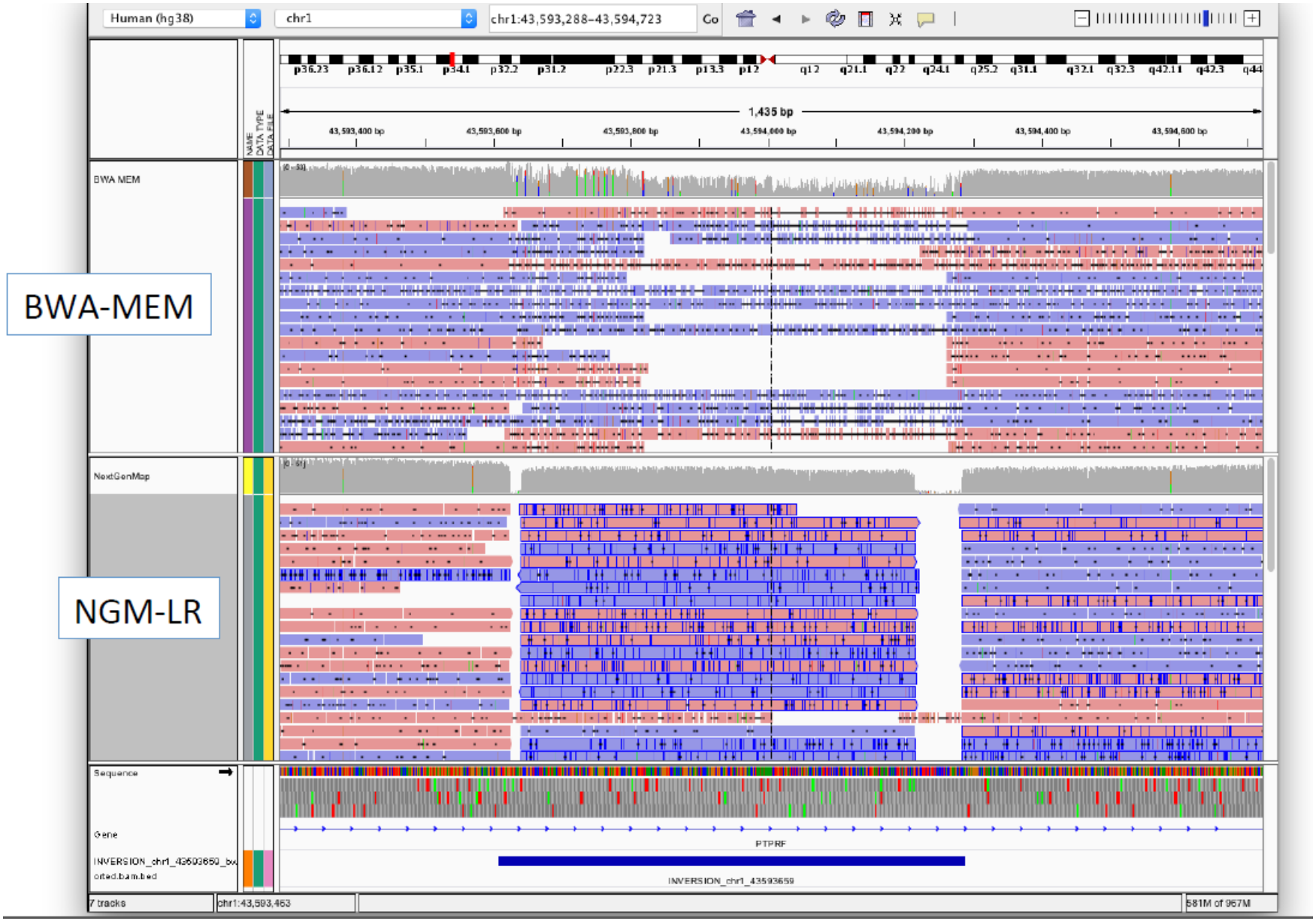
Structural variations with 3rd gen sequencing

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



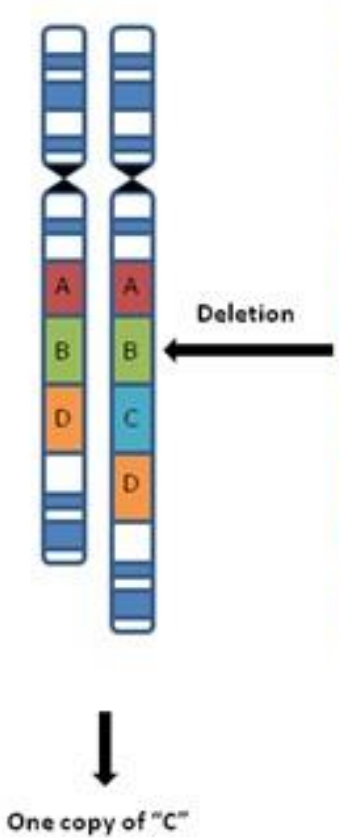
NGM-LR complex SV



CNV detection

Copy Number Variation - CNV

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

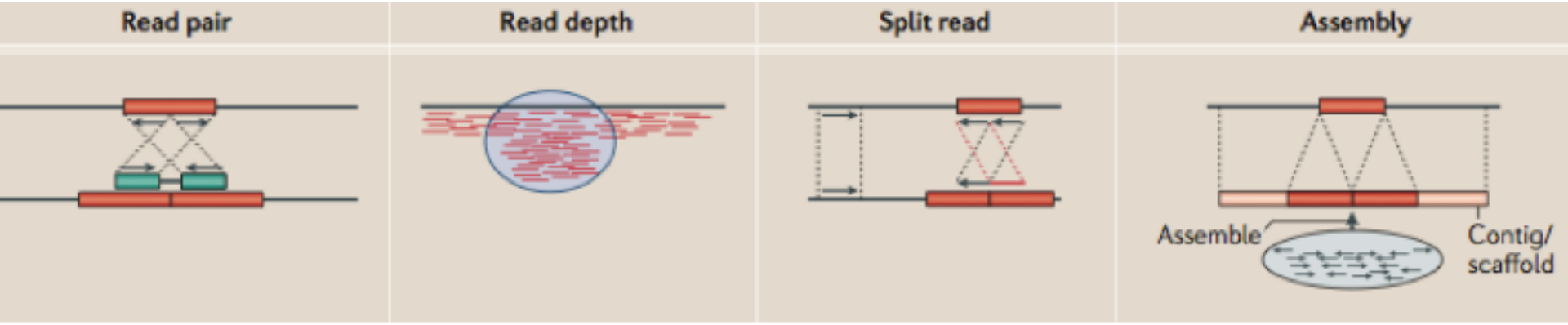


CNVs associated with complex diseases

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

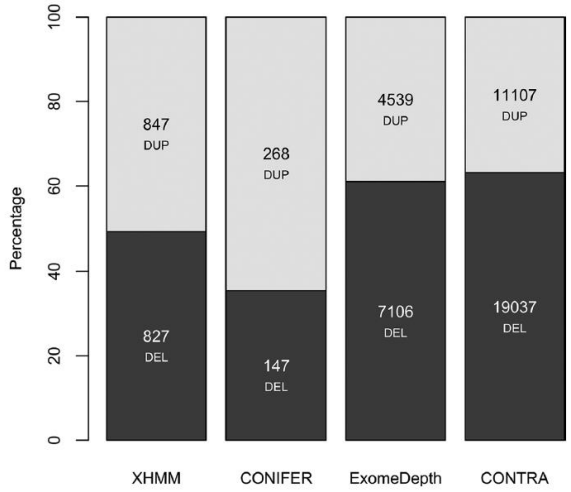
http://readingroom.mindspec.org/wp-content/genetics_CNV.jpg
Human Copy Number Variation and Complex Genetic Disease. Annual Review of Genetics. 2011

