Medizinische Genomanalysen

LE 5 (20.04.2021 – 18:00)

Stephan Pabinger
stephan.pabinger@gmail.com

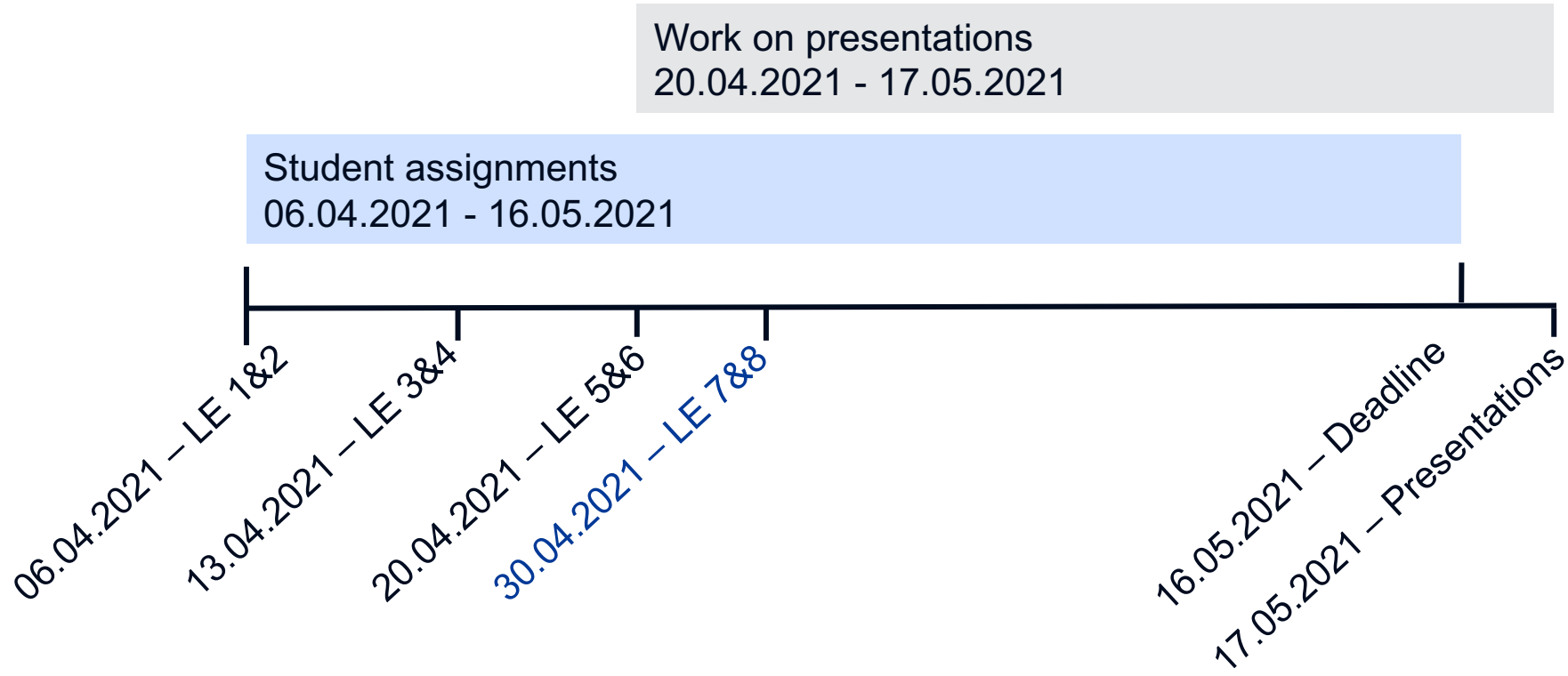
Recap

Assignment 1 & Assignment 2

Status update

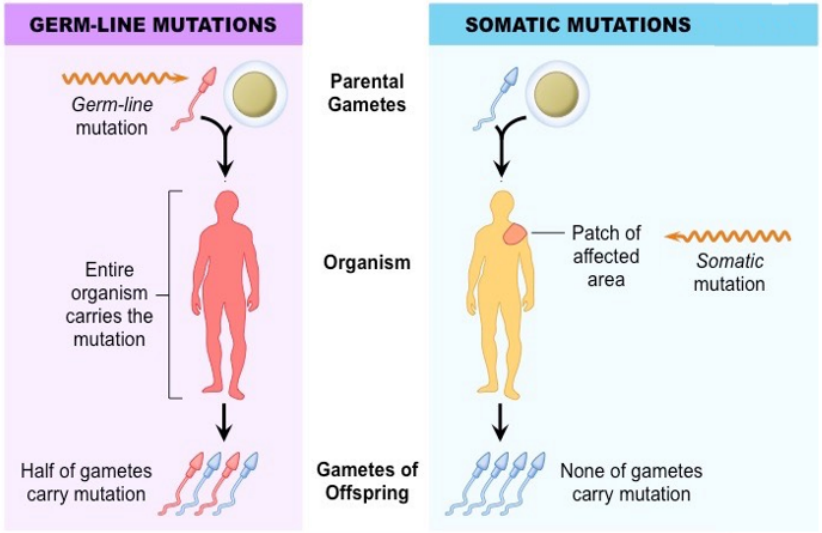
github.com/spabinger/medizinische_genomanalysen_2021

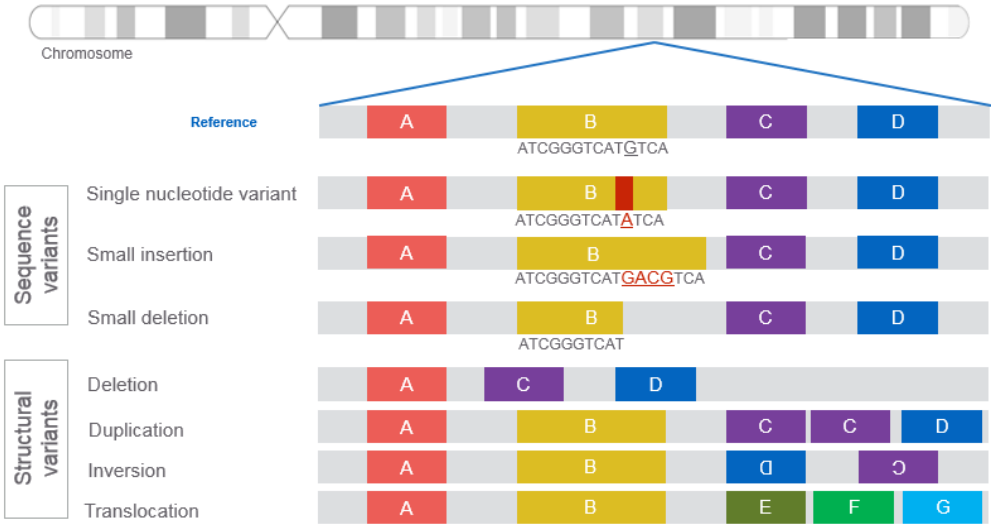
Timeline


Work on presentations
20.04.2021 - 17.05.2021
Student assignments
06.04.2021 - 16.05.2021
06.04.2021 – LE 1&2
13.04.2021 – LE 3&4
20.04.2021 – LE 5&6
30.04.2021 – LE 7&8
16.05.2021 – Deadline
17.05.2021 – Presentations

Recap

- Samtools
- Genetic variations
- Variant calling
- Structural variant calling





Questions

Annotation – graph genomes

- Seven Bridges -> https://www.sevenbridges.com/graph-genome-academic-release/
  - Fast and accurate genomic analyses using genome graphs, Nature Genetics 2019
  - “The read alignments against a path in the graph are projected to the standard reference genome and output to a standard BAM file, with the alignment path along the graph reported using custom annotation tags.“
- Vg
  - Uses a reference and a vcf file to build the graph; Variants are then reported based on the reference

Reverse strand

- SAM flag 16

Sequencing DNA input quantities

- https://dnatech.genomecenter.ucdavis.edu/sample-requirements/
- https://genomics.ed.ac.uk/resources/sample-requirements
- Lijuan Zhang

Structure

LE 1

- History, terms, sequencing, SAM/BAM

LE 2

- Reference genome, FASTQ, cleaning

LE 3

- SAMtools, genetic variation, variant calling

LE 4

- Variant callers, structural variations & callers

LE 5

- CNVs, somatic mutations, filtering, annotation

LE 6

- Visualization, pipelines

LE 7 & 8

- Albert Kriegner

LE 9 & 10

- Presentations

Structural variations – combination/evaluation

SURVIVOR

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

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

- Filter and combine the calls from VCF files

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

Meta SV-caller

Parliament2

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

https://github.com/dnanexus/parliament2

Structural variations - challenges

Often many false positives

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

The false negative rate is also typically high

- Most current datasets have low to moderate ***physical*** coverage (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- Too stringent filtering removes TP

Evaluation

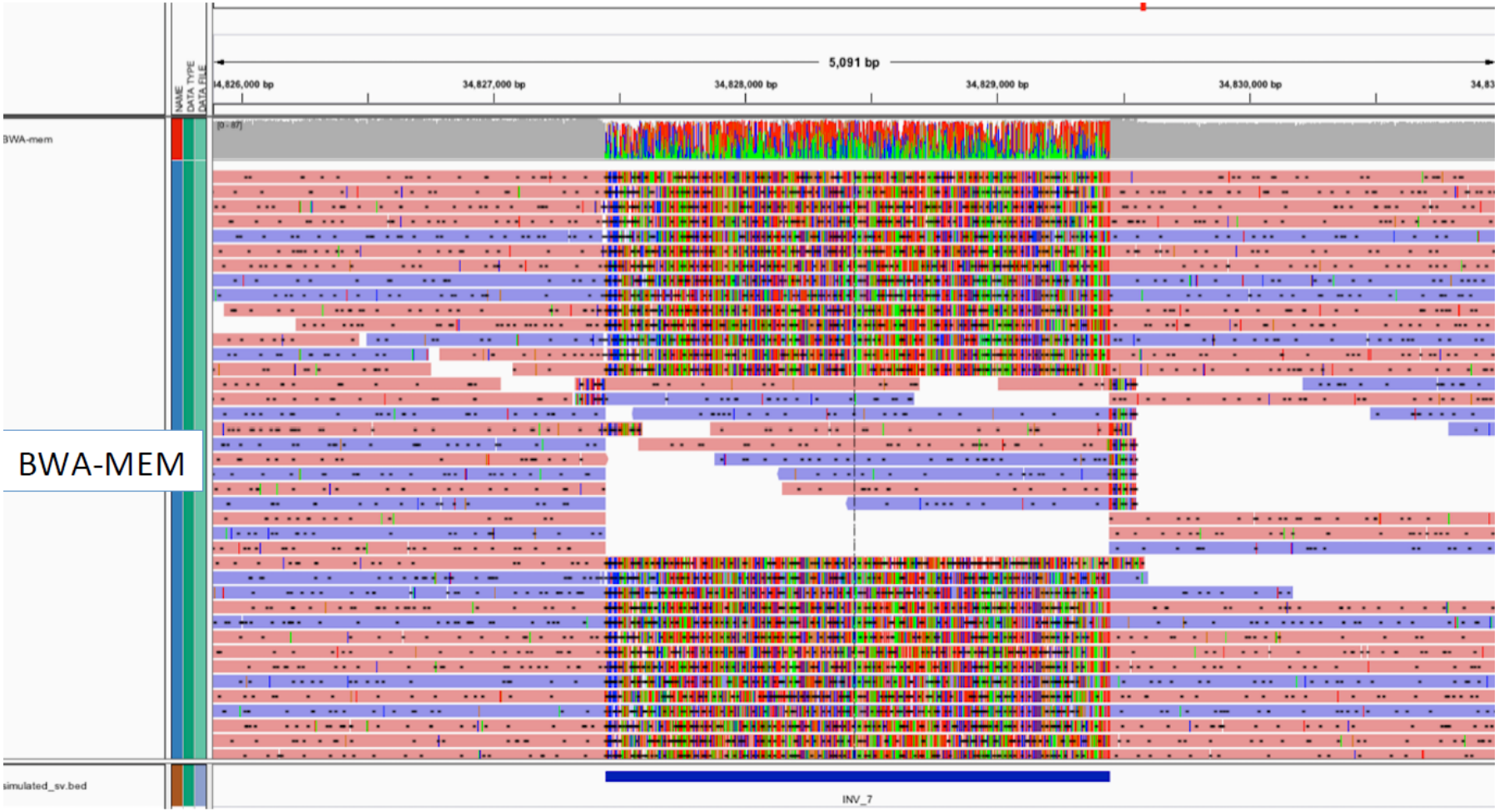
The false negative rate is usually hard to measure, but is thought to be extremely high for most paired-end mapping studies (>30%)

