AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

Bioinformatics and Genome Analyses Course

Lecture 4 (Nov 15)

Stephan Pabinger
stephan.pabinger@ait.ac.at

http://pabinger.site44.com

Recap

Variant calling

- VCF format
- GATK
- Samtools

Structural variant calling

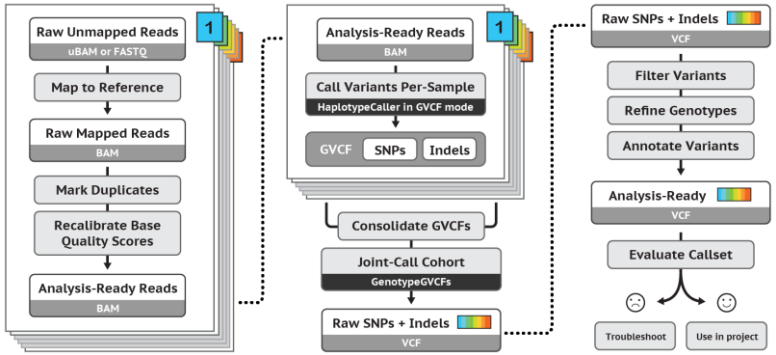
- Different methods and types

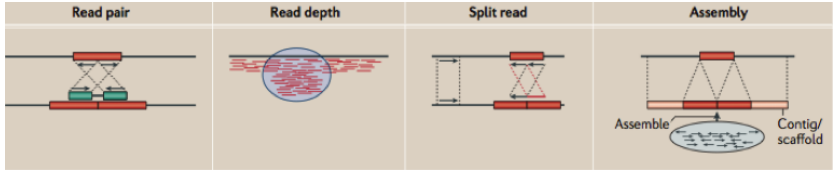
CNV calling in liquid biopsies

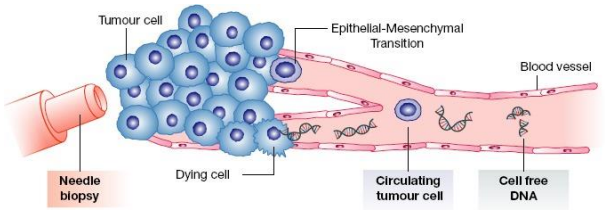
- Basic workflow

SV calling with long reads

- Sniffles







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

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 ....

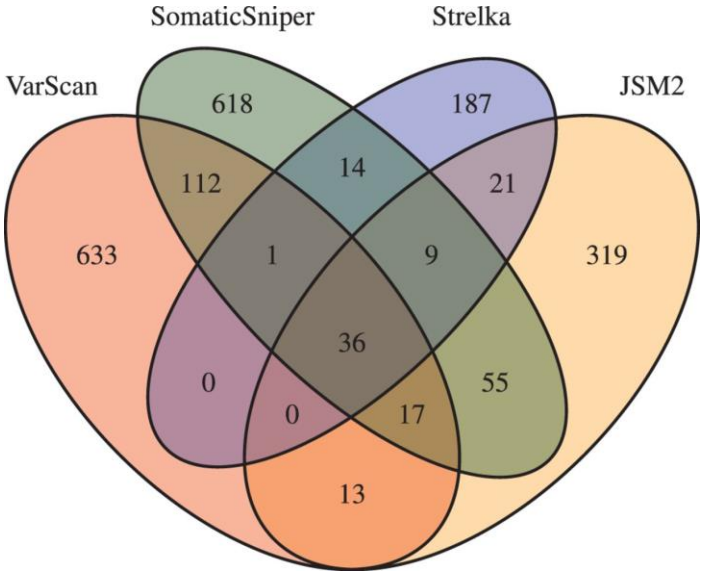
https://www.biostars.org/p/19104/

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

Roberts et al., Bioinformatics (2013)