CNVs – how can we detect them



CNV detection

- XHMM
- CoNIFER
- ExomeDepth
- CONTRA



33 individual WES data (cumulatively)
Deletion and duplication CNVs

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

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/

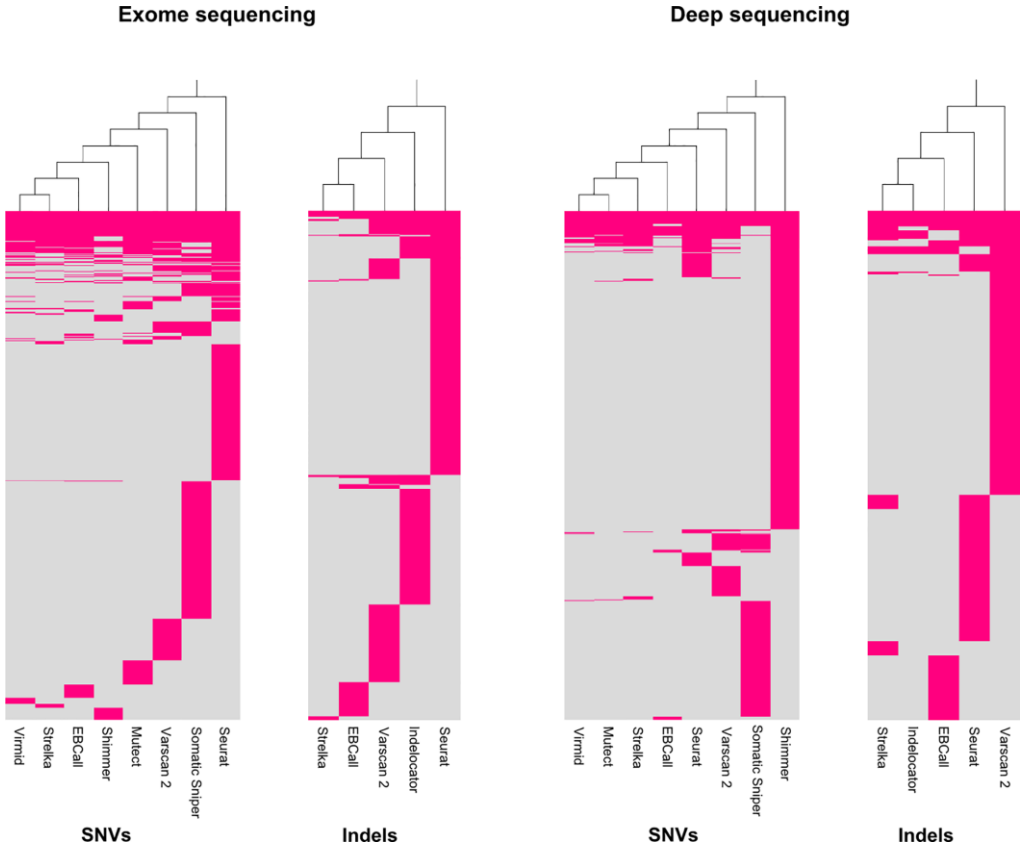
Here are a few more, a summary of the other answers, and updated link

- 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)

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



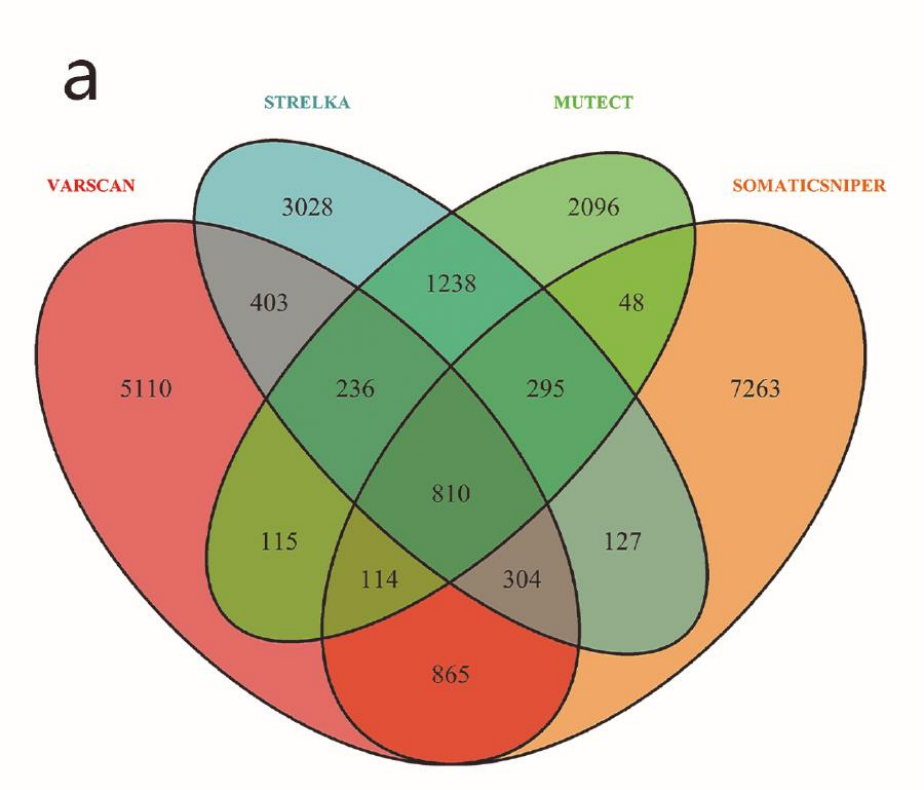
Each red line represents a called somatic mutation

Hierarchical cluster analysis of mutations called by the somatic variant callers in exome and deep sequencing data in left and right panel, respectively.

Cai et al, Sci Rep. 2016

- 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

Variant filtering

Variant filtering

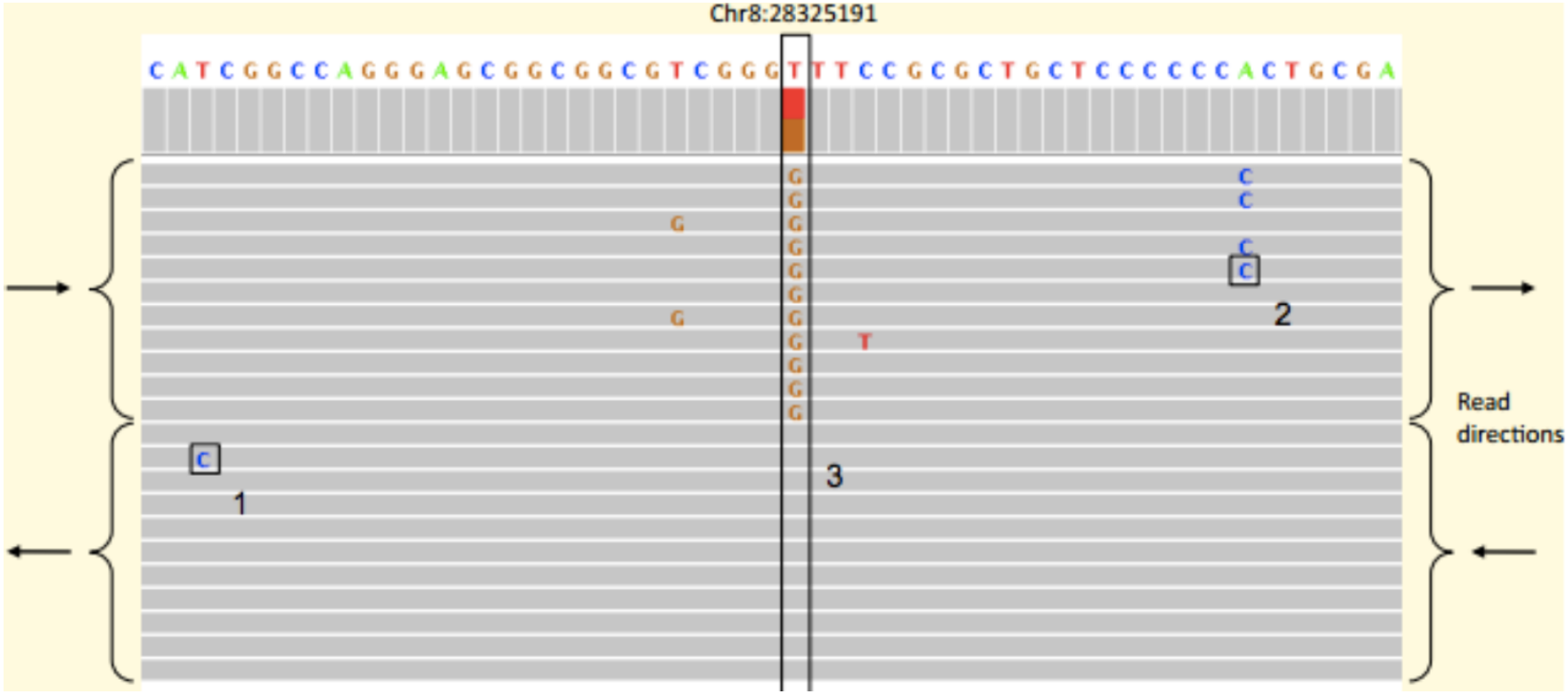
- QUAL (depends on MQ of reads and base qualities) is a useful measure
  - there will also be FP with high QUAL