Long Read Technologies

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

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

Hard to align



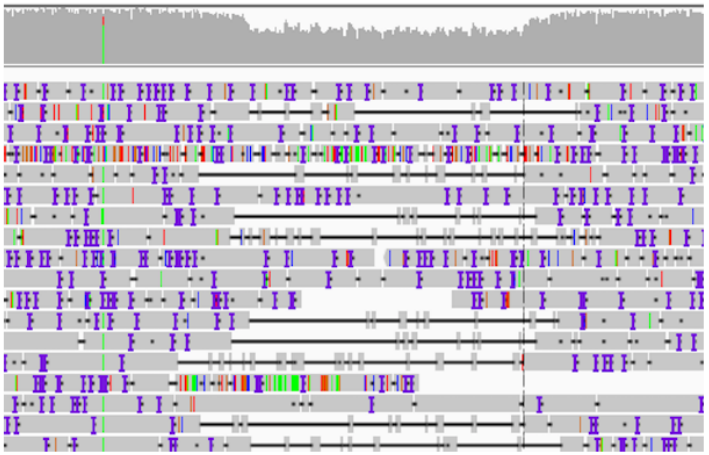
Human genome: 1kb Inversion

Improving long read alignment

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

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

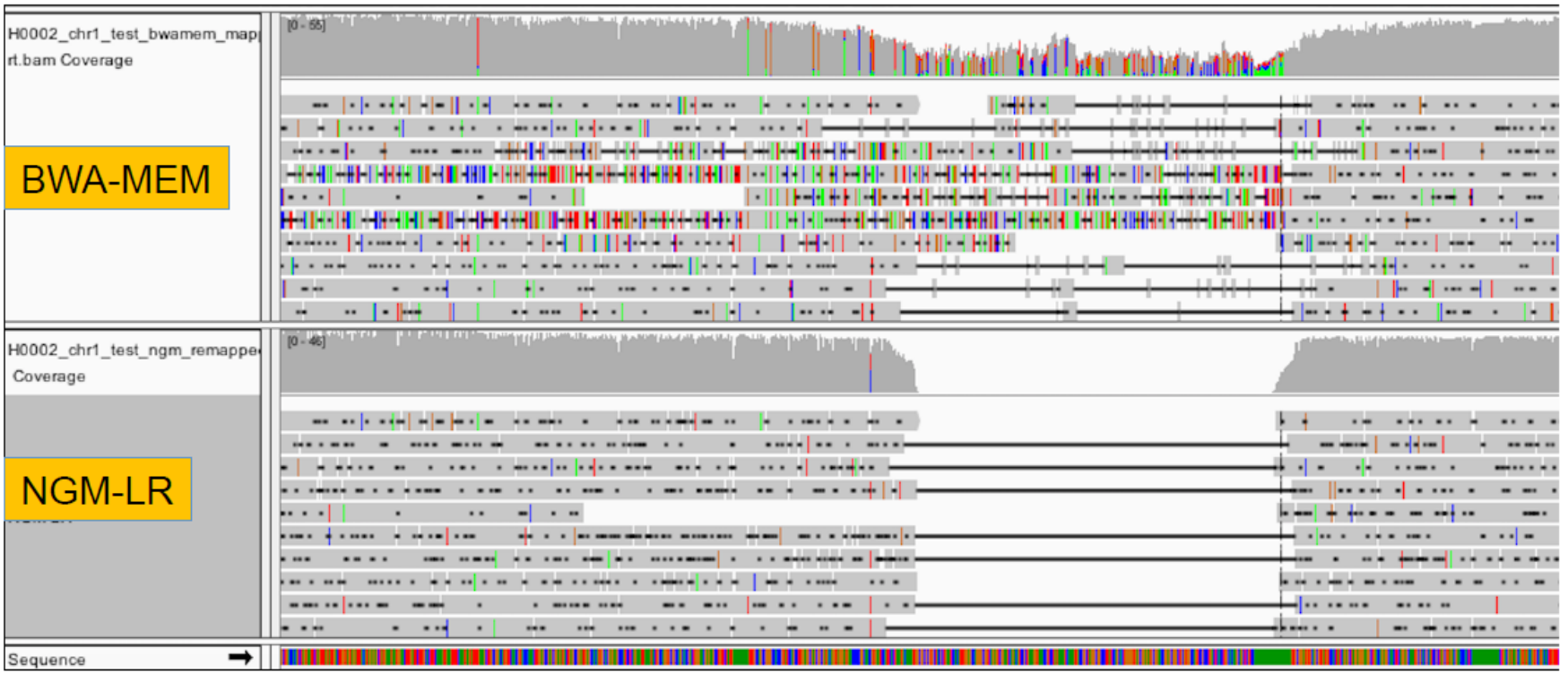
2. Improve alignment:
   - Insertions
   - Deletions



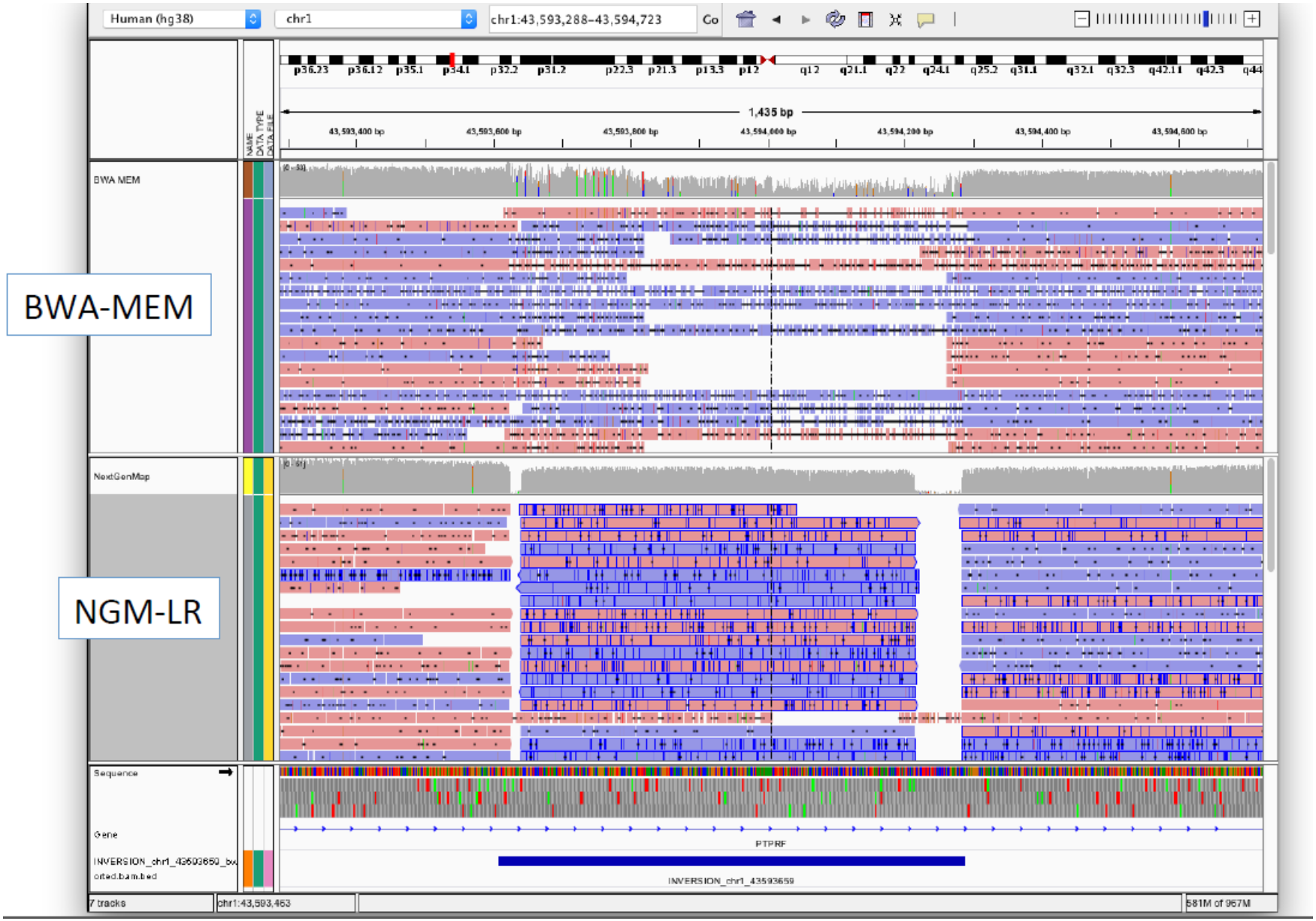
Structural variations with 3rd gen sequencing

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



Benchmark for structural variant calling

12,745 isolated events

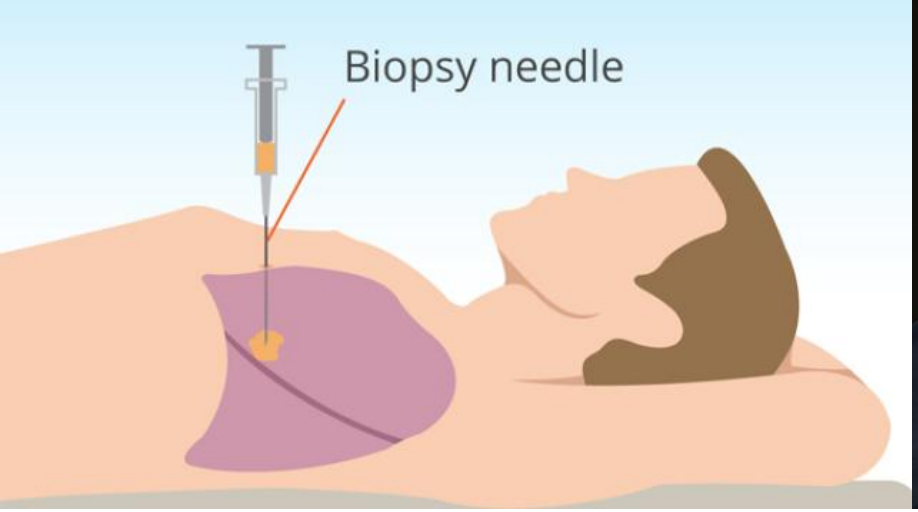
- 7,281 insertions
- 5,464 deletions

Used 19 sequence-resolved variant calling methods from diverse technologies

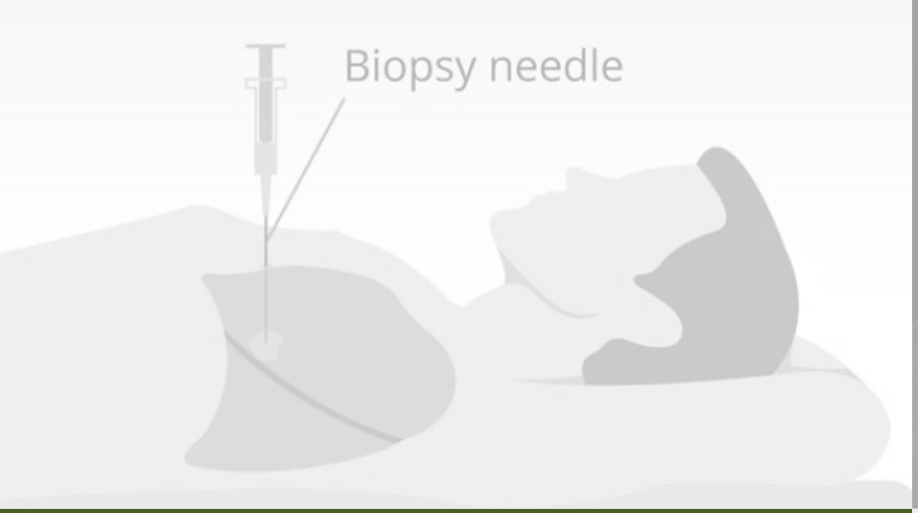
Already used to evaluate new tools (ClinSV, SVIM-asm, PopDel, ClipSV …)

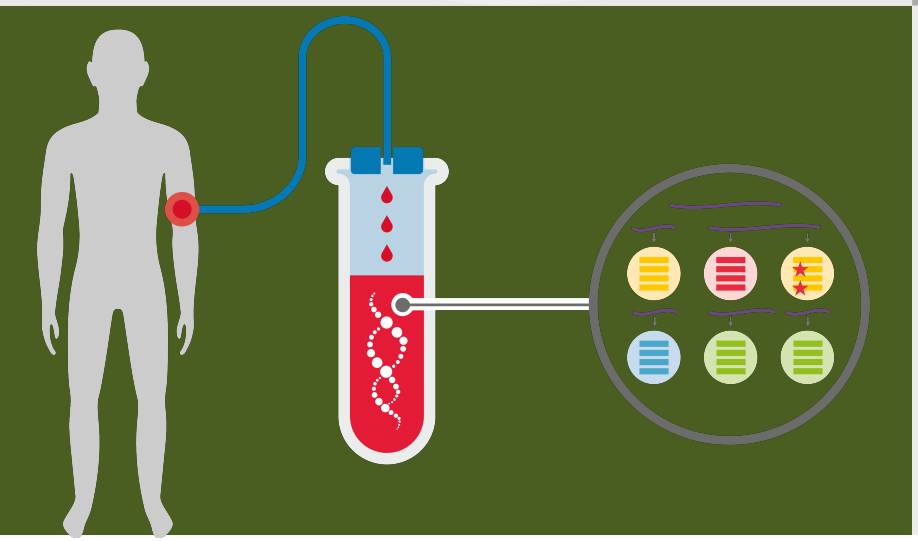
Zook, J.M., Hansen, N.F., Olson, N.D. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38,** 1347–1355 (2020).