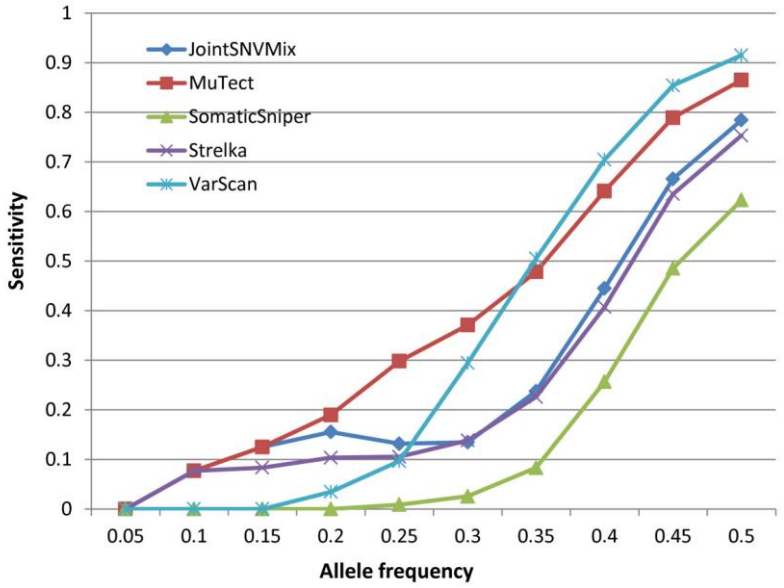
VarScan, SomaticSniper, JSM2 and Strelka **revealed substantial differences** as to the number and character of sites returned, the somatic probability scores assigned to the same sites, their susceptibility to various sources of noise, and their sensitivities to low-allelic-fraction candidates



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

Wang et al., Genome Medicine (2013)

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



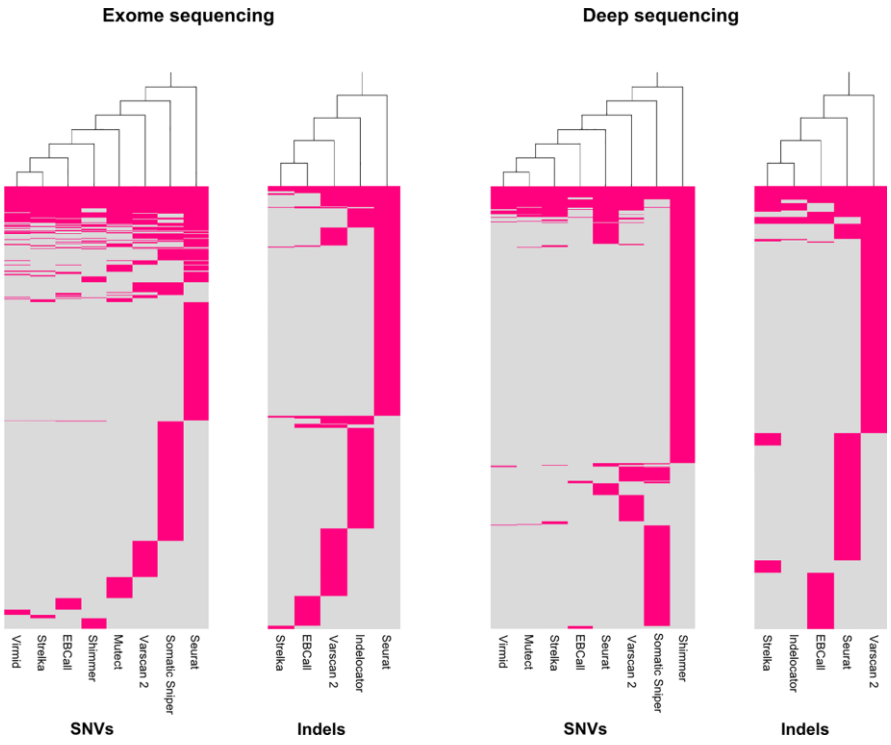
Wang et al. **Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers**
Genome Medicine (2013), **5**:91