Signs of suspicious variants

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

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

Beware of Systematic Errors



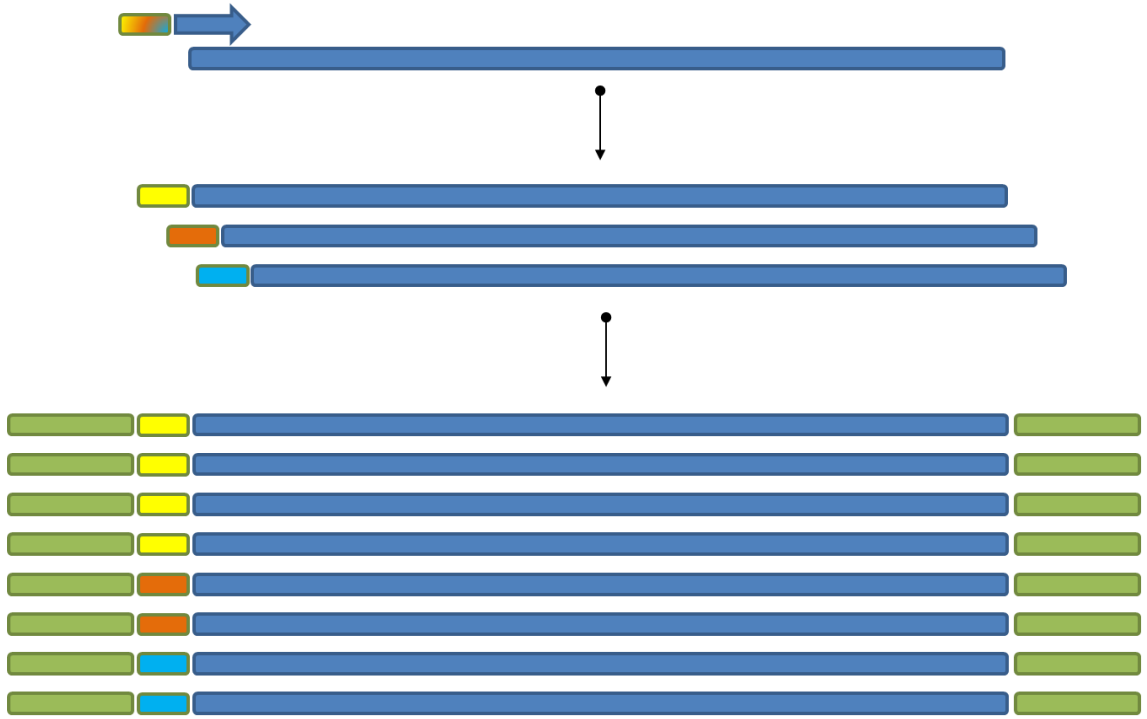
- **Identification and correction of systematic error in high-throughput sequence data** Meacham et al. (2011) *BMC Bioinformatics.* 12:451
- **A closer look at RNA editing.** Lior Pachter (2012) *Nature Biotechnology.* 30:246-247

Molecular barcoding

Library generation

STEP 1 - Barcoding

STEP 2 - Amplification



Consensus sequence generation

random barcode mix unique barcodes sequencing adaptors

Adapted from http://www.imgm.com

AUSTRIAN INSTITUTE OF TECHNOLOGY

Variant annotation

Variant annotation

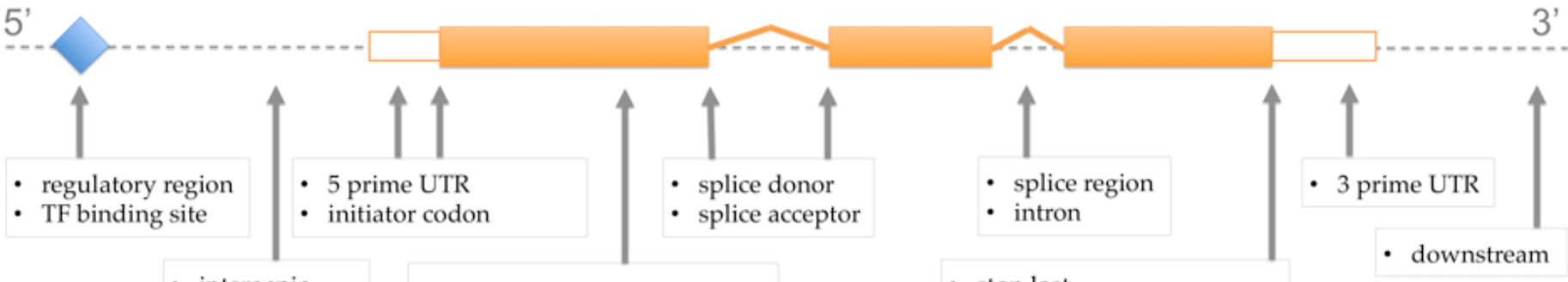
Annotation

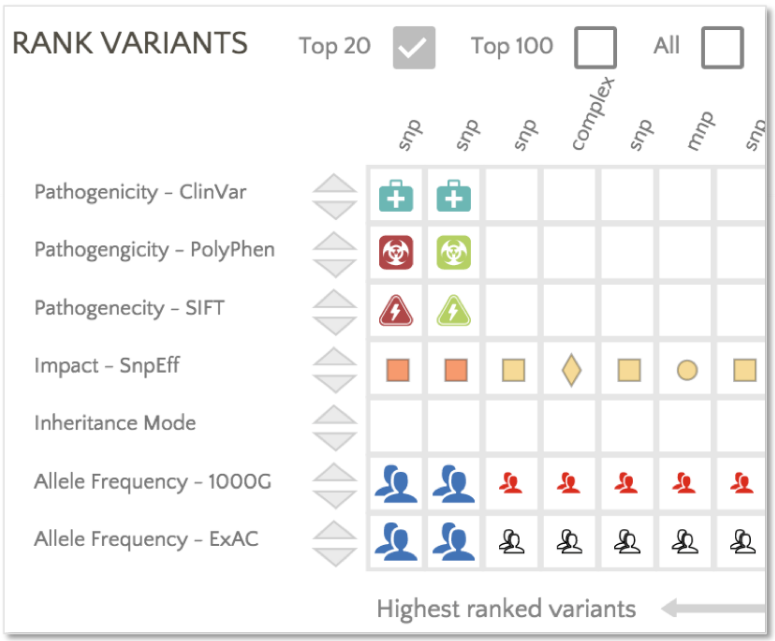
- Basis for filtering and prioritizing potential disease-causing mutations
- Most tools focus on the annotation of SNPs
- Many provide database links to various public variant databases (dbSNP...)
- Functional prediction of the variants
  - sequence-based analysis
  - region-based analysis
  - structural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