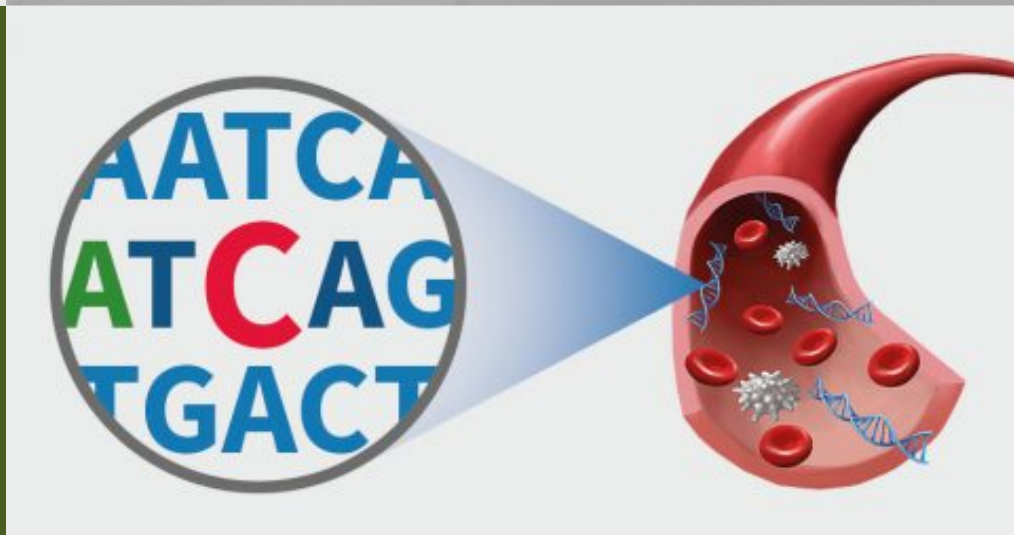
Liquid biopsies – CNV detection



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

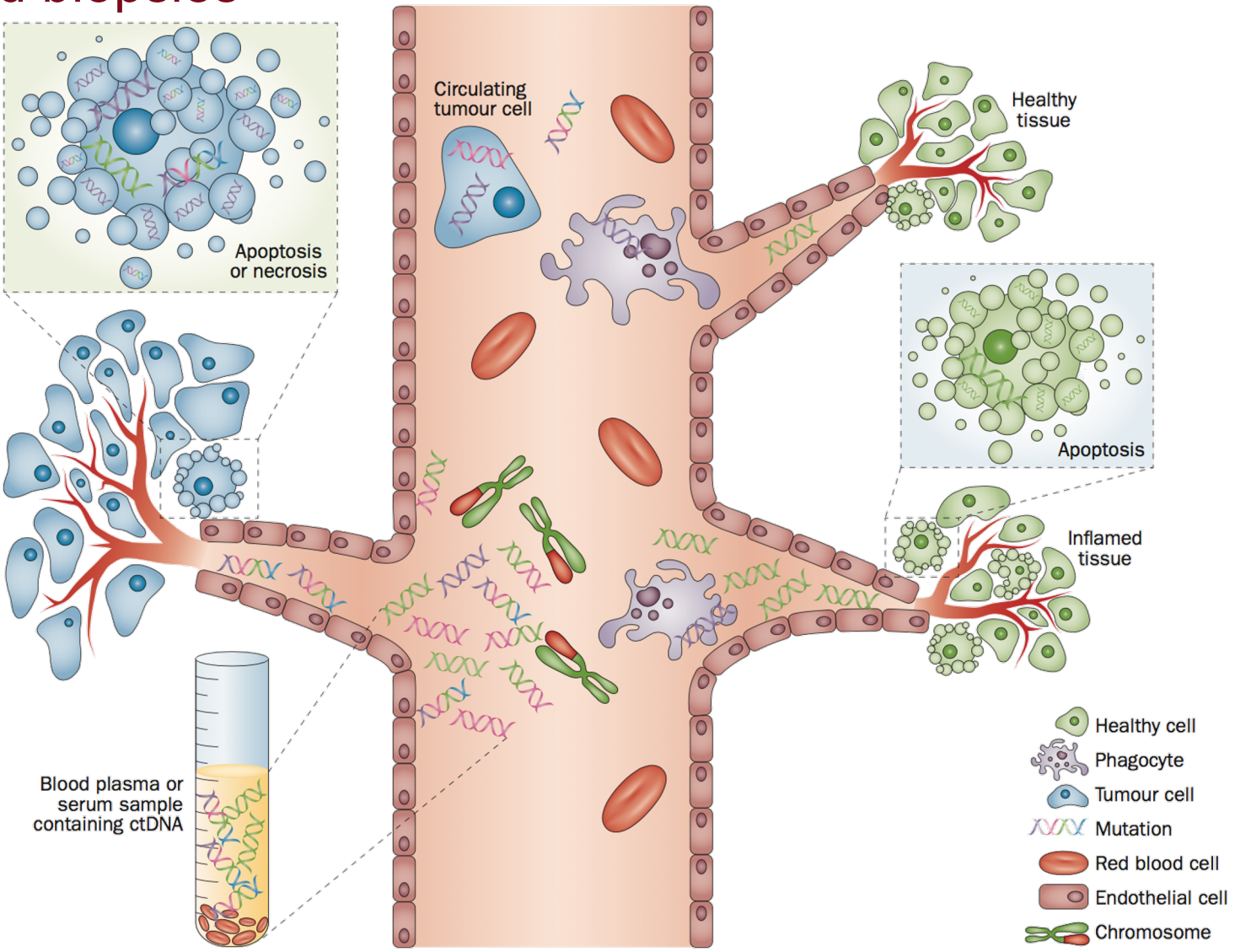






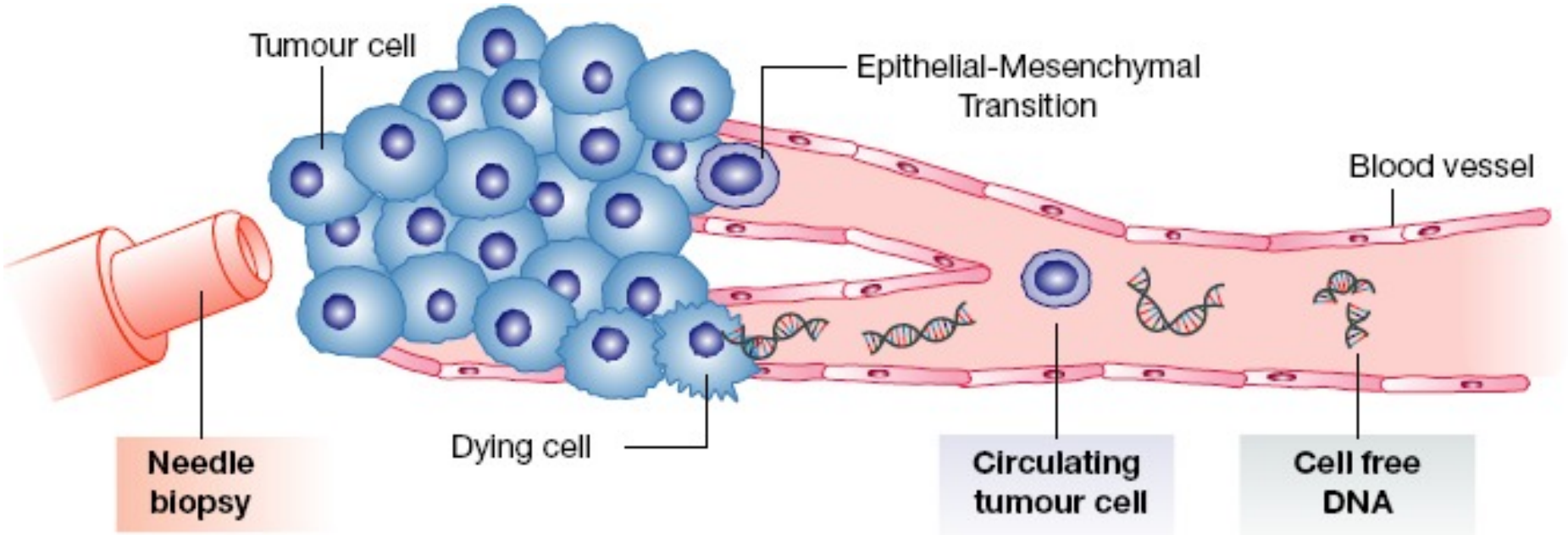
http://medifitbiologicals.com/wp-content/uploads/2015/11/BIOPSY-2.jpg
https://fthmb.tqn.com/WQWRqsQfxwQKOCgZXALeLoYcTow=/768x0/filters:no_upscale()/about/iStock_000001373663_Large-56a5c6145f9b58b7d0de6b32.jpg
http://www.gene-quantification.de/wyatt-gleave-liquid-biopsy-2105.jpg

Liquid biopsies



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

Liquid biopsies



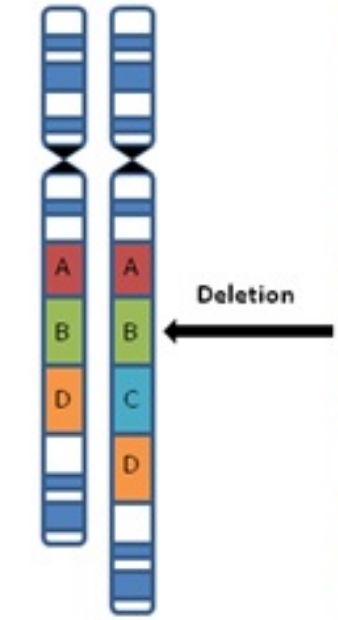
	Biopsy	CTC	cfDNA
Invasive	+	–	–
All patients eligible	–	+	+
Instrumentation required	+	+	–
Biomarker applicability	–	++	+++

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

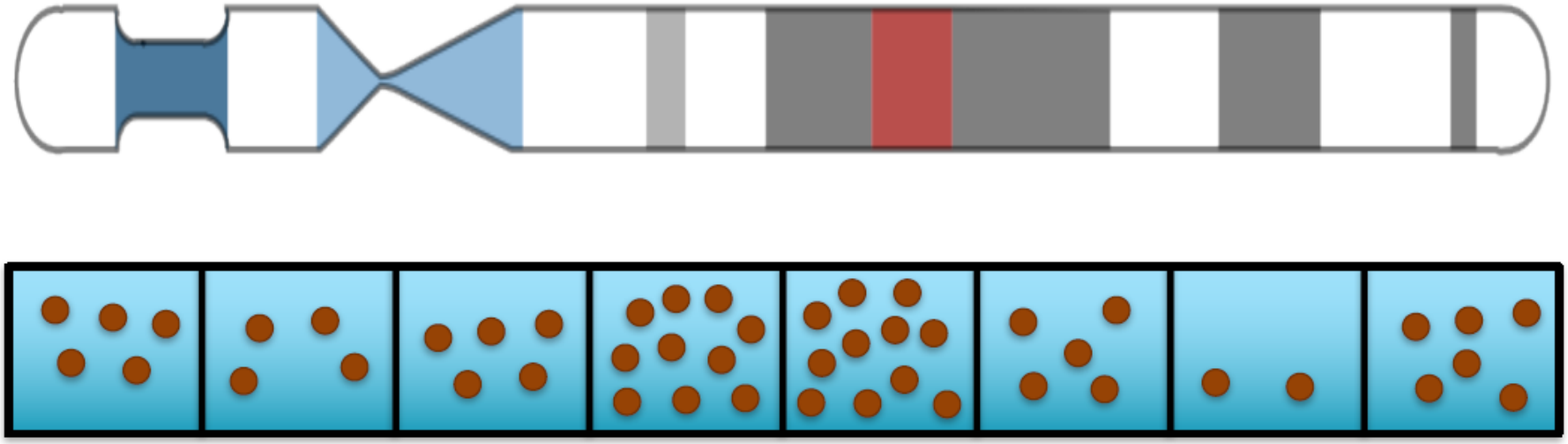


CNVs associated with complex diseases

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

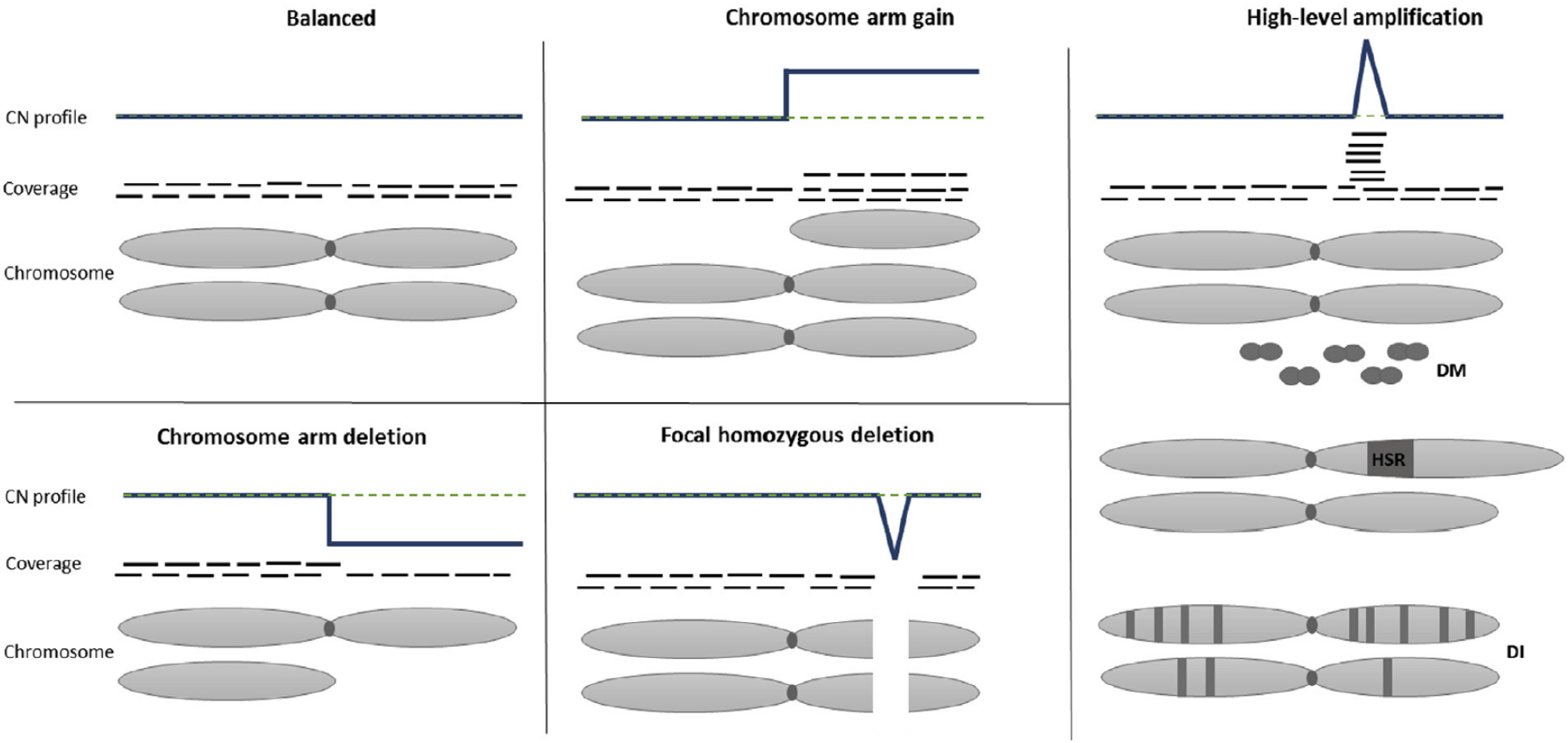
http://readingroom.mindspec.org/wp-content/genetics_CNV.jpg
Human Copy Number Variation and Complex Genetic Disease. Annual Review of Genetics. 2011

CNV Analysis - Overview



- Divide genome into bins
- Map reads
- Count reads
- Normalization against healthy controls
- Segmentation -> identify gains and losses

What can be detected

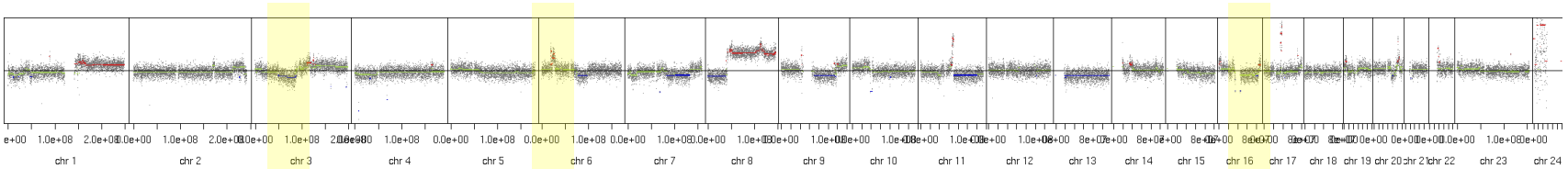


Non-invasive detection of genome-wide somatic copy number alterations by liquid biopsies

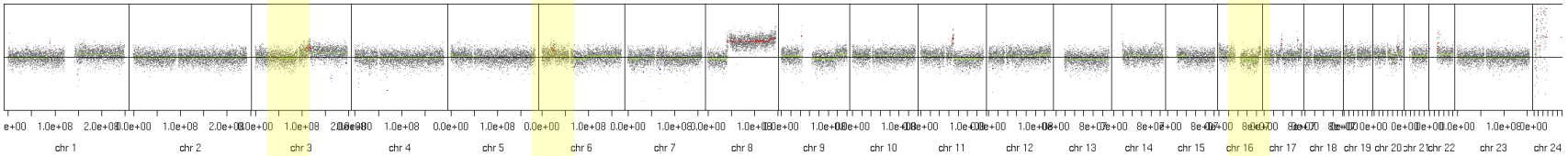
Time series analysis

Breast cancer patient (F)