sSNV – somatic SNV

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



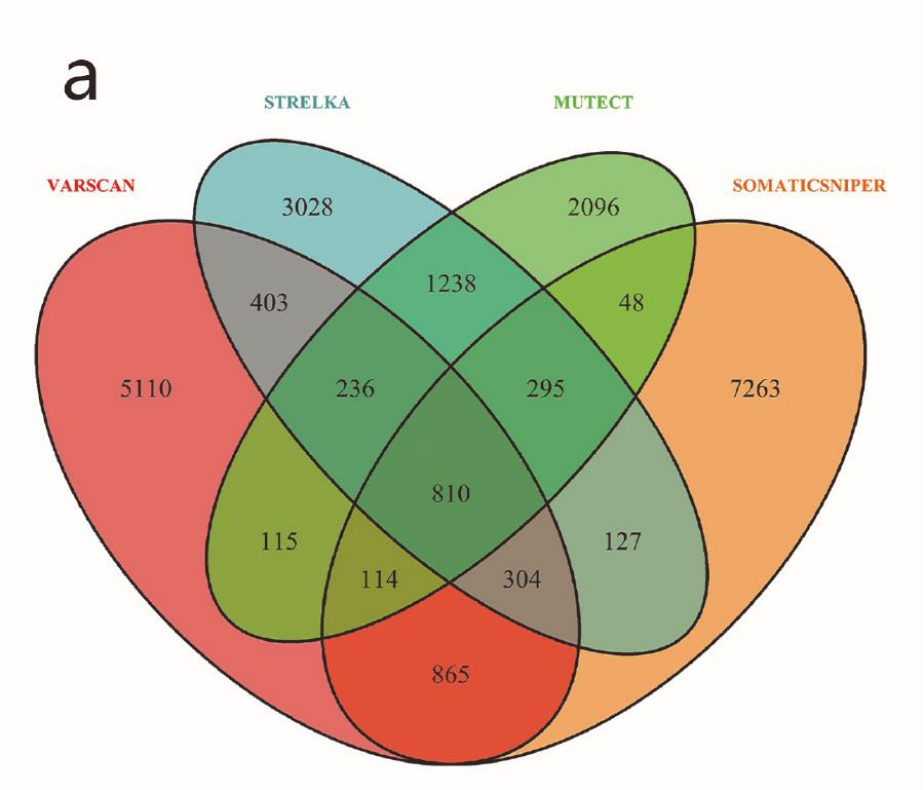
Each red line
represents a called somatic mutation

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

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 individuals)
- 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

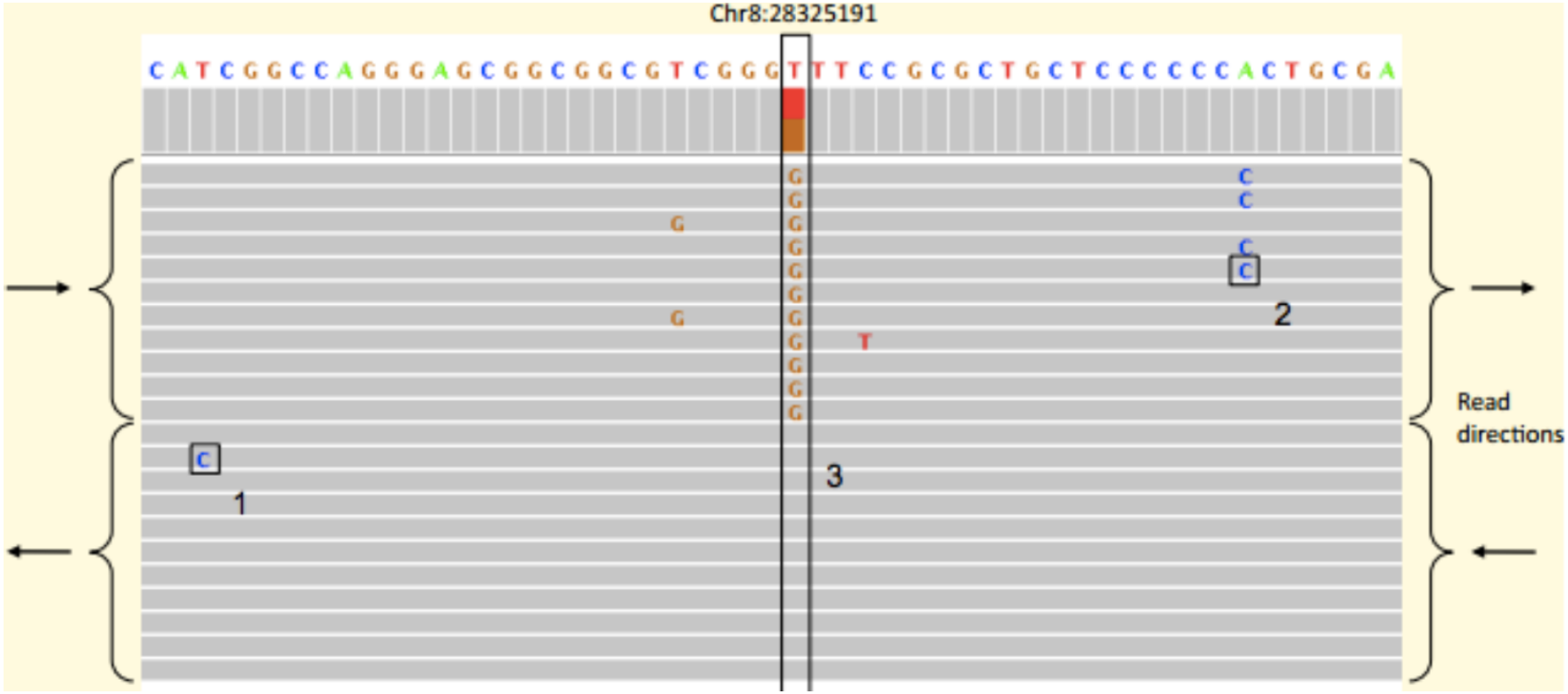
https://www.nature.com/articles/s41588-018-0257-y