Interpretation of Variants

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

Solution: effect prediction





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

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

Make use of phylogenetic conservation

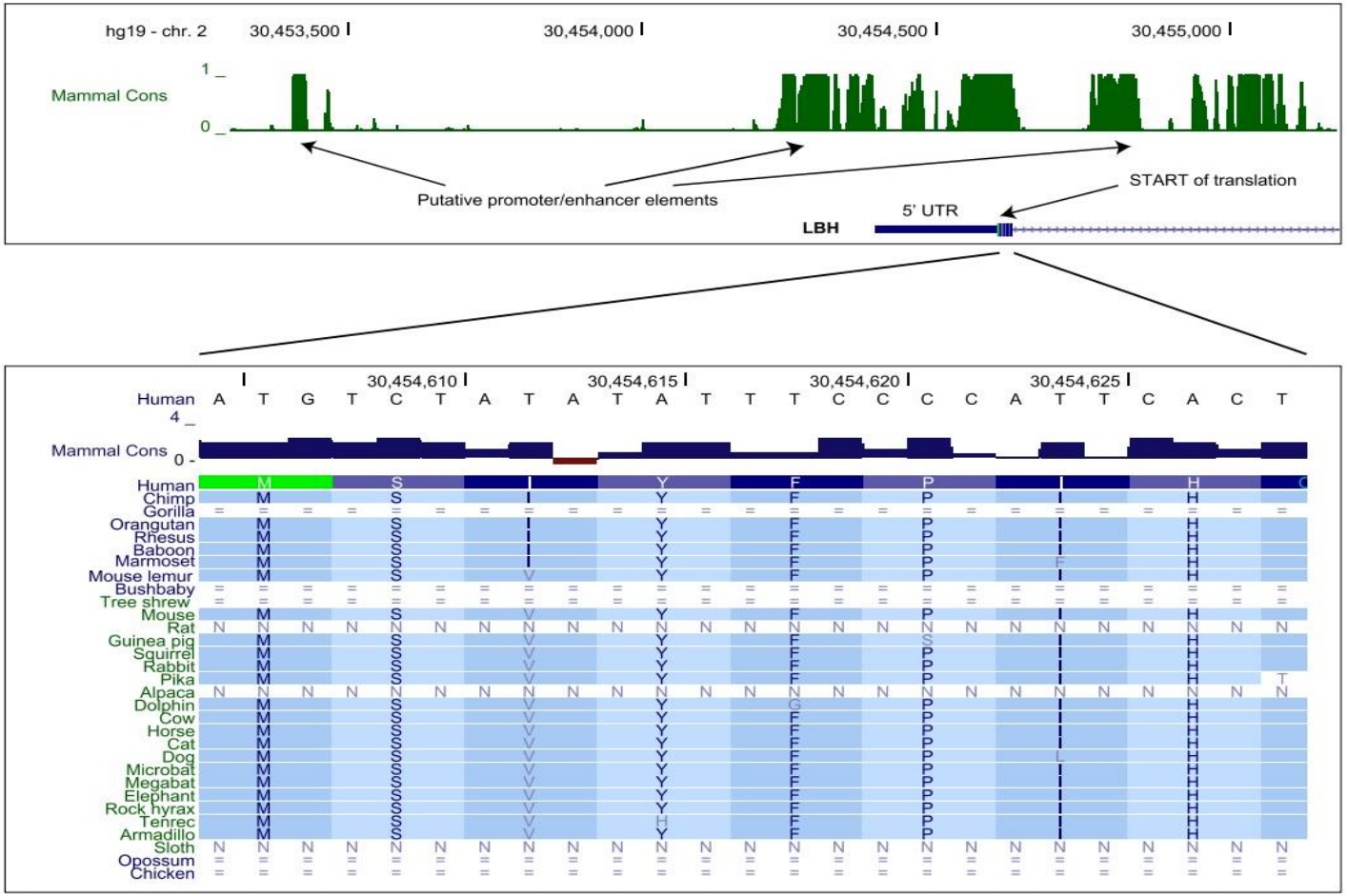
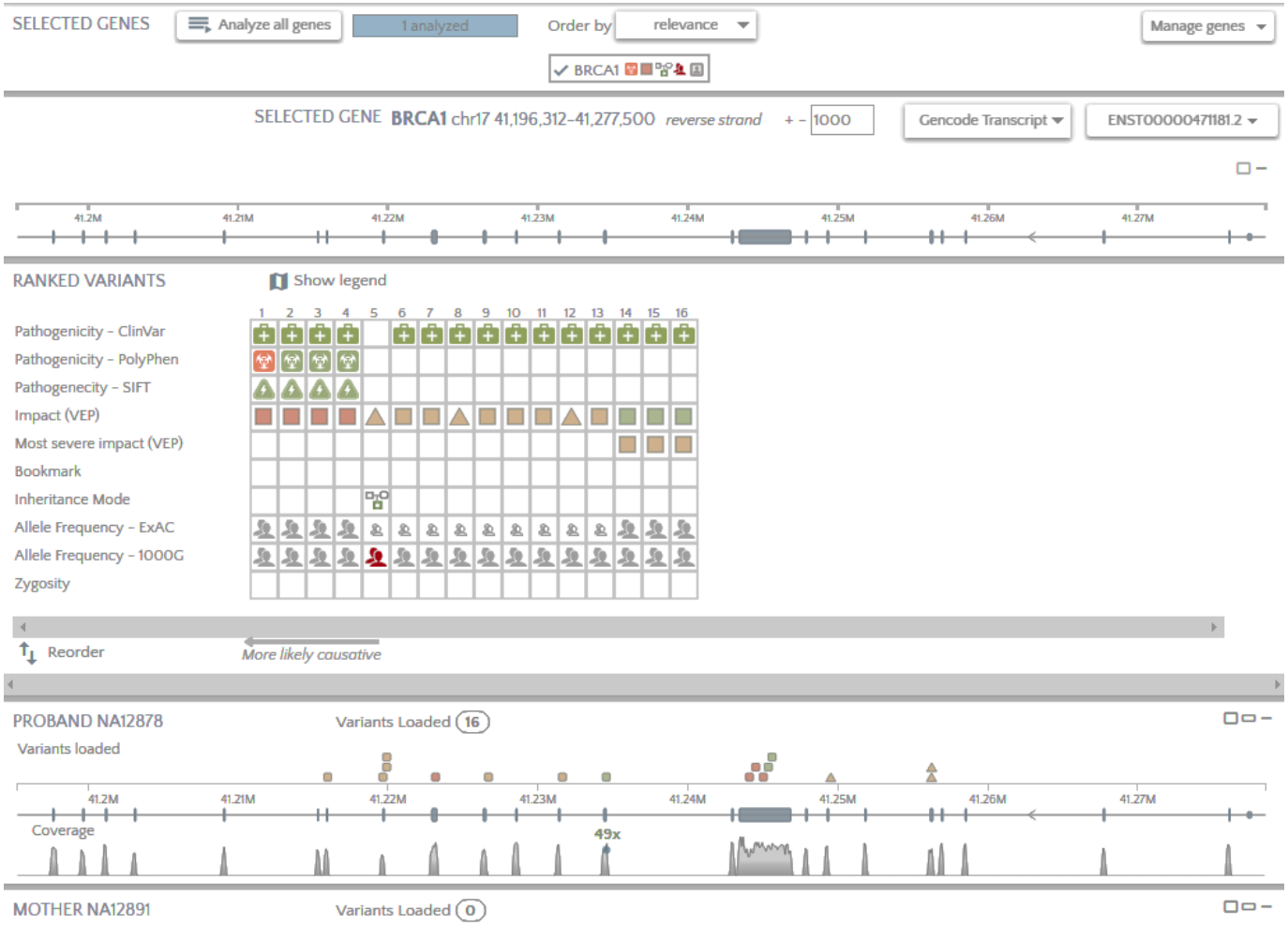


Figure 1. A comparative genomics display derived from the UCSC Genome Browser (Meyer et al. 2013). The *top* panel depicts the genomic region surrounding the 5′ end of the gene *LBH* (limb bud and heart development homolog) in the human genome. The *top* track indicates mammalian conservation as determined by phastCons (Pollard et al. 2010). Putative promoter and enhancer elements are indicated. The second track shows the intron/exon structure of the 5′ end of *LBH*. The 5′ untranslated region (UTR) and start site are indicated. The *bottom* panel shows a close up on the protein-coding portion of the first exon of *LBH*. Here, the *top* track shows the human DNA sequence, and the second track shows the degree of mammalian conservation as determined by PhyloP (Pollard et al. 2010). The *bottom* series of tracks shows the homologous protein sequence in selected vertebrate genomes. (N) Gaps in sequence; (=) unalignable sequence.

Gene annotation

Gene.iobio.io

- Interactive
- Load custom data



dbSNP

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

- Single Nucleotide Polymorphism Database
- Central repository for SNPs and INDELs
- Information for variants: Population, Sample Size, allele frequency, genotype frequency, heterozygosity, …
- ~907m submissions, ~320m variants (stats only for human) [v150]

Problems

- high FP rate
- not many validated SNPs (~30%)

→ careful when filtering out variants based on dbSNP information

In order for a RefSNP(rs) to be validated, at least one of its clustered submitted SNPs (ss) must either have been ascertained using a non-computational method or have frequency information associated with it.

dbSNP

AUSTRIAN INSTITUTE OF TECHNOLOGY

Minor Allele Frequency (MAF)

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

Consider a variation with the following alleles and allele frequencies:

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

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

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

Variant annotation - tools

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

Variant annotation - tools

ANNOVAR

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

SeattleSeq Annotation server

- Online tool
- Human SNPs and INDELs

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

ExAC

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

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

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

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

CADD

Combined Annotation Dependent Depletion

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

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

Variant annotation – what to consider

- 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
- List of tools
  https://docs.google.com/spreadsheets/d/1JRVrNniAoiraR8Jv22ZmqIWSztUomwcHlHsaHOhoJBU/edit#gid=0