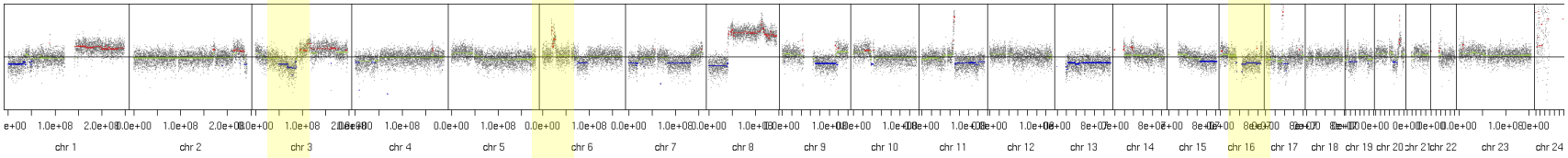
Timepoint 1



Timepoint 2



Timepoint 3



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/

Lancet

- Uses colored de Bruijn graphs
- Better accuracy, especially for indel detection, than widely used somatic callers

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/

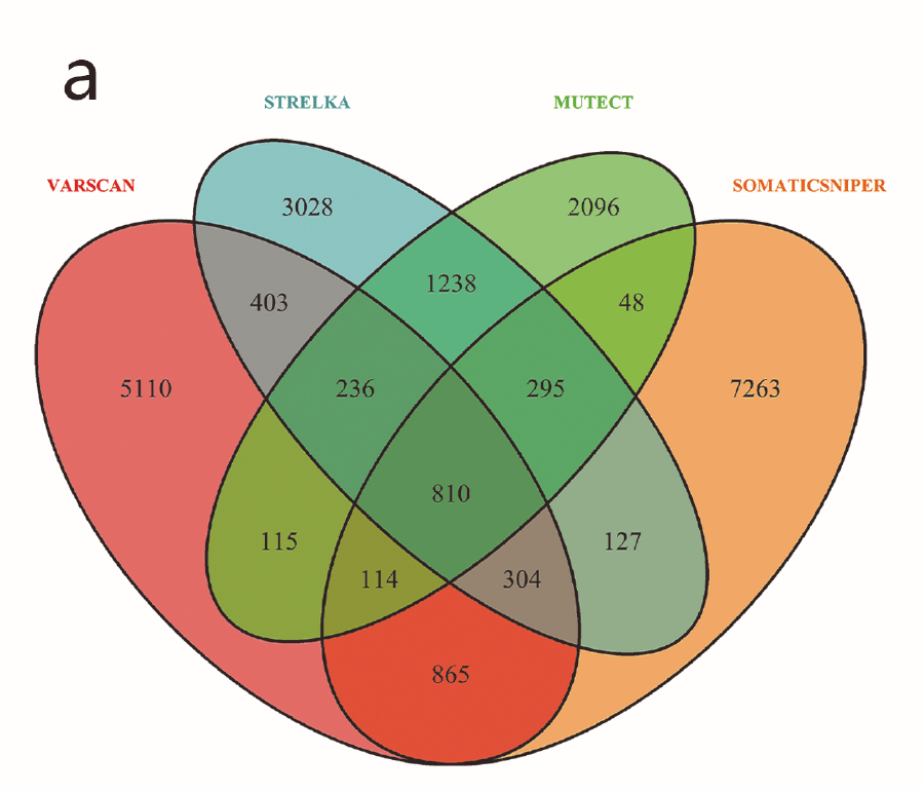
Here are a few more, a summary of the other answers, and updated links

- 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)

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

Deep learning → refinement of somatic variant calling from cancer sequencing data

- Training dataset of 41,000 variants from 21 studies, with 440 cases derived from nine cancer subtypes (manually reviewed by researchers)
- Is used to remove FP variant calls
- Random forest and deep learning models → both high classification performance

When employing the classifier on **new datasets**
→ manually reviewing or performing validation sequencing for a small subset of variants called via statistical variant callers (for example, 5% of all data)
→ re-train the classifier and improve performance

https://www.nature.com/articles/s41588-018-0257-y

Variant filtering

What information is needed to decide if a variant exists?

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

Variant filtering

The biggest problem is large numbers of FPs:

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

FPs and FNs, may result in:

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

→ **Filter**

Variant filtering – how to

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

But - there will also be FP with high QUAL

Signs of suspicious variants

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

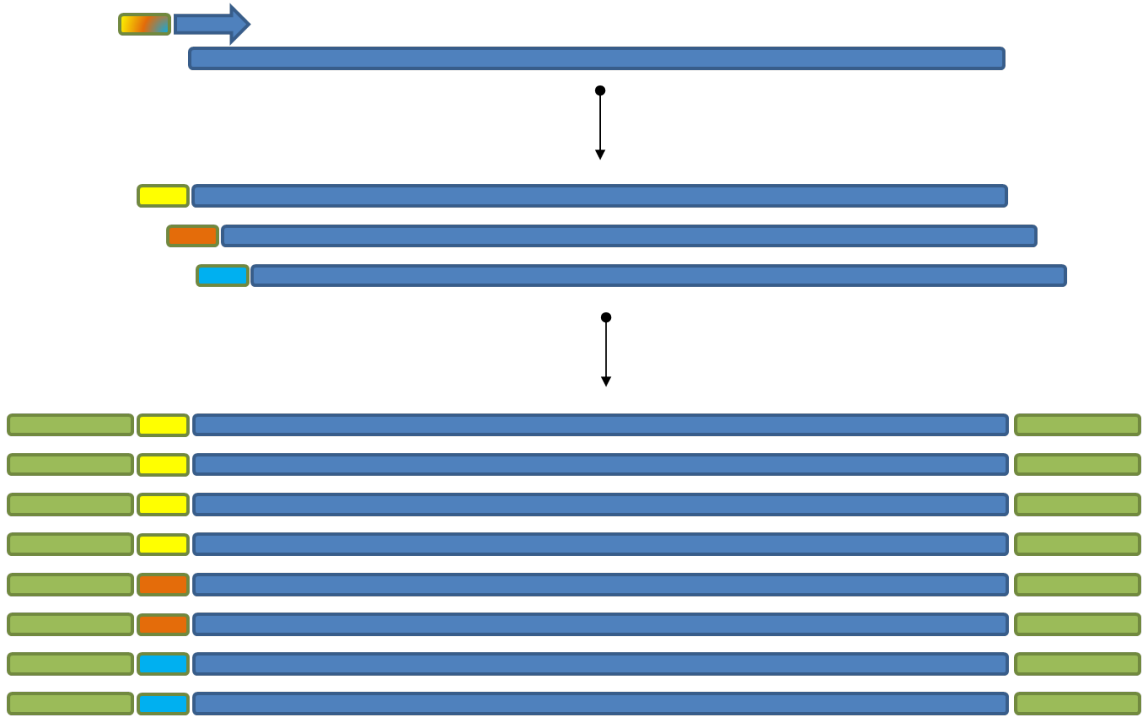
More information: https://www.broadinstitute.org/gatk/guide/tagged?tag=VQSR

Molecular barcoding

Library generation

STEP 1 - Barcoding

STEP 2 - Amplification



Consensus sequence generation

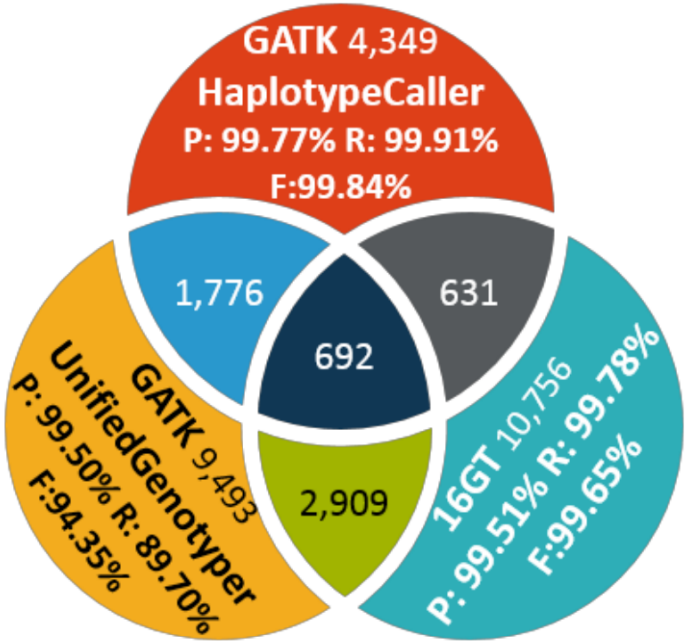
random barcode mix unique barcodes sequencing adaptors

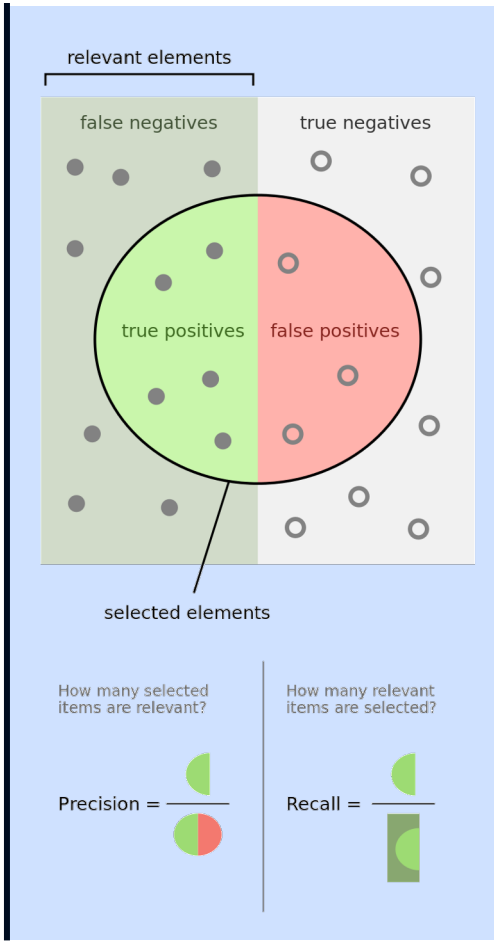
Adapted from http://www.imgm.com

Skyhawk/Clairvoyante

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

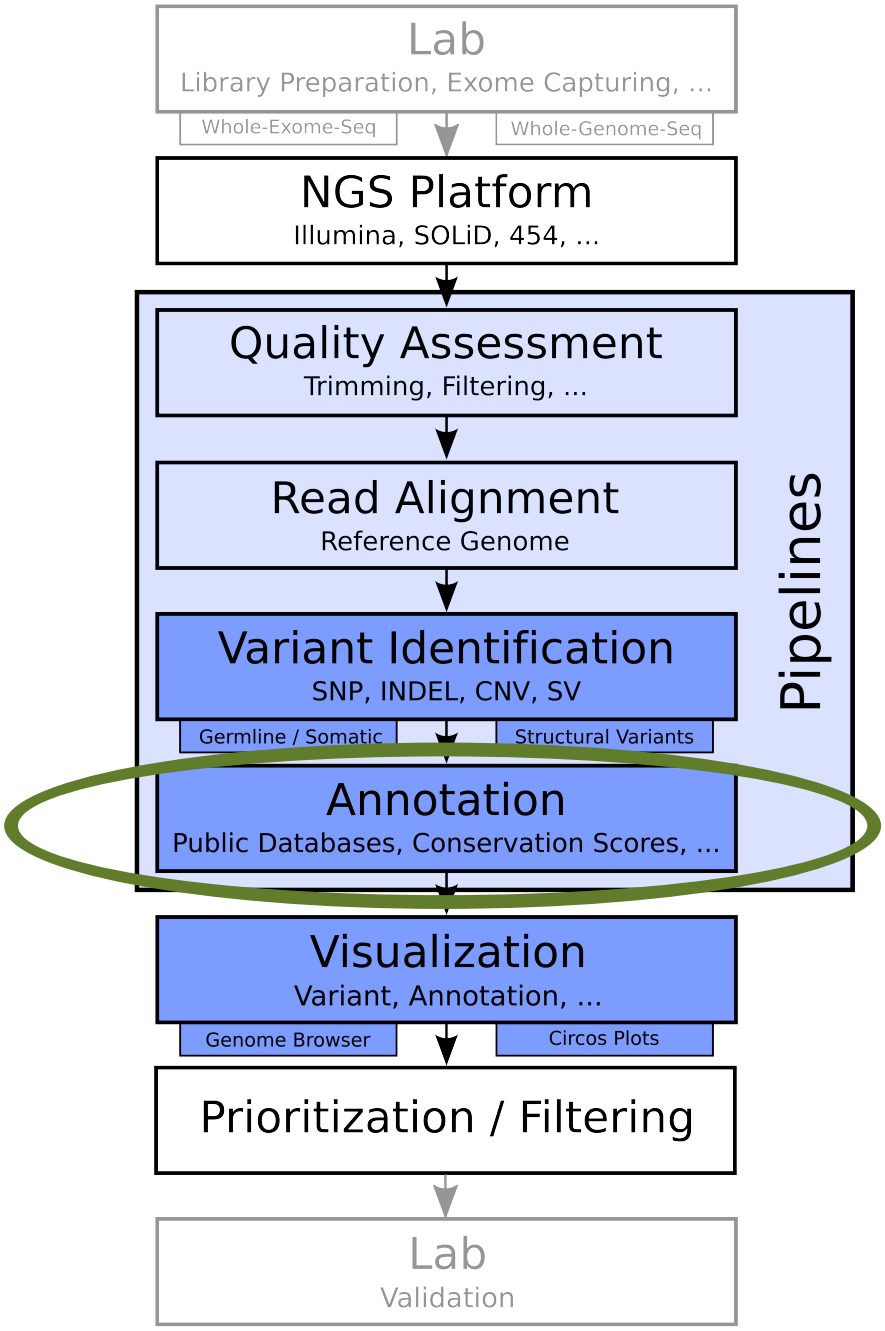
- Helps removing FP results





Luo, R., Sedlazeck, F.J., Lam, TW. et al.
A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun 10, 998 (2019).