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

Beware of Systematic Errors



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

Variant filtering

The biggest problem is large numbers of FPs and FNs:

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

FPs and FNs, may result in:

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

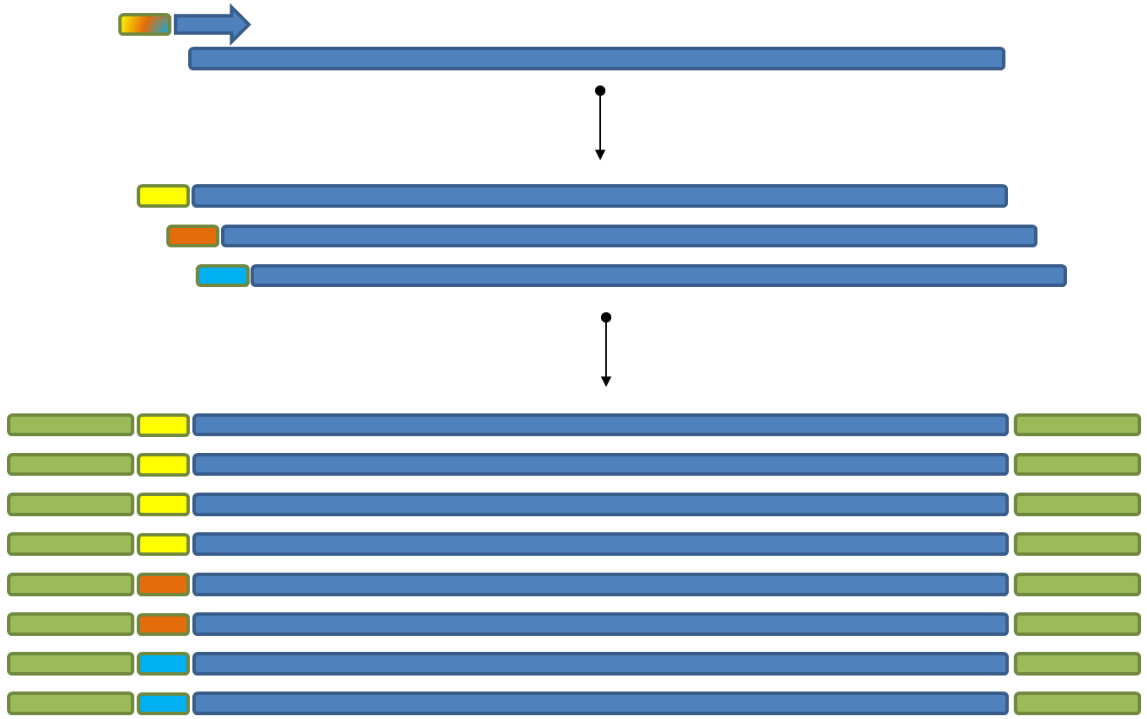
→ **Filter**

Molecular barcoding

Library generation

STEP 1 - Barcoding

STEP 2 - Amplification



Consensus sequence generation

random barcode mix unique barcodes sequencing adaptors

Adapted from http://www.imgm.com

Variant filtering – how to

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

But - there will also be FP with high QUAL

Signs of suspicious variants

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

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

Variant filtering – (some) tools

AUSTRIAN INSTITUTE OF TECHNOLOGY

vcfutils (with samtools)