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

Visualization

Visualization

Genome browsers - most widely-used tools

Read

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

Able to

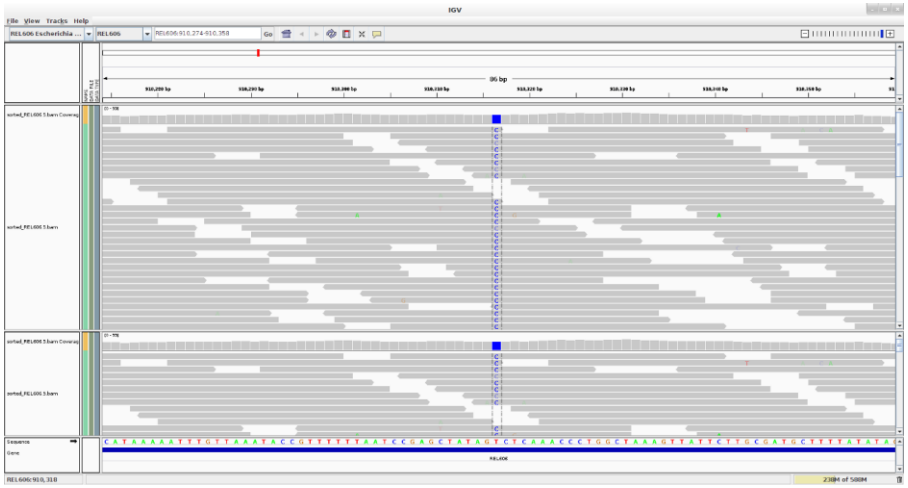
- Browse/zoom genome
- Display multiple samples / multiple tracks
- Colorize/mark features of your data (paired reads, SNPs, ...)

Genome Browsers

IGV (Integrative Genomics Viewer)

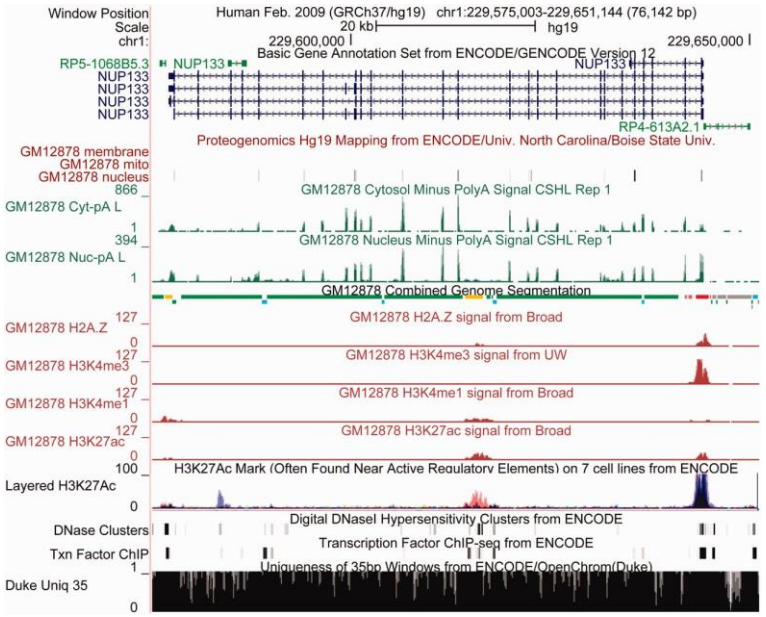
- Widely used viewer
- Java based – standalone tool
- Easy and fast to view own data
- IGV3 **supports long-reads**

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



UCSC

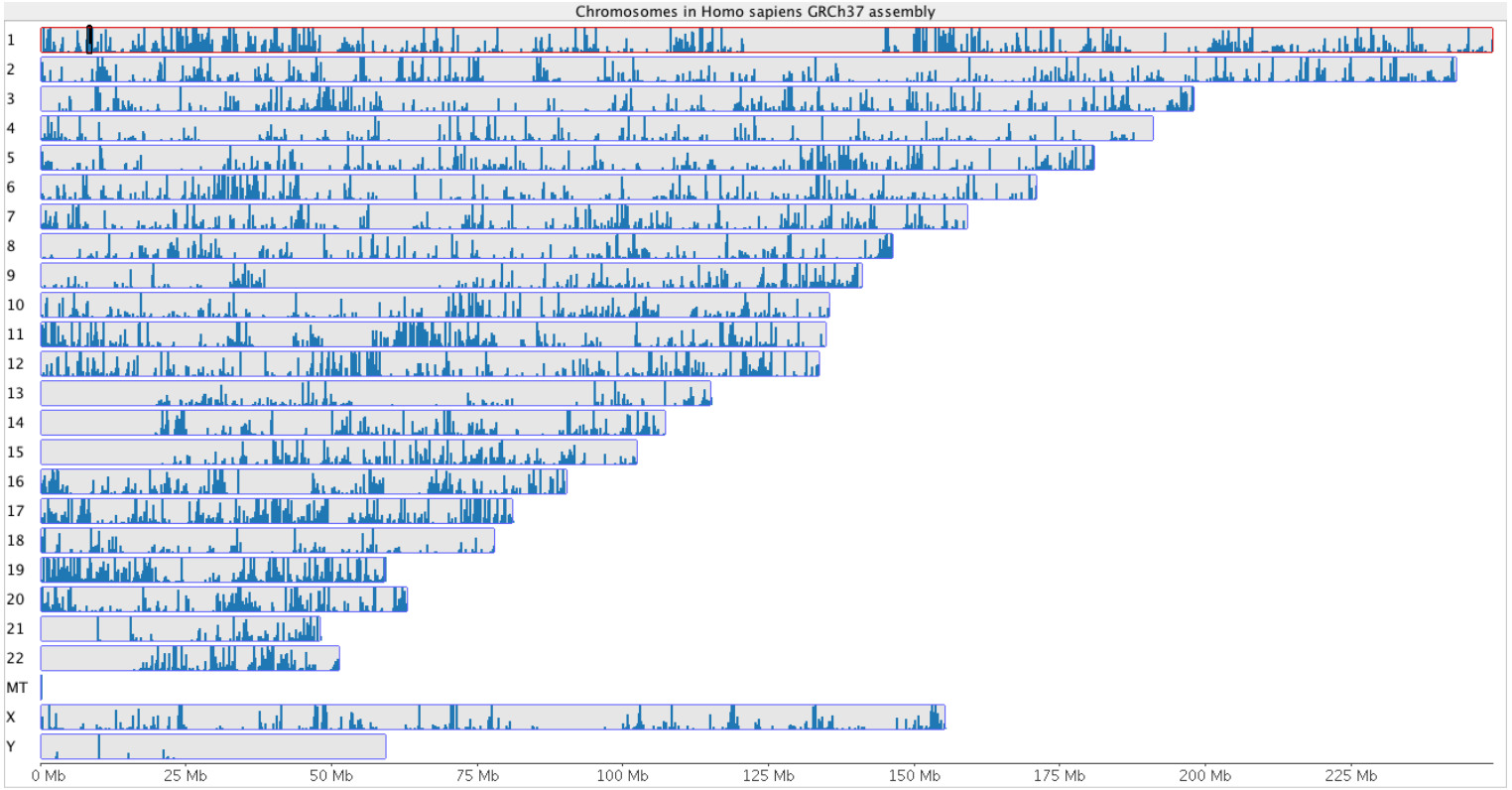
- Web based tool
- Offers many different annotation tracks
- Needs some configuration to display own data