Variant annotation


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

Basis for Molecular Assay: Pathogenesis

Understanding molecular pathogenesis of human disease enables effective utilization of molecular assays

Diagnostic

- Distinguishing variants of human disease based on presence of specific molecular markers (chromosome translocations in Burkitt's lymphoma)

Prognostic

- Prediction of likely patient outcomes based on presence of specific molecular markers (gene mutations predicting clinical course in cancer)

Therapeutic

- Prediction of response to specific therapies based on presence of specific molecular markers (gene mutations predicting poor drug sensitivity in lung cancer: p53, k-ras)

Variant annotation

After variant calling → **many** variants

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

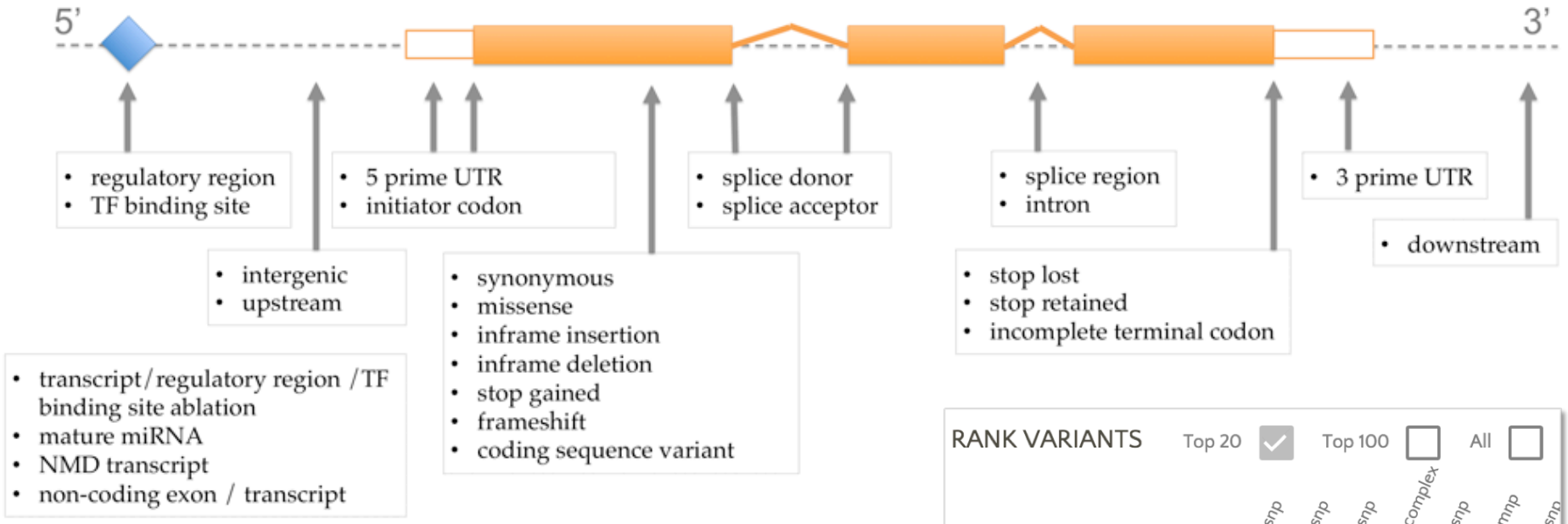
Annotation

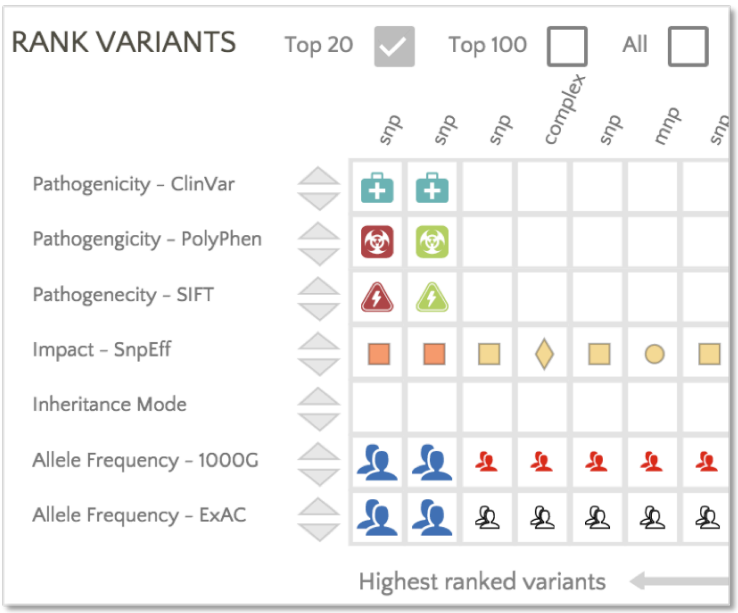
- Basis for filtering and prioritizing potential disease-causing mutations
- Most tools focus on the annotation of SNPs
- Many provide database links to various public variant databases (dbSNP...)
- Functional prediction of the variants
  - Sequence-based analysis
  - Region-based analysis
  - Structural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

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

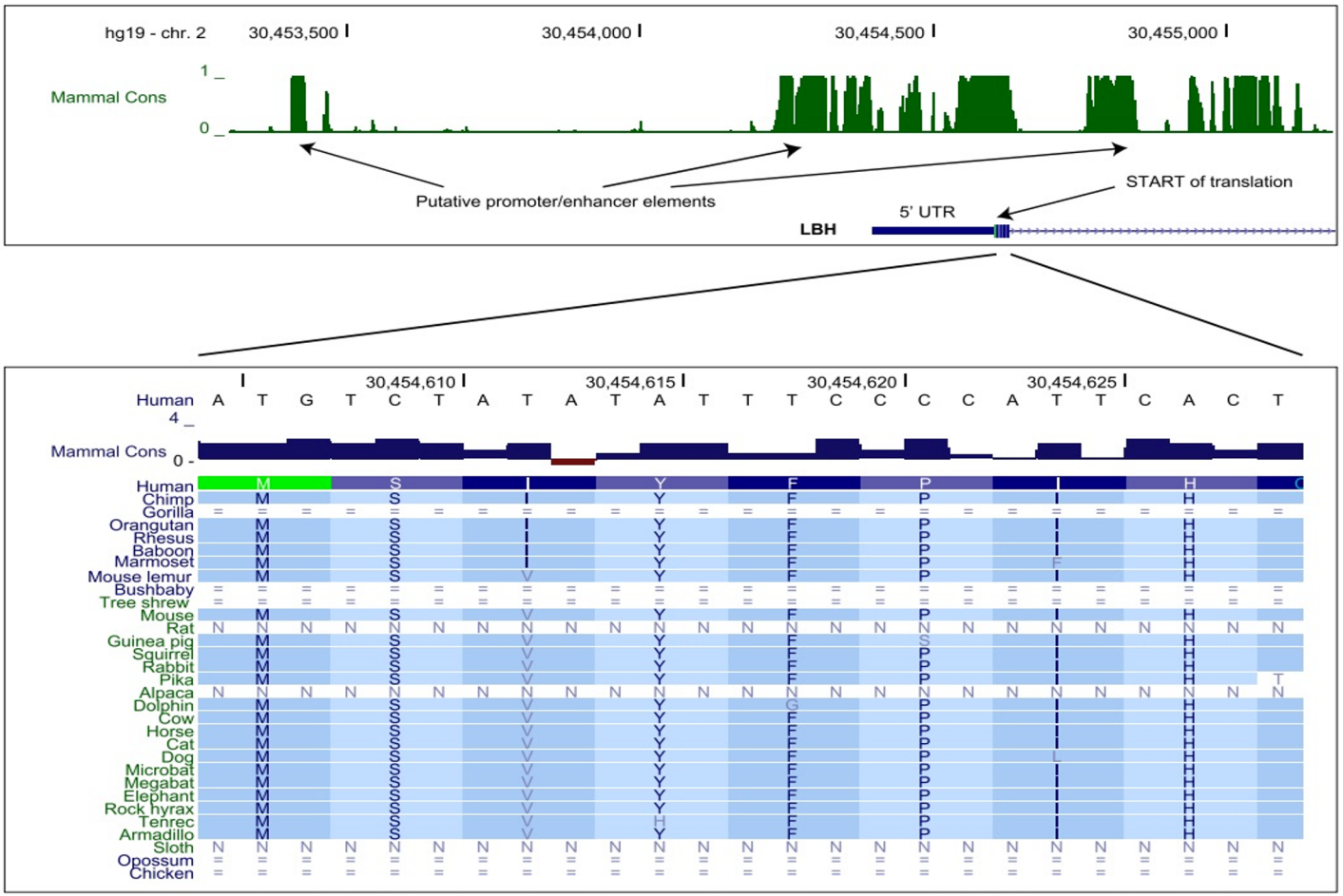


Figure 1. A comparative genomics display derived from the UCSC Genome Browser (Meyer et al. 2013). The *top* panel depicts the genomic region surrounding the 5′ end of the gene *LBH* (limb bud and heart development homolog) in the human genome. The *top* track indicates mammalian conservation as determined 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.

Consequences of mutations

Missense mutations differ in severity

- Conservative amino acid substitution:
  substitutes chemically similar amino acid, less likely to alter function
- Nonconservative amino acid substitution:
  substitutes chemically different amino acid, more likely to alter function
- Consequences for function; often context-specific

Nonsense mutation results in premature termination of translation

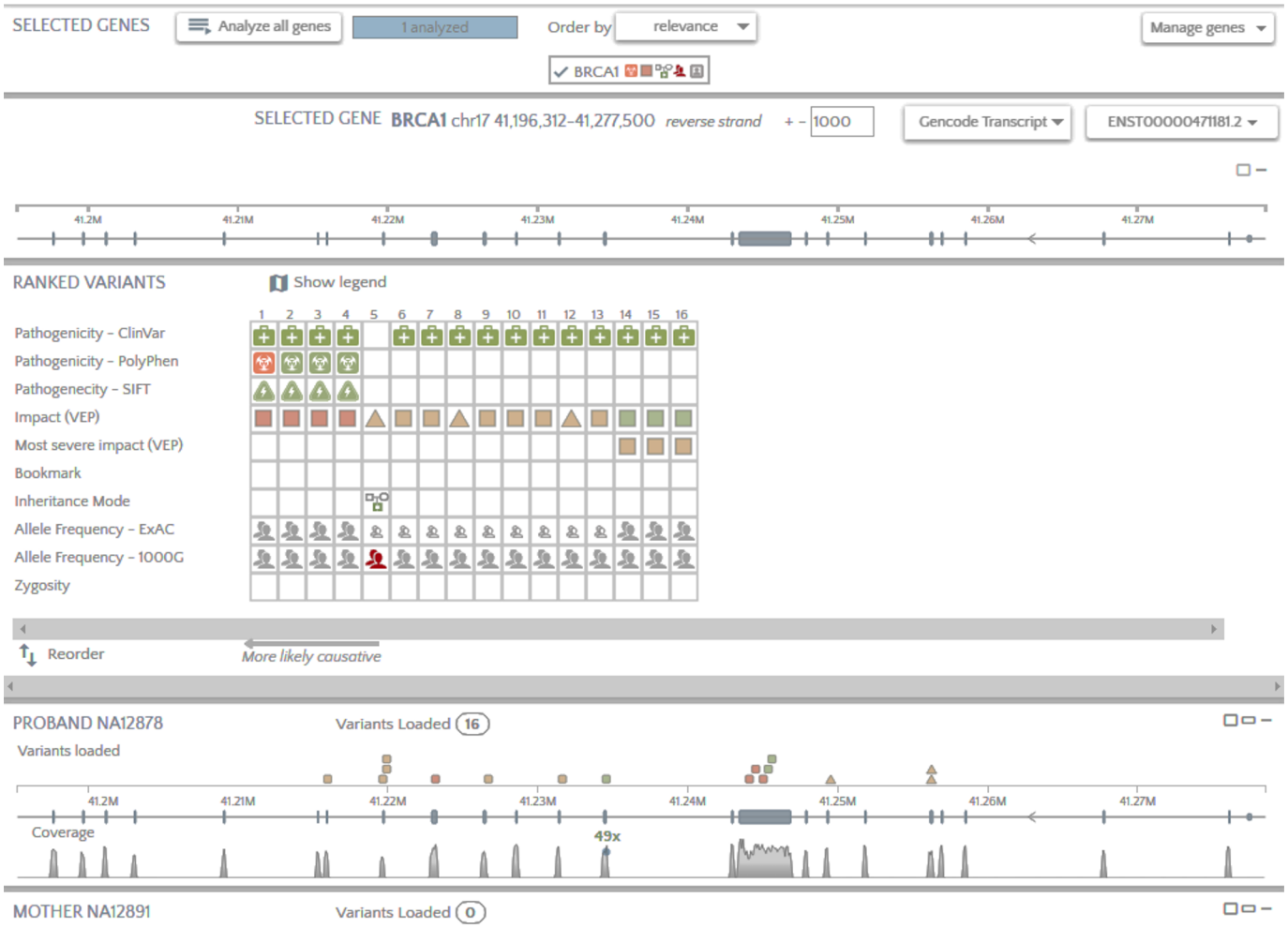
- Truncated polypeptides often are nonfunctional

Point mutation in non-coding region may affect transcription, RNA splicing, and protein assembling

Gene annotation

Gene.iobio.io

- Interactive
- Load custom data



dbSNP

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

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

Submissions

- ~150m variants (stats only for human) [v147, 2016]
- ~660m variants (stats only for human) [v151, 2018]
- ~730m variants (stats only for human) [v154, 2020]

https://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi
https://ncbiinsights.ncbi.nlm.nih.gov/2020/06/11/dbsnp-human-build-154-release-alfa-data/

Variant annotation - tools

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

Variant annotation - tools

ANNOVAR

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

SeattleSeq Annotation server

- Online tool
- Human SNPs and INDELs

snpEff

- Integrated within Galaxy and GATK.
- SNPs and INDELs

MARRVEL

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

Variant annotation - tools

SVScore

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

VEP

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

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

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

Variant annotation - tools

ONCOTATOR

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

Exomiser

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

LOFTEE

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

Vcfanno

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

ExAC

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

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

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

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

gNOMAD

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

CADD

Combined **A**nnotation **D**ependent **D**epletion

- 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

Pathway Commons

- Collect and disseminate biological pathway and interaction data
- Includes
  - biochemical reactions
  - assembly of biomolecular complexes
  - transport and catalysis events
  - physical interactions involving proteins, DNA, RNA, and small molecules

Pathway Commons 2019 Update: integration, analysis and exploration of pathway data

Igor Rodchenkov, Ozgun Babur, Augustin Luna, Bulent Arman Aksoy, Jeffrey V Wong, Dylan Fong, Max Franz, Metin Can Siper, Manfred Cheung, Michael Wrana ... Show more

Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D489–D497,

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)
- start (starting position on the chromosome)
- end (end position of the chromosome, note this base is not included!)