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

VCFtools

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

GATK pipeline

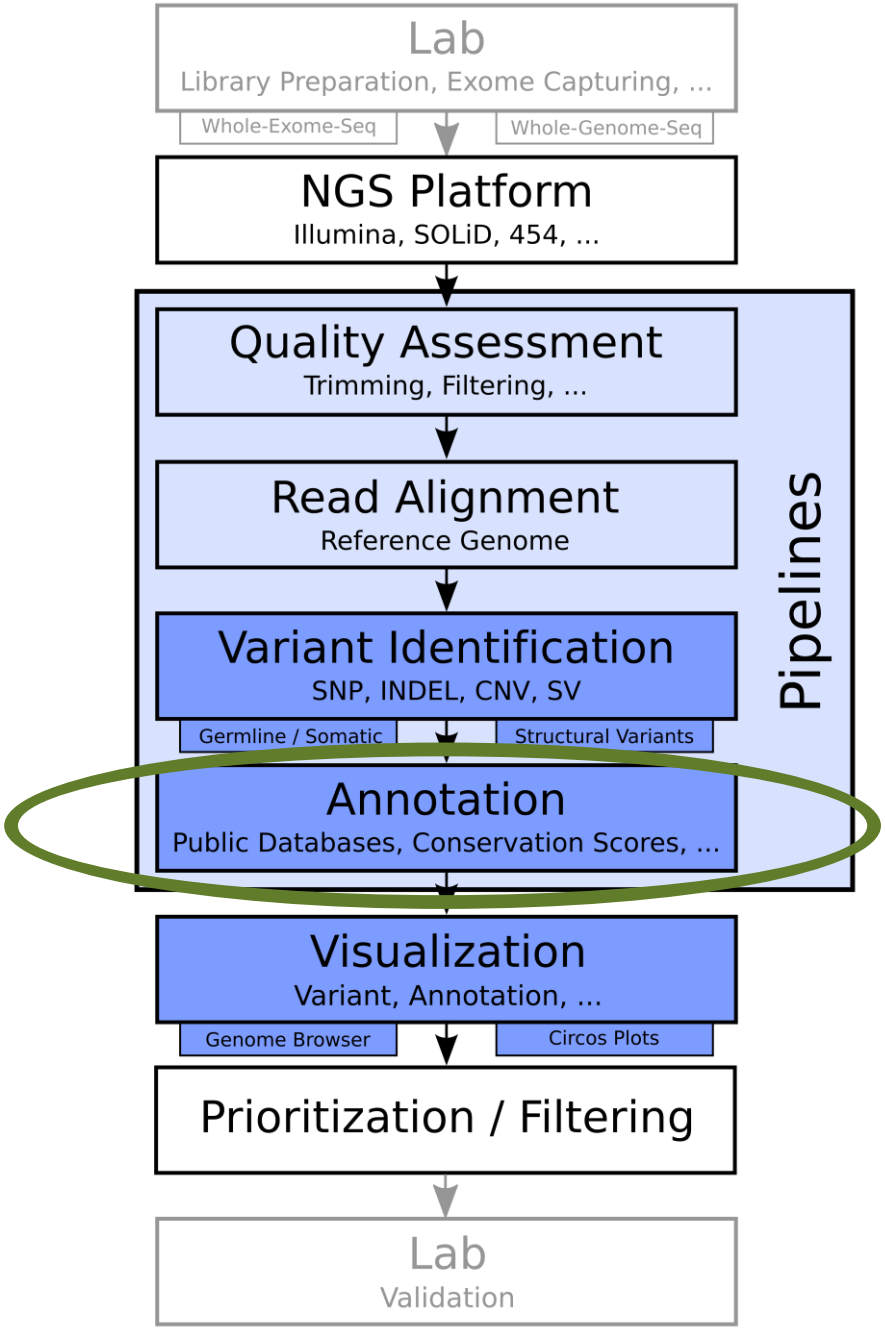
```
java -jar GenomeAnalysisTK.jar -T VariantFiltration -R
<reference.fa> -V <file.vcf>  --filterExpression "QD < 2.0 || MQ <
40.0 || MappingQualityRankSum < -12.5 --filterName "my_snp_filter"
-o <filtered_snps.vcf>
```

vcffilter (vcflib)

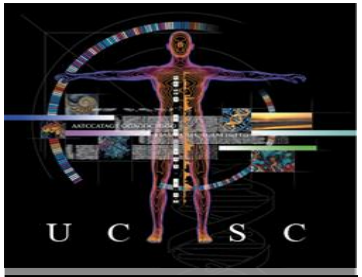
```
vcffilter -f "DP > 10 & MQ > 30 & QD > 20" file.vcf
```

AUSTRIAN INSTITUTE OF TECHNOLOGY

Variant annotation


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

Annotation provides context for interpretation



Conservation
Repeat elements
Genome Gaps
Cytobands
Gene annotations
"Mappability"
DeCIPHER
ISGA

dbSNP
Short Genetic Variations

gnomAD browser

ClinVar

OMIM
Online Mendelian Inheritance in Man

Genetic variation

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

1000 Genomes
A Deep Catalog of Human Genetic Variation

ENCODE

Chromatin marks
DNA methylation
RNA expression
TF binding

Pfam

KEGG
Kyoto Encyclopedia of Genes and Genomes

HGMD®

Human Protein Reference Database

Annotation

Why important?