Coverage visualization

Coverage histogram for chromosomes

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

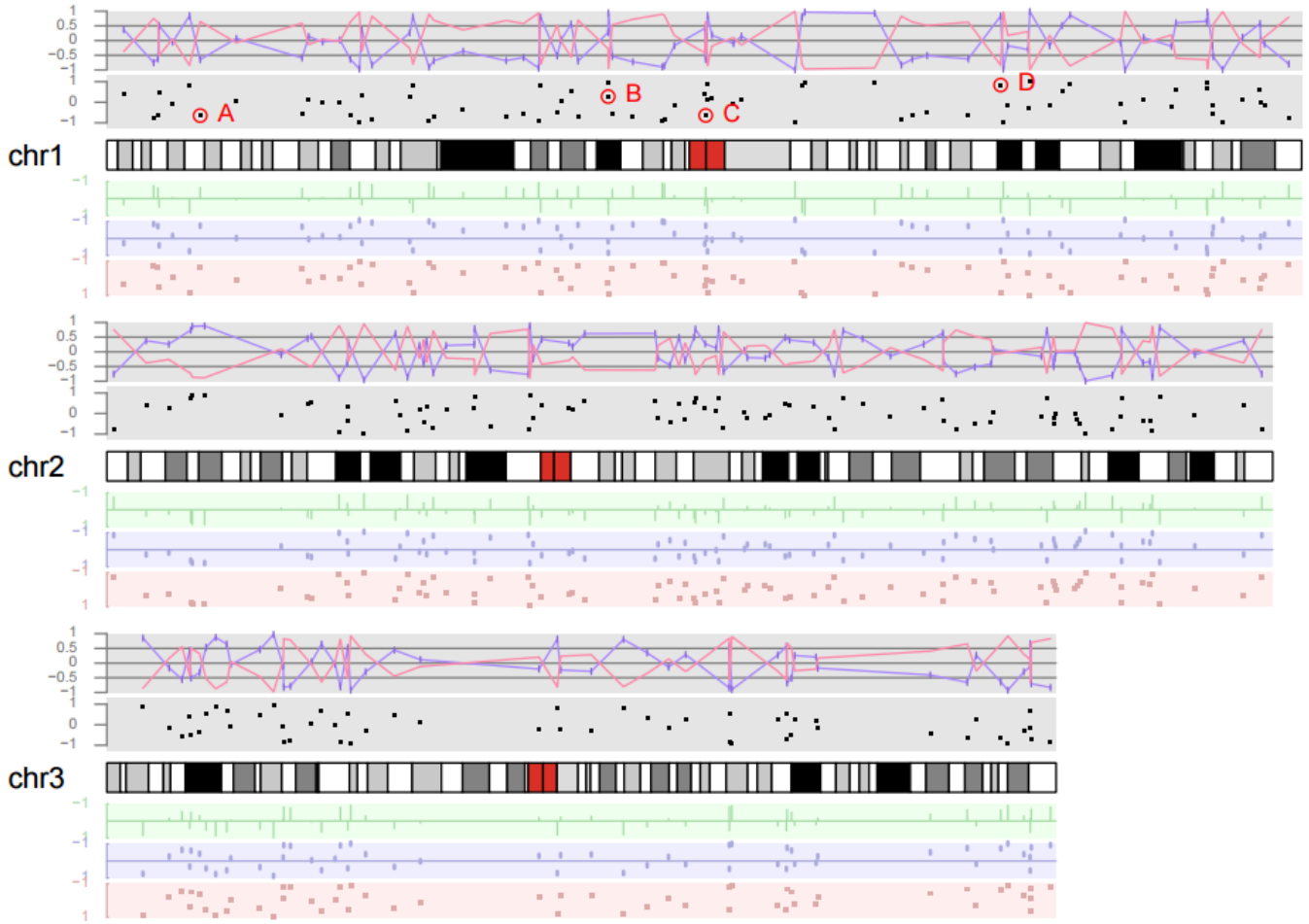


http://seqanswers.com/forums/attachment.php?attachmentid=2118&d=1364889859

Genome-wide visualization

karyoploteR

- Customizable karyotypes with arbitrary data

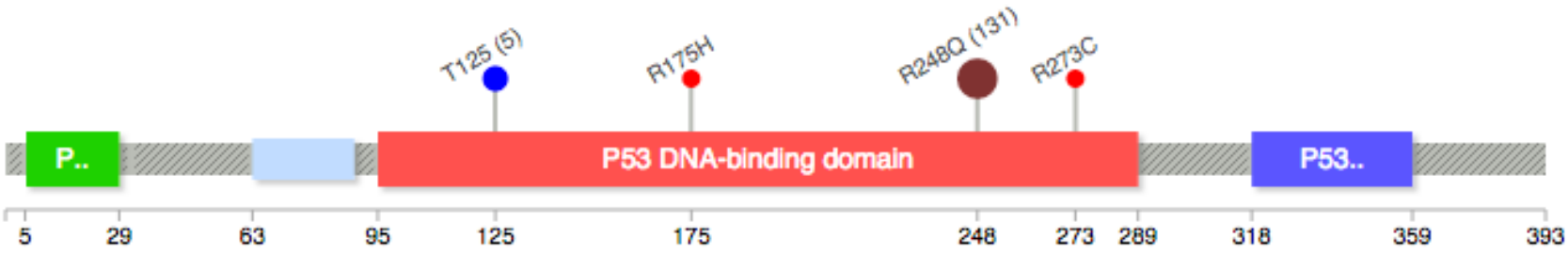


Lollipop

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
```



Other useful information

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

Genome in a Bottle (GIAB)

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

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

IGSR: International Genome Sample Resource

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

IGSR and the 1000 Genomes Project

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

Recommendations

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

Where can you get help and information?

Biostar

- A high quality question & answer Web site.

SEQanswers

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

http://omictools.com/

- An informative directory for multi-omic data analysis

Rosalind (http://rosalind.info/)

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

Collection of helps

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

Useful information

List of one liners

https://github.com/stephenturner/oneliners

Basic awk & sed

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

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

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

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

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

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

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

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

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

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

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

SAM and BAM filtering oneliners

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

bamfilter_oneliners.md Raw

SAM and BAM filtering one-liners

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

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

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

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

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

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

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

Count with flagstat for additional information:

```
samtools flagstat filename.bam
```

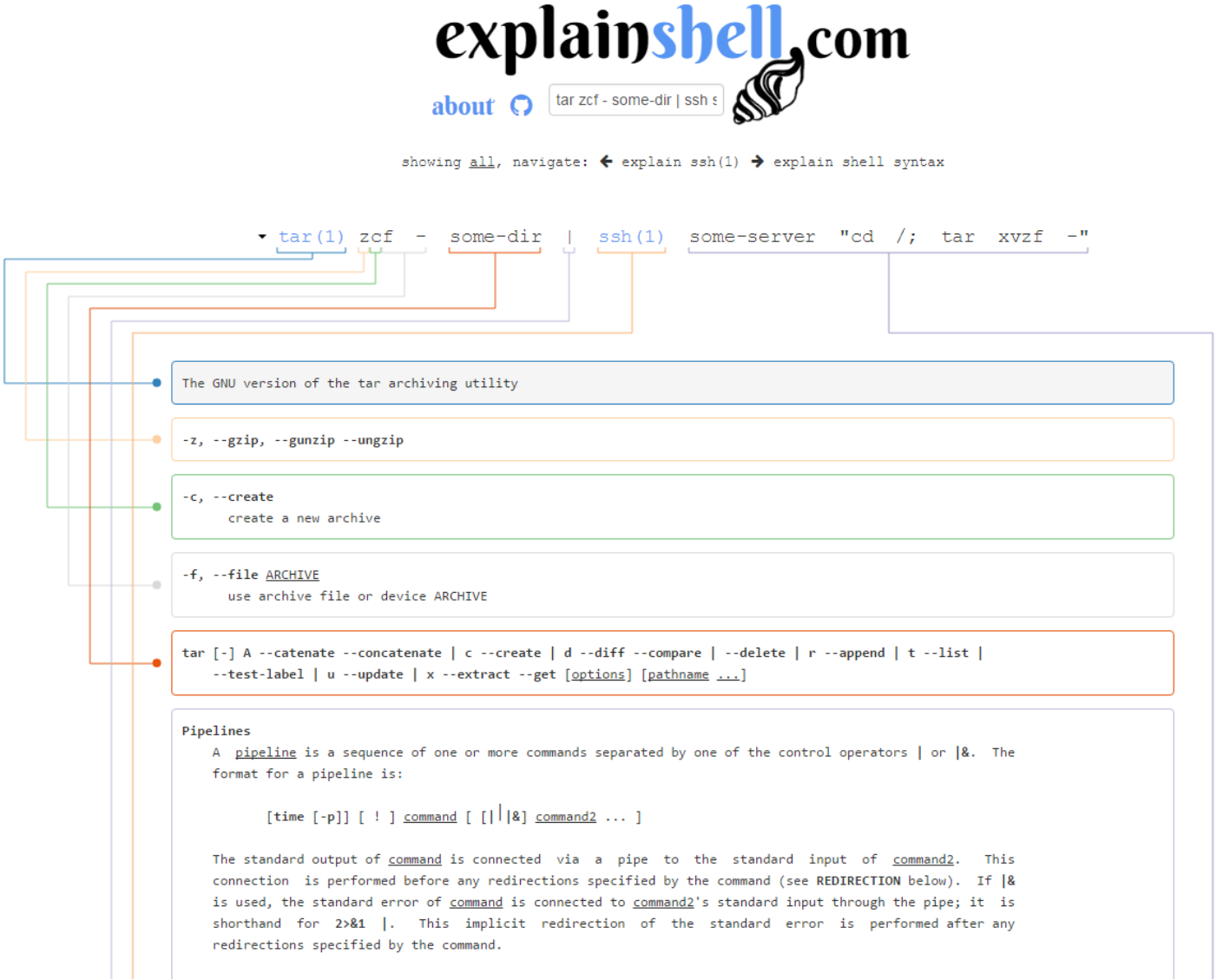
Count the number of alignments (reads mapping to multiple locations counted multiple times)

Collection of published “guides” for bioinformaticians

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

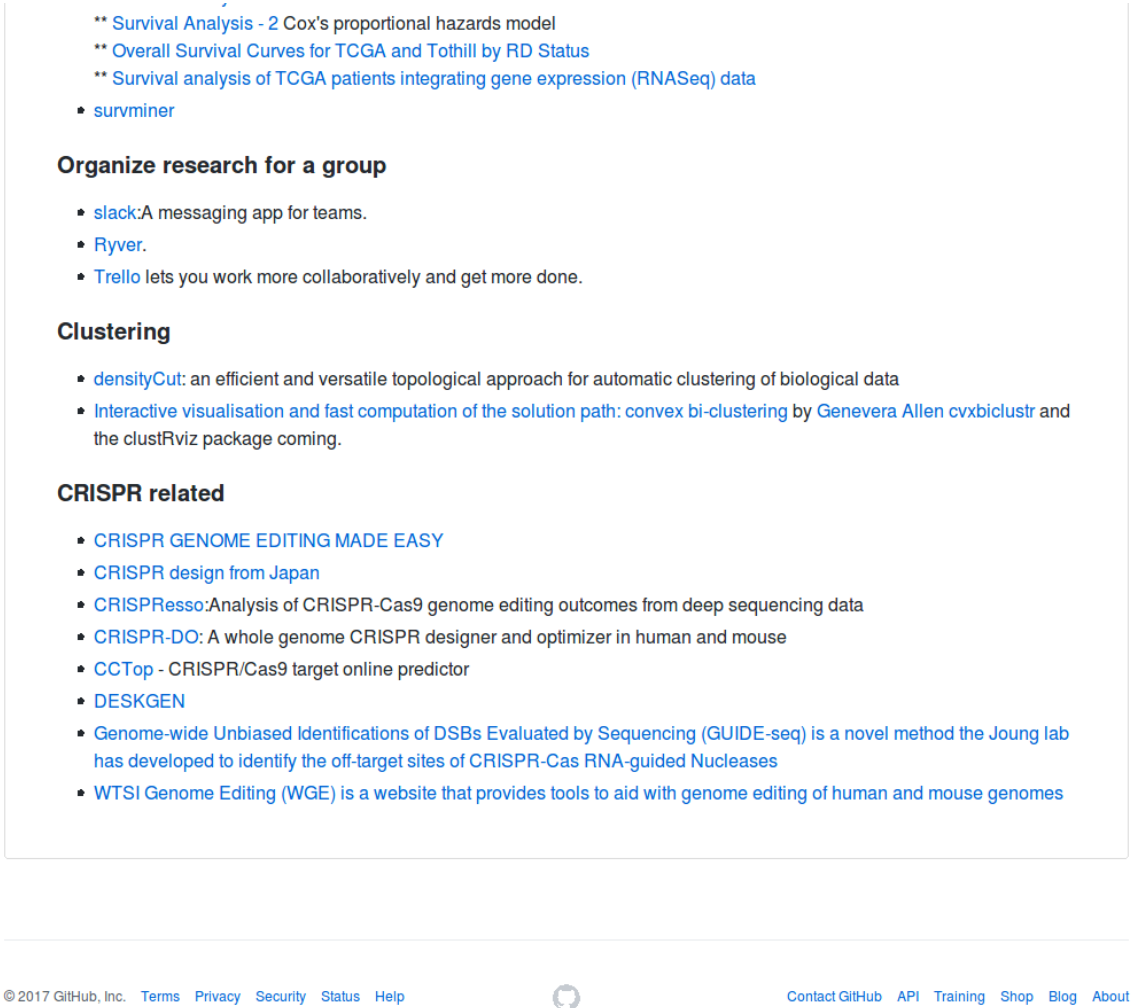
1. Loman N and Watson M (2013) So you want to be a computational biologist? *Nature Biotech* **31(11)**:996-998 [link]
2. Corpas M, Fatumo S, Schneider R. (2012) How not to be a bioinformatician. *Source Code Biol Med.* **7(1)**:3. [link]
3. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy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

Explain Shell - https://explainshell.com/



Huge resource

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



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

- survminer

Organize research for a group

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

Clustering

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

CRISPR related

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

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)

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Tools for variant analysis of Next Generation Genome Sequencing data

Lecture

Stephan Pabinger