Used for

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

Bedtools

Tool suite

- Contains multiple tools for a wide-range of genomics analysis tasks
- *Genomic intervals*
  - *Intersect*
  - *Merge*
  - *Count*
  - *Complement*
  - *Shuffle*
- https://bedtools.readthedocs.io/en/latest/

Tutorial

- http://quinlanlab.org/tutorials/bedtools/bedtools.html

Assignment 3

Visualization

Visualization

Genome browsers - most widely-used tools

Read

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

Able to

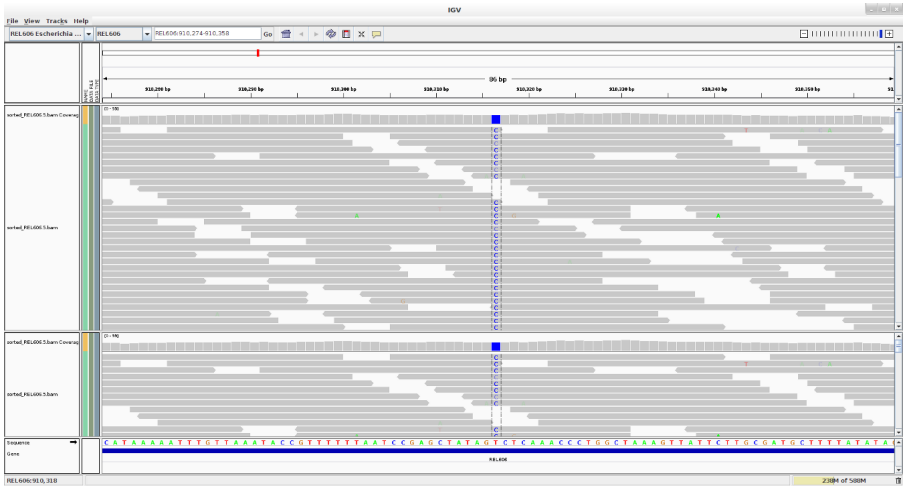
- Browse/zoom genome
- Display multiple samples / multiple tracks
- Colorize/mark features of your data (paired reads, SNPs, ...)

Genome Browsers

IGV (Integrative Genomics Viewer)

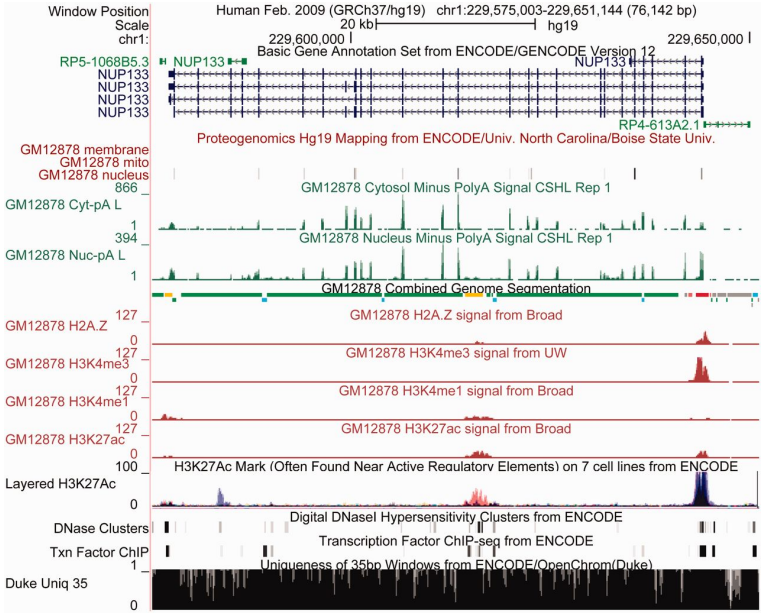
- Widely used viewer
- Java based – standalone tool
- Easy and fast to view own data
- IGV3 **supports long-reads**

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



UCSC

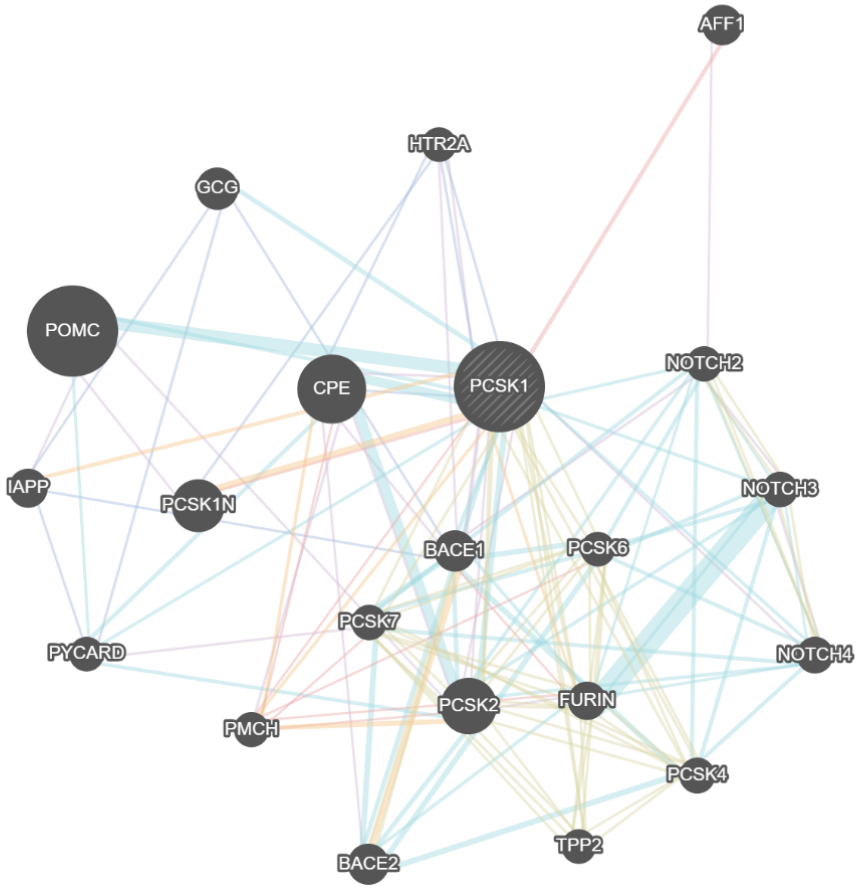
- Web based tool
- Offers many different annotation tracks
- Needs some configuration to display own data