Molecular Diagnosis

Molecular diagnosis of human disorders is referred to as the **detection of the various pathogenic mutations in DNA and /or RNA samples** in order to facilitate **detection, diagnosis, sub-classification, prognosis, and monitoring response to therapy.**

Aspects

- Information revolution → impacting every aspect of medical practice
- Rate of disease gene discovery is increasing exponentially
  → facilitates understanding diseases at molecular level
- Molecular understanding of disease is translated into diagnostic testing, therapeutics, and eventually preventive therapies

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)

Defining Association

- Variants are not always causal!
  - SNPs sometimes only serve as markers
  - Can play absolutely no role in the disease and even be located on different chromosomes from the gene actually responsible for the phenotype

- Population stratification
  - Variants differ by population
  - Variants **important** markers of disease in **one population** or ethnicity may **not** be effective markers in **another**
  - For GWA studies to be effective predictors in multiple populations, large datasets for each ethnicity must be obtained

Variant annotation

After variant calling → **many** variants

- Synonymous vs. non-synonymous
- Frameshift mutation?
- Impact of variant?

Annotation

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

Variant annotation

Two broad categories of annotations

Annotations depending on gene models

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

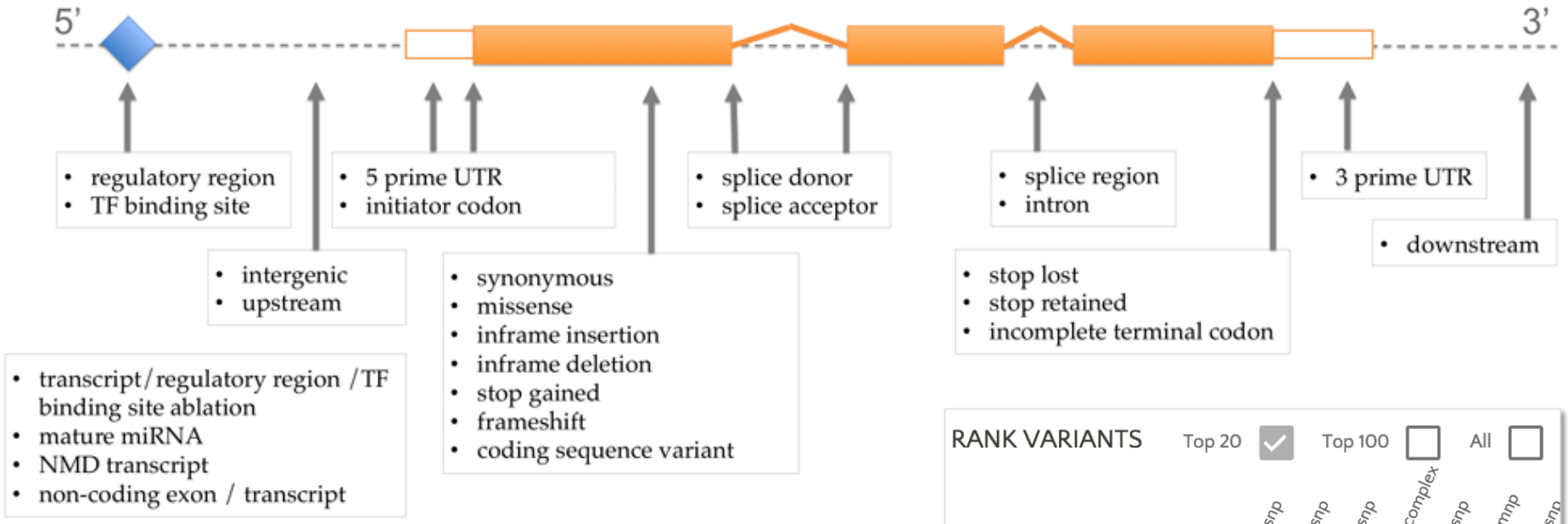
Annotations that do not depend on gene models