stephan.pabinger@ait.ac.at

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Tools for variant analysis of Next Generation Genome Sequencing data

Practicals

Stephan Pabinger
stephan.pabinger@ait.ac.at

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

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)
- ...

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
- …

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

Realignment around INDELs

Why?

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

Realignment principles

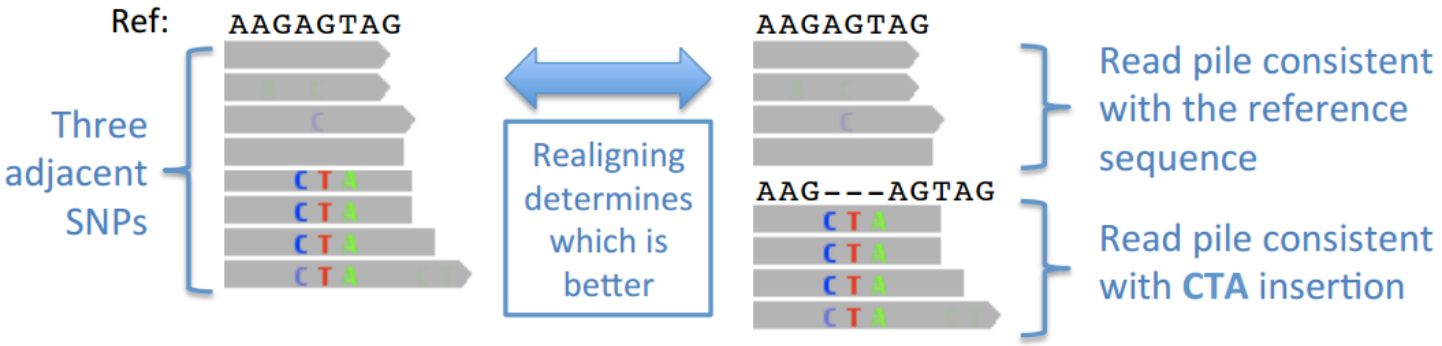
- Realign locally around INDELs→ GATK

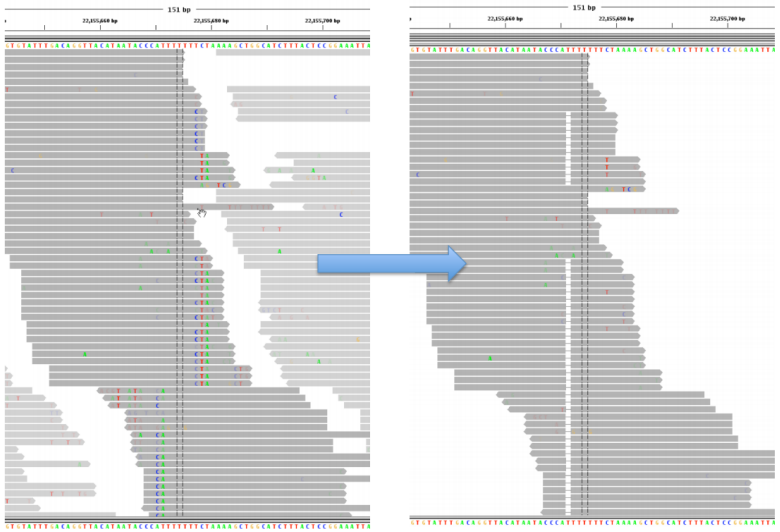
- Input set of known INDEL sites (dbSNP, 1000genomes)
- INDELs from alignments
- Presence of mismatches & softclips (BAM file)

Realignment around INDELs

Realignment

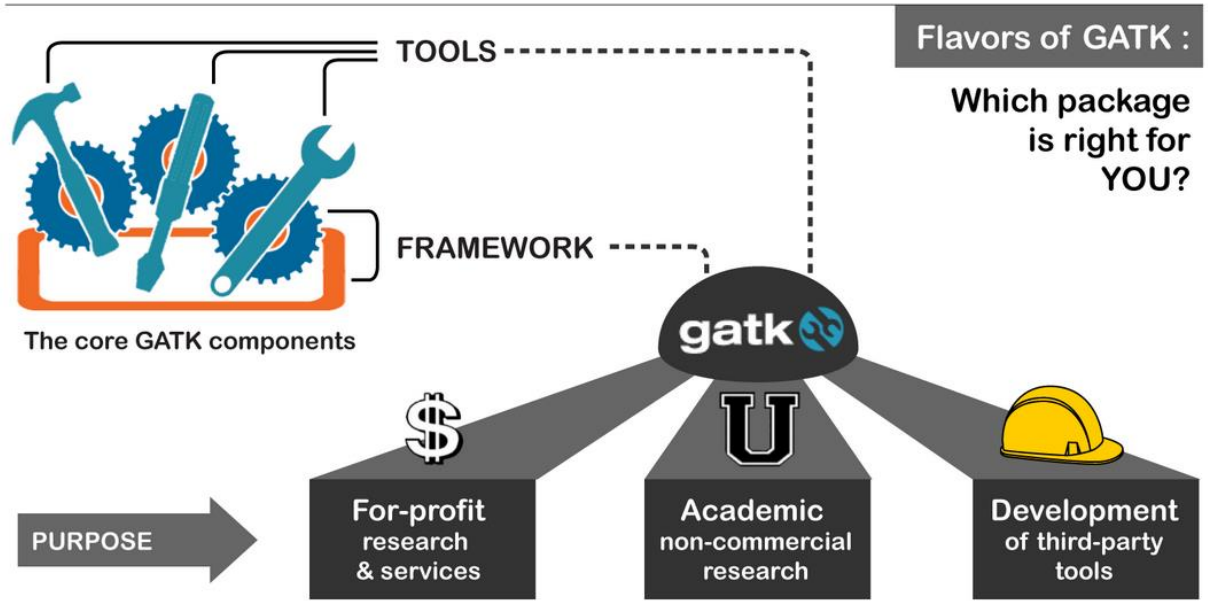
- Model the INDEL haplotype
- → if score of alternative consensus is better than original use realigned



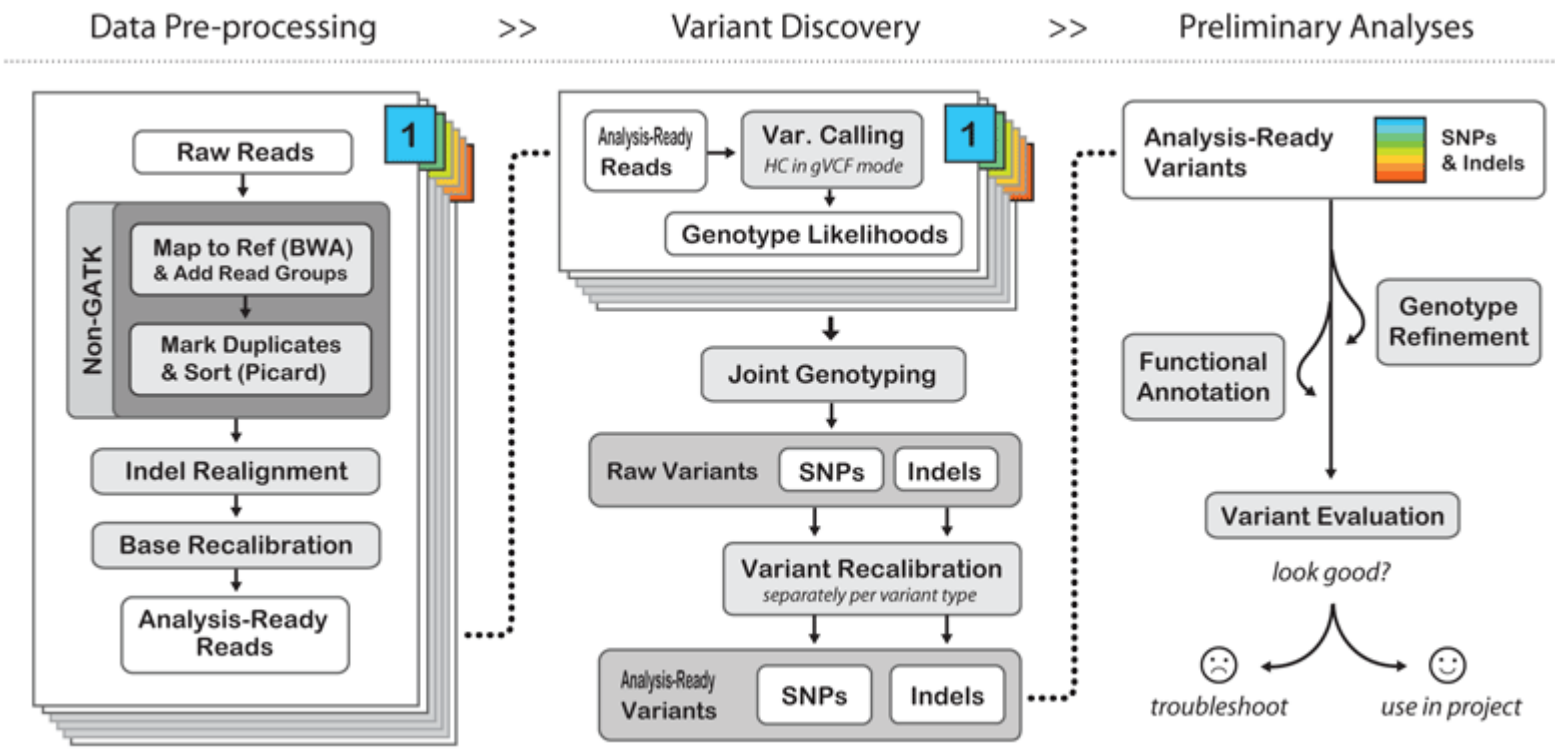


GATK

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



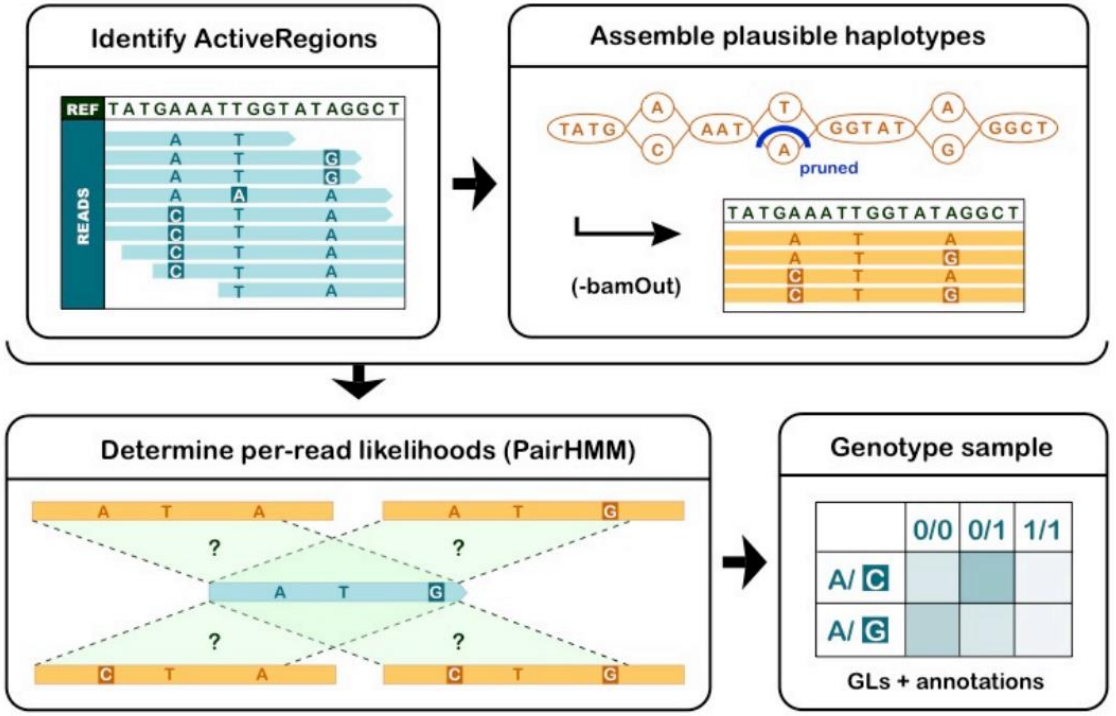
GATK


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.

Practicals

https://github.com/spabinger/BioinformaticsAndGenomeAnalyses2017