GeneMANIA - Network of genes

- Enter multiple genes

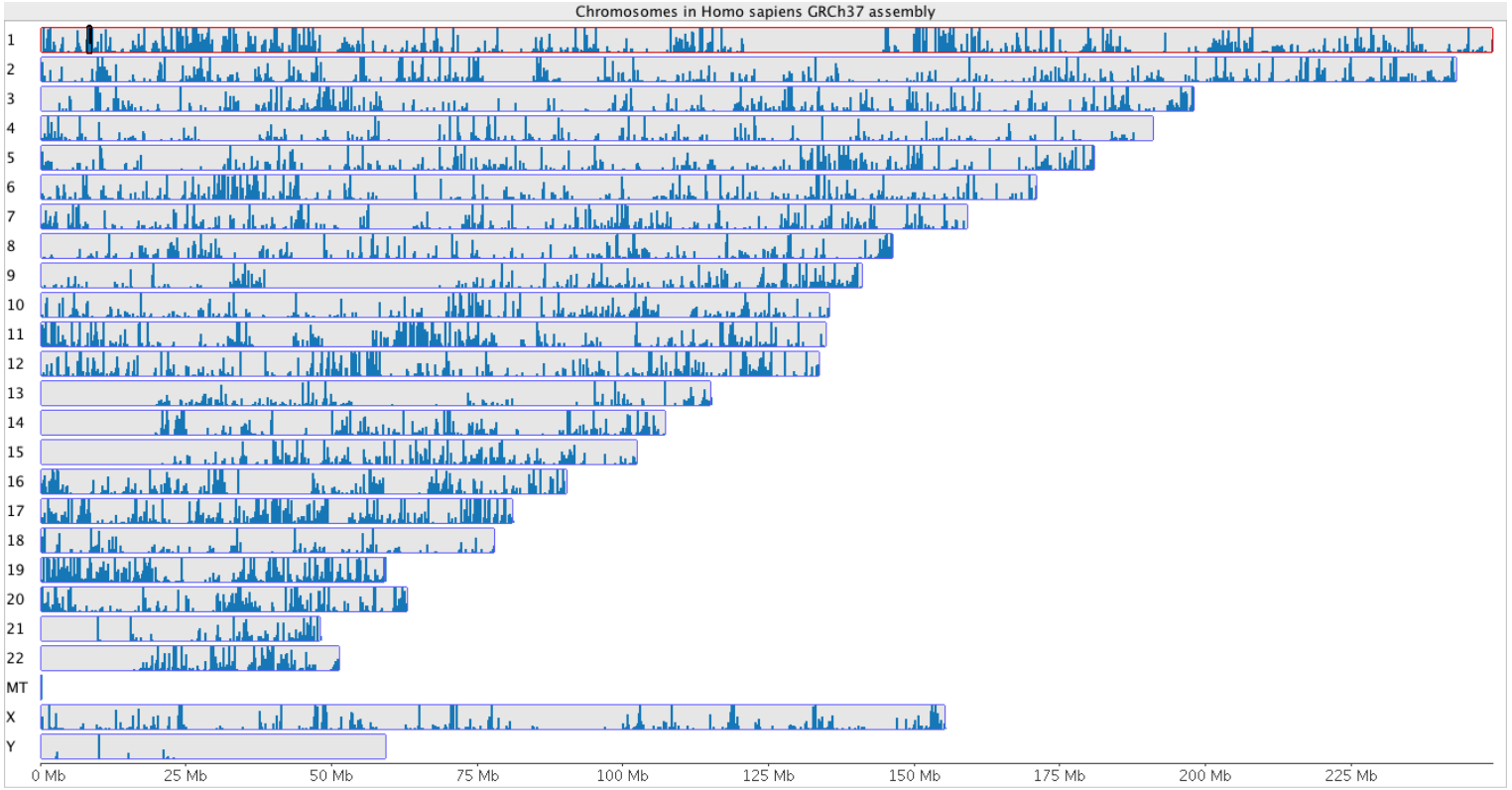
- Interpret their context



Coverage visualization

Coverage histogram for chromosomes

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

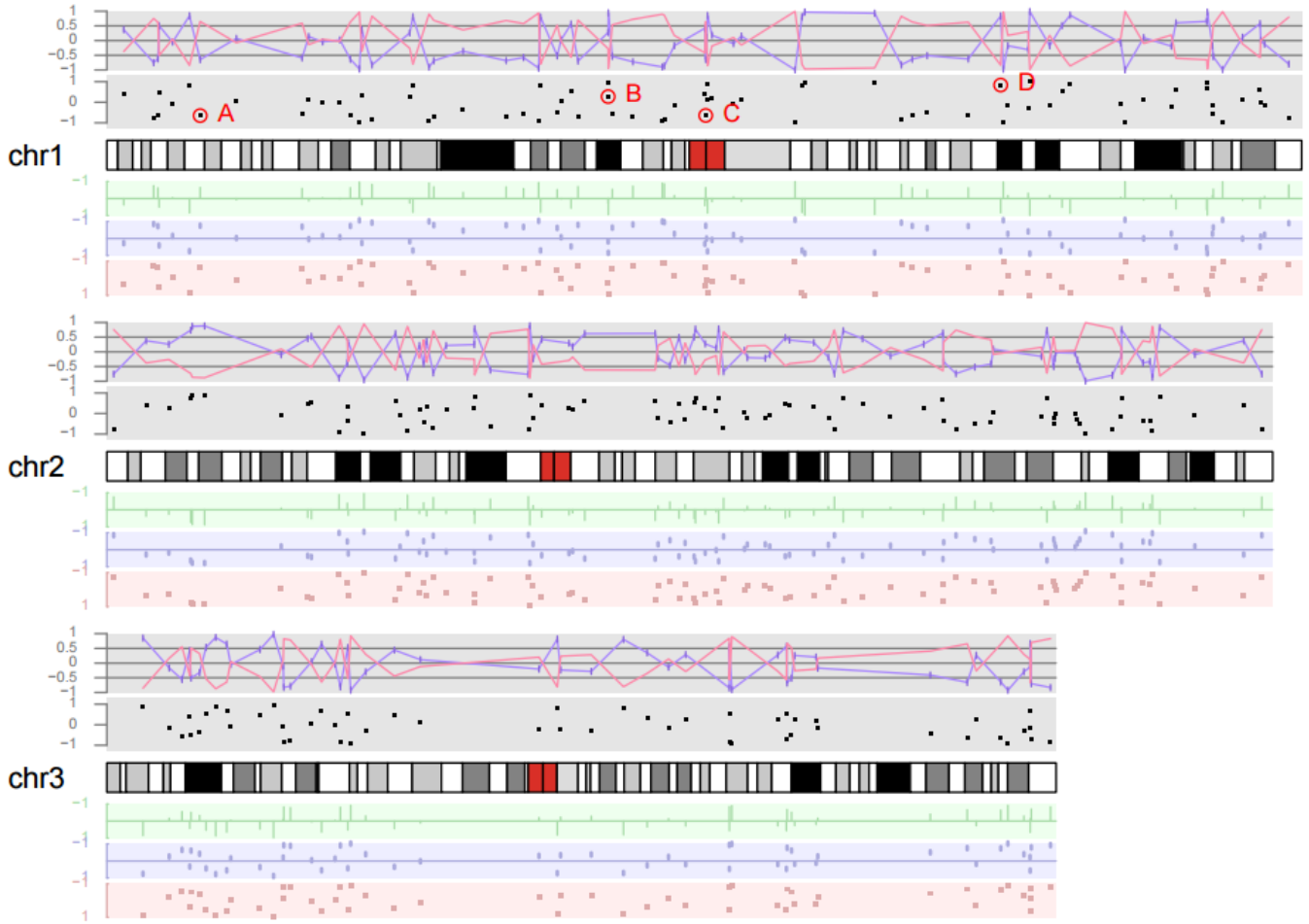


http://seqanswers.com/forums/attachment.php?attachmentid=2118&d=1364889859

Genome-wide visualization

karyoploteR

- Customizable karyotypes with arbitrary data

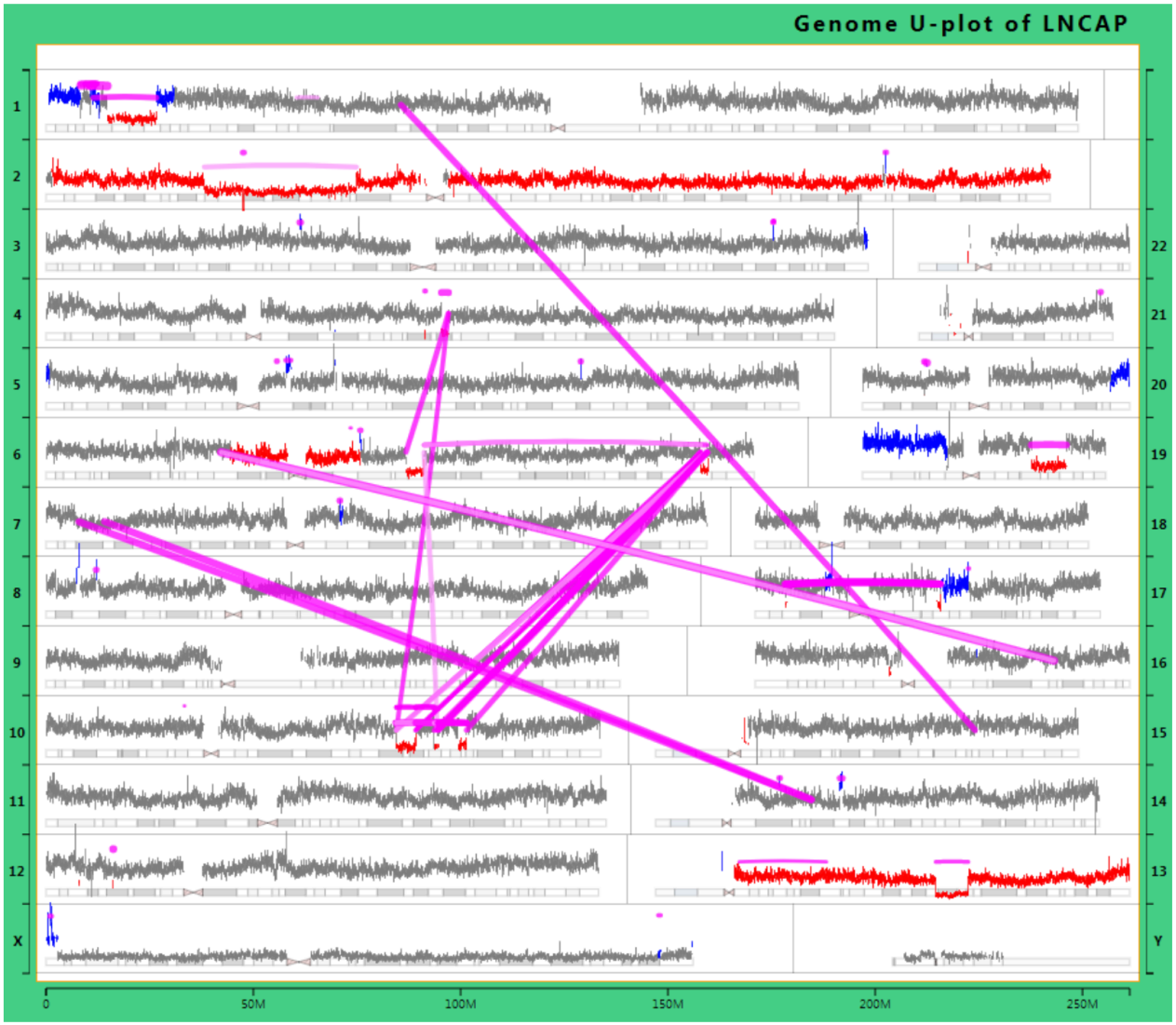


CNV Visulization

GenomeUPlot

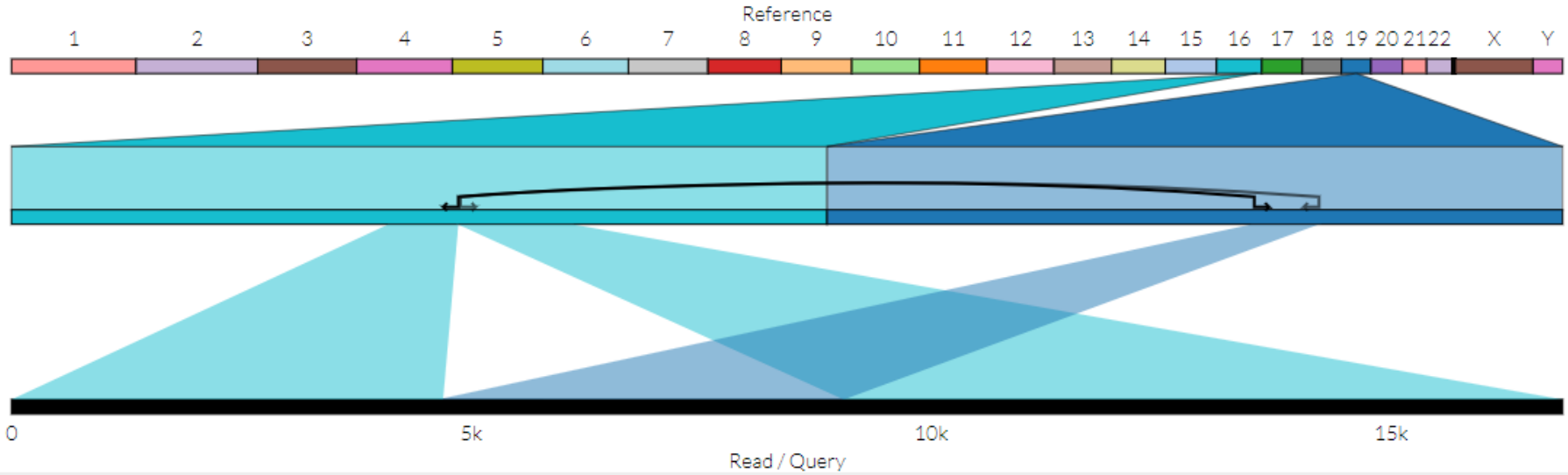
Visualize Chromosomal abnormalities

https://github.com/gaitat/GenomeUPlot



Visualization of structural variants

Ribbon: Visualizing complex genome alignments and structural variation



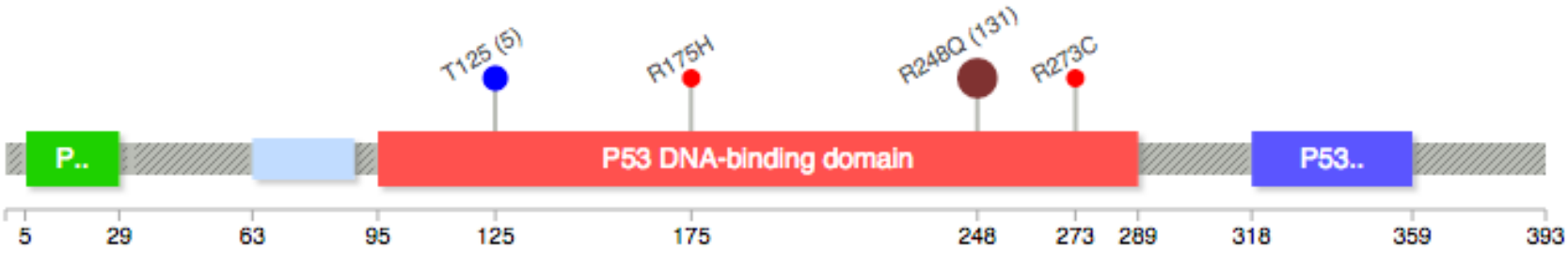
Maria Nattestad, Chen-Shan Chin, Michael C. Schatz; https://www.biorxiv.org/content/10.1101/082123v1

Lollipop

Lollipop-style mutation diagrams for annotating genetic variations

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

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



CircosVCF

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

- Interactive tool
- Web-based

Analysis pipelines and workflow systems

Pipelines

Bcbio pipeline

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

Nextflow

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

DNAp

- DNA and RNA analysis
- BWA, GATK, SV calling

ngs_backbone

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

Pipelines

SeqWare

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

Firehose

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

iGenomics - DNA Analysis on your smartphone

- Align reads
- Call variants
- Visualize the results
  - → entirely on an iOS device
- Benchmarked using real and simulated Nanopore sequencing datasets
  - Viral and bacterial genomes

iGenomics: Comprehensive DNA sequence analysis on your Smartphone

Aspyn Palatnick, Bin Zhou, Elodie Ghedin, Michael C Schatz ✉ Author Notes

GigaScience, Volume 9, Issue 12, December 2020, giaa138,
https://doi.org/10.1093/gigascience/giaa138
Published: 07 December 2020 **Article history** ▾

Galaxy

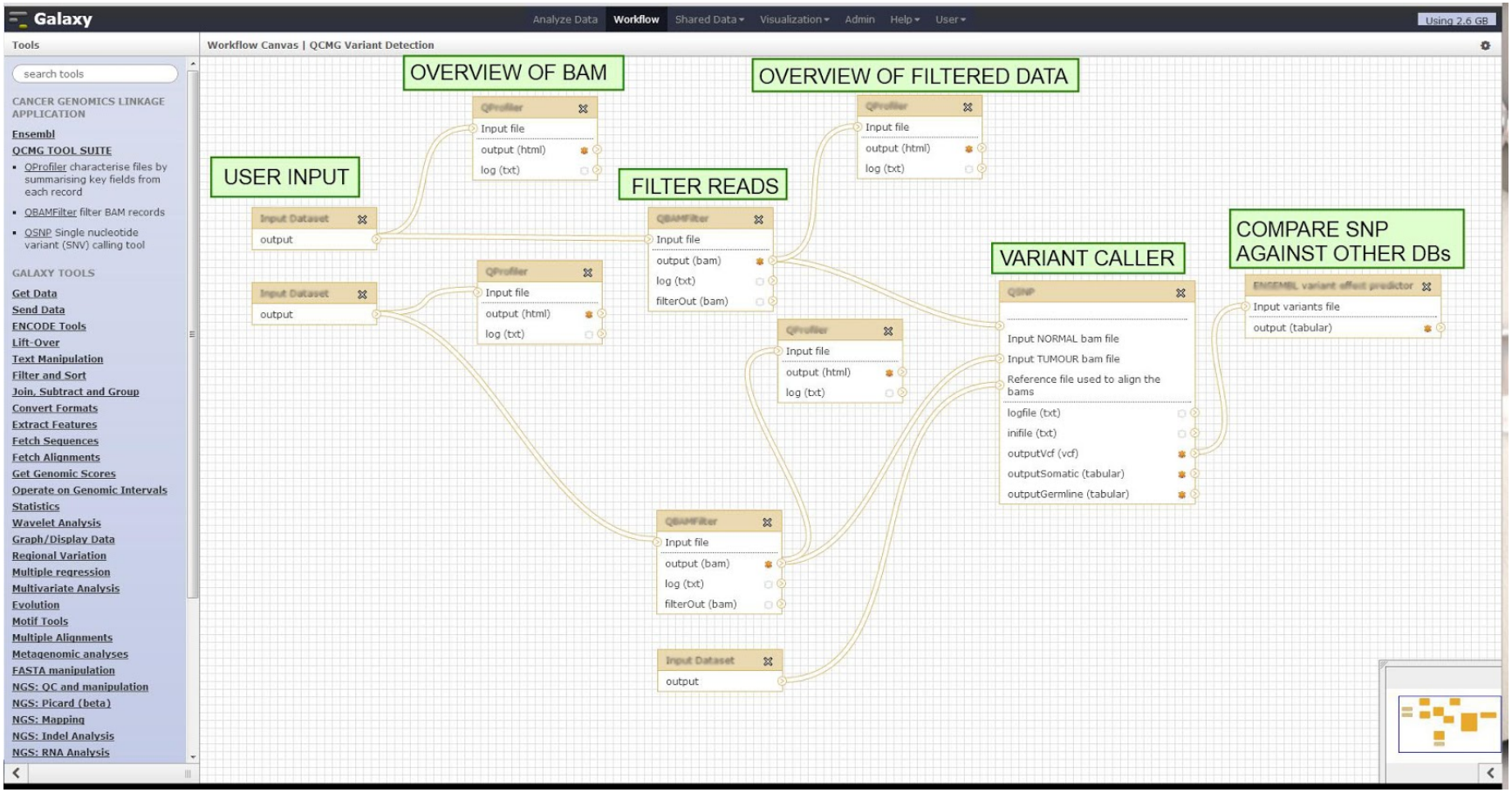
"Galaxy is an open, web-based platform for data intensive biomedical research"

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

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

- Can be locally installed and used

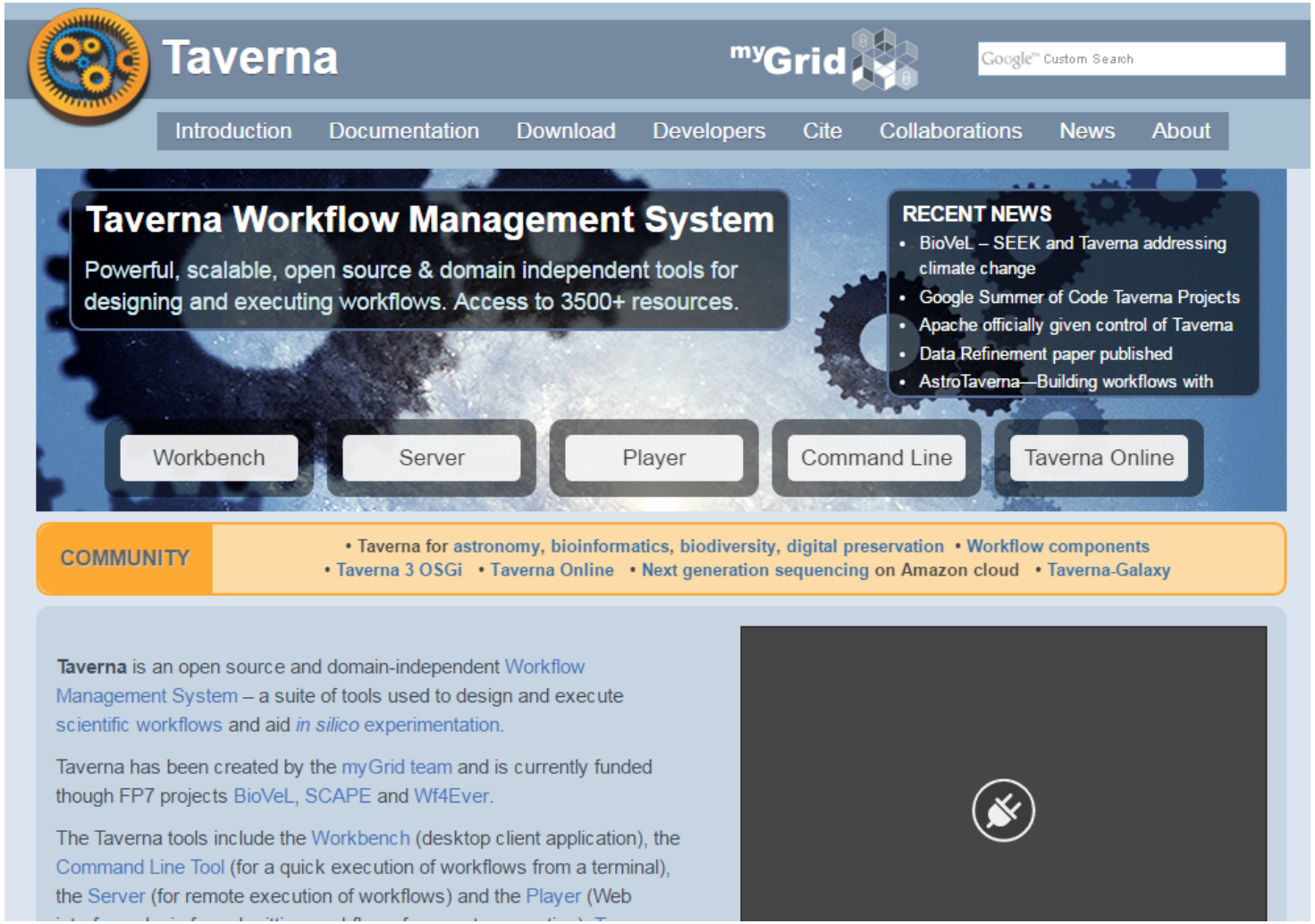
Galaxy



„The Cancer Genomics Linkage Application”