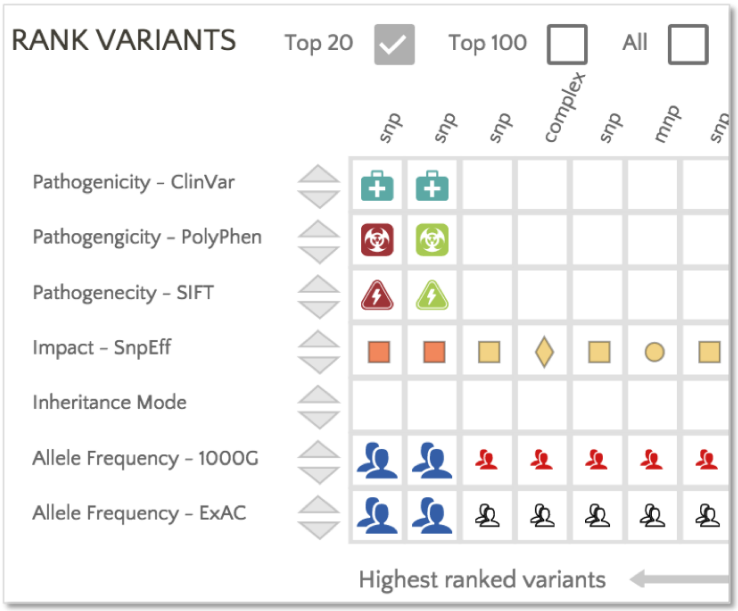
- Variant frequency in different database / different populations
- Degree of conservation across species