Taverna

- Workflow management system



Student projects

https://github.com/spabinger/medizinische_genomanalysen_2021

Assignments

- Send
  - Your name
  - The name of your Github repository
- to stephan.pabinger@gmail.com (or post it in the chat)

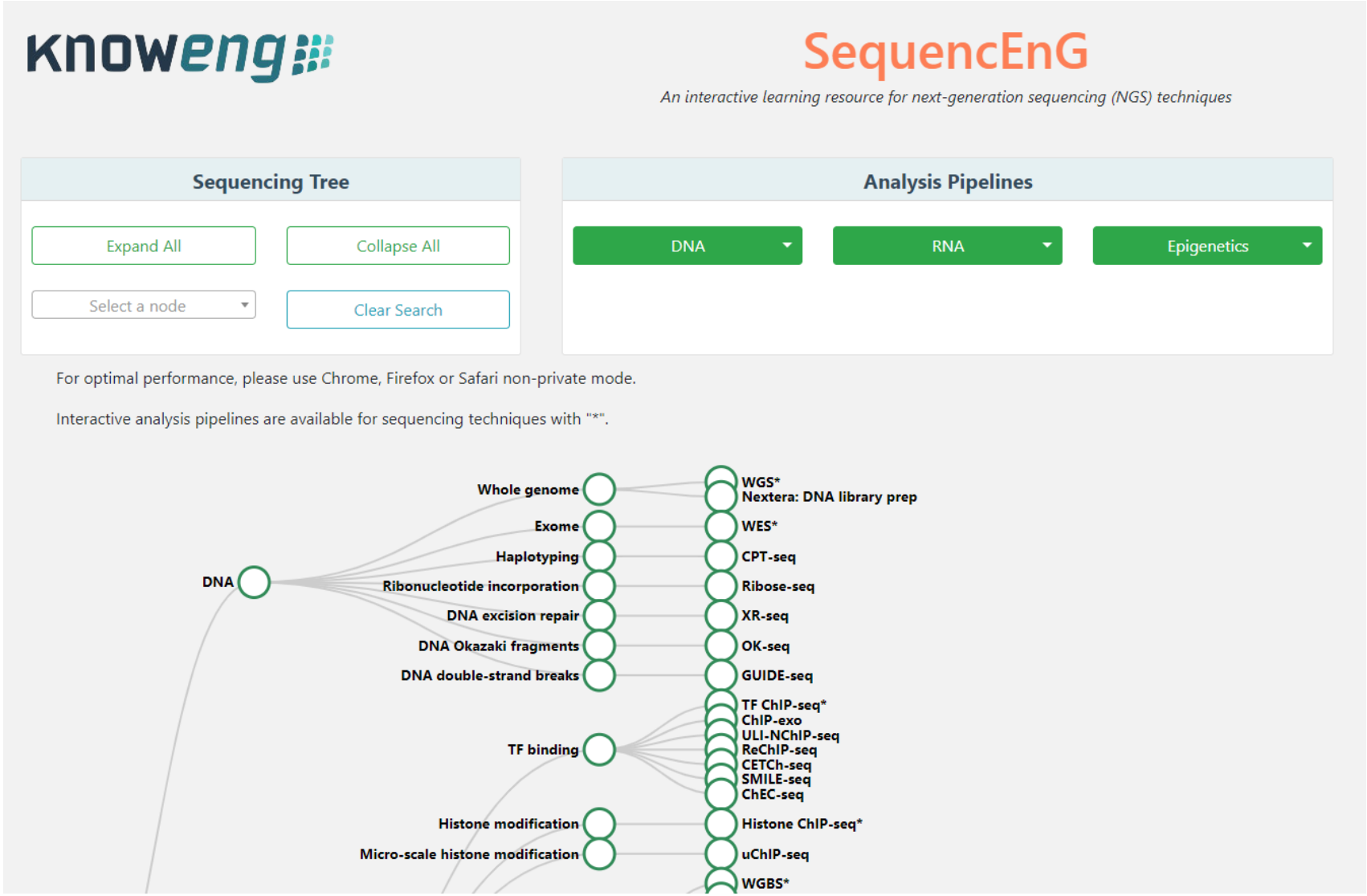
- **Deadline: 16.05.2021**

Email
- To: stephan.pabinger@gmail.com
- Subject: Medizinische Genomanalysen 2021

Other useful information

SequencEnG

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



Genome Austria

- Sequencing of 20 volunteers (currently 11 datasets)
- Data publicly available
- Interpretation of data

rs1799930(A;A) -> Increased risk of Age Related Hearing Impairment(ARHI), also known as presbycusis.
http://www.snpedia.com/index.php/Rs1799930
http://www.ncbi.nlm.nih.gov/pubmed/17513527
http://www.ncbi.nlm.nih.gov/pubmed/16369173

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.

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
```

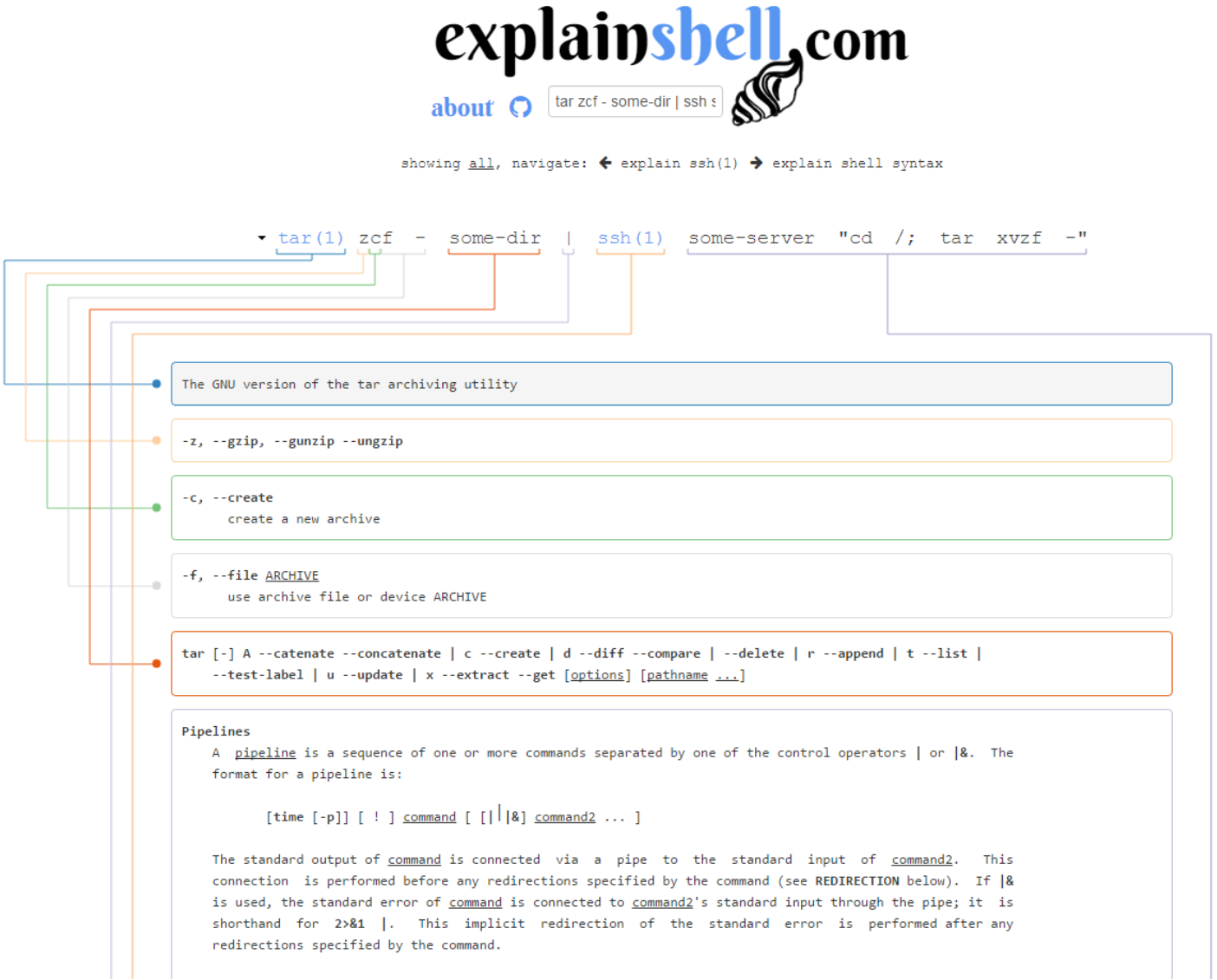
Count the number of alignments (reads mapping to multiple locations counted multiple times)

Collection of published “guides” for bioinformaticians

http://www.opiniomics.org/collection-of-published-guides-for-bioinformaticians/

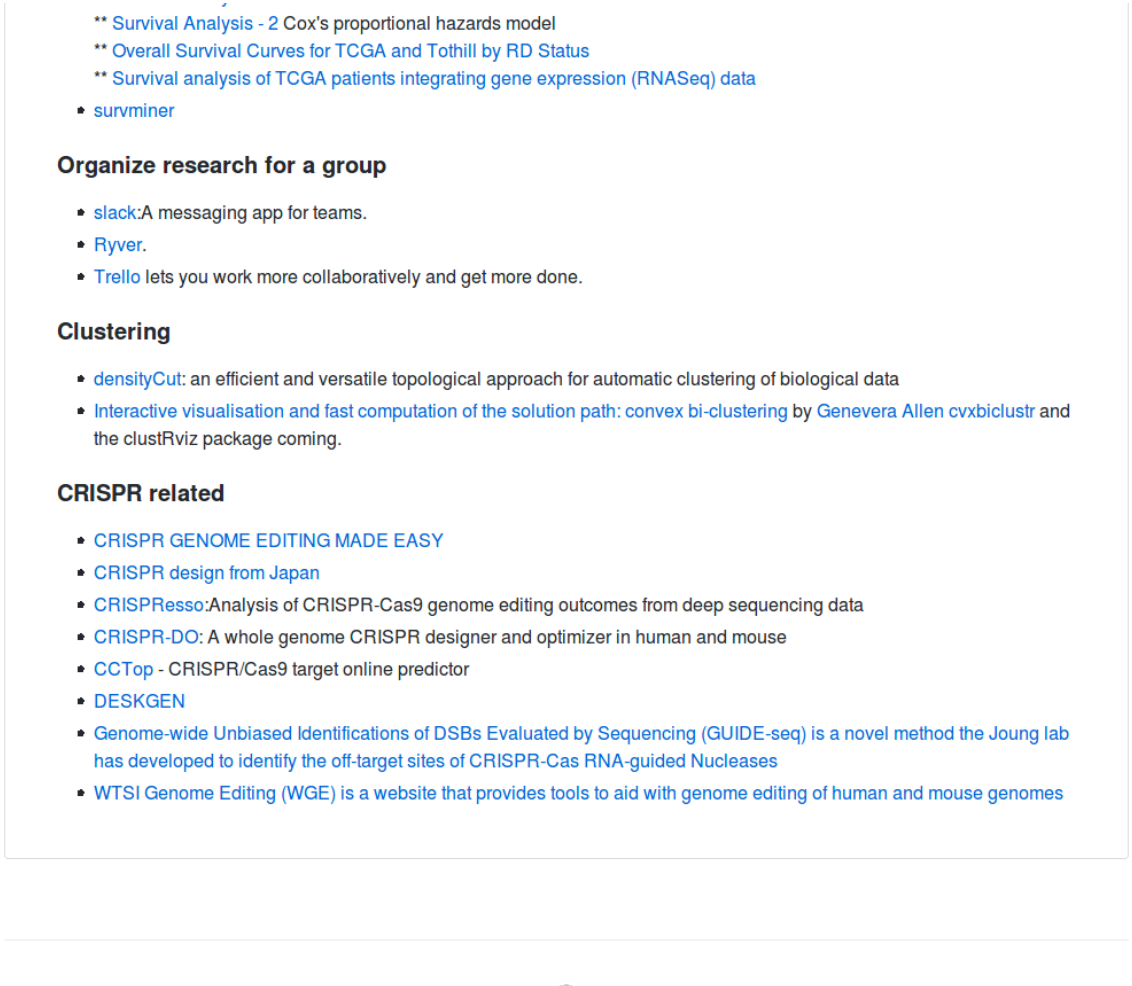
1. Loman N and Watson M (2013) So you want to be a computational biologist? *Nature Biotech* **31(11)**:996-998 [link]
2. Corpas M, Fatumo S, Schneider R. (2012) How not to be a bioinformatician. *Source Code Biol Med.* **7(1)**:3. [link]
3. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy 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 for the work / assignments

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)

Appendix

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/

GFF file format

GFF3 – Generic Feature Format
http://www.sequenceontology.org/gff3.shtml

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

Used for describing

- genes
- features of DNA
- protein sequences
- …

GFF columns

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

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

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

GTF

General Transfer Format (GTF)

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

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

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

Interpretation of variants (technical)

Variants

- Check strand-bias
- Check coverage
- Homopolymer region

Analysis system

- Be careful with stringent default filtering settings
- Know your analysis system (avoid black-boxes)
- Ability to use own databases

Sources of error

- Contaminations through barcodes
- PCR amplification
- FP through sampling (e.g.: skin tissue when taking blood)

-> Clinical interpretation