Interpretation of Variants

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

Solution: effect prediction





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

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

Make use of phylogenetic conservation

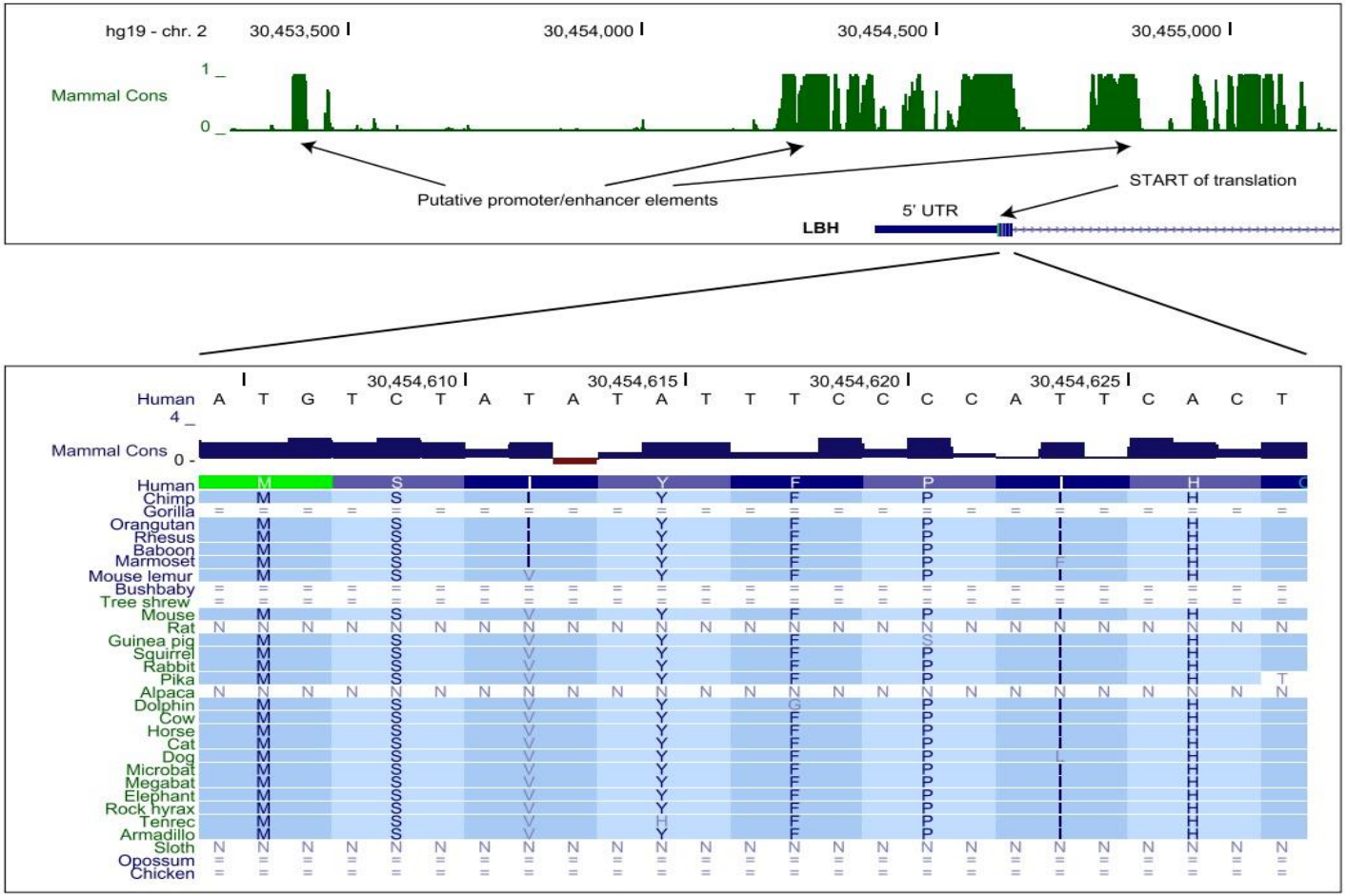


Figure 1. A comparative genomics display derived from the UCSC Genome Browser (Meyer et al. 2013). The *top* panel depicts the genomic region surrounding the 5′ end of the gene *LBH* (limb bud and heart development homolog) in the human genome. The *top* track indicates mammalian conservation as determined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

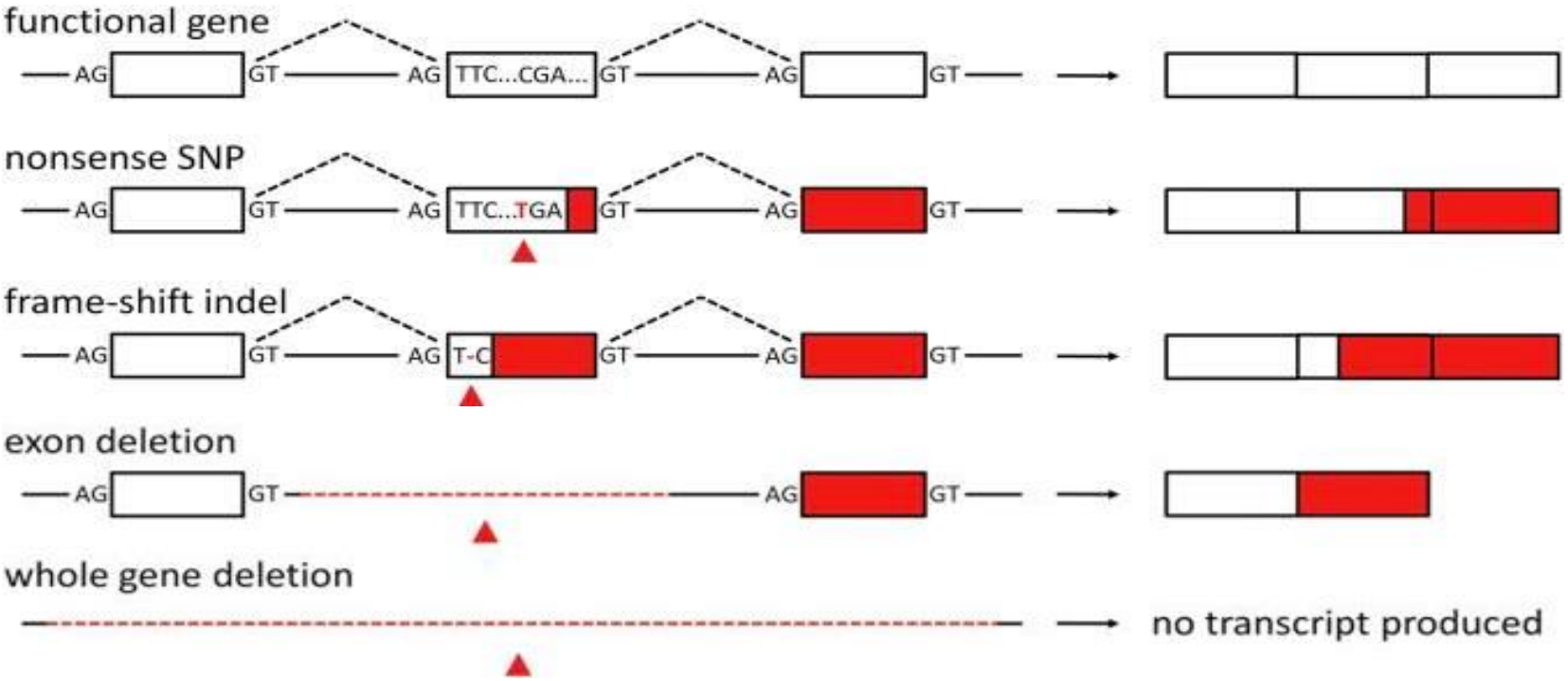
- Conservative amino acid substitution:
  substitutes chemically similar amino acid, less likely to alter function
- Nonconservative amino acid substitution:
  substitutes chemically different amino acid, more likely to alter function
- Consequences for function; often context-specific

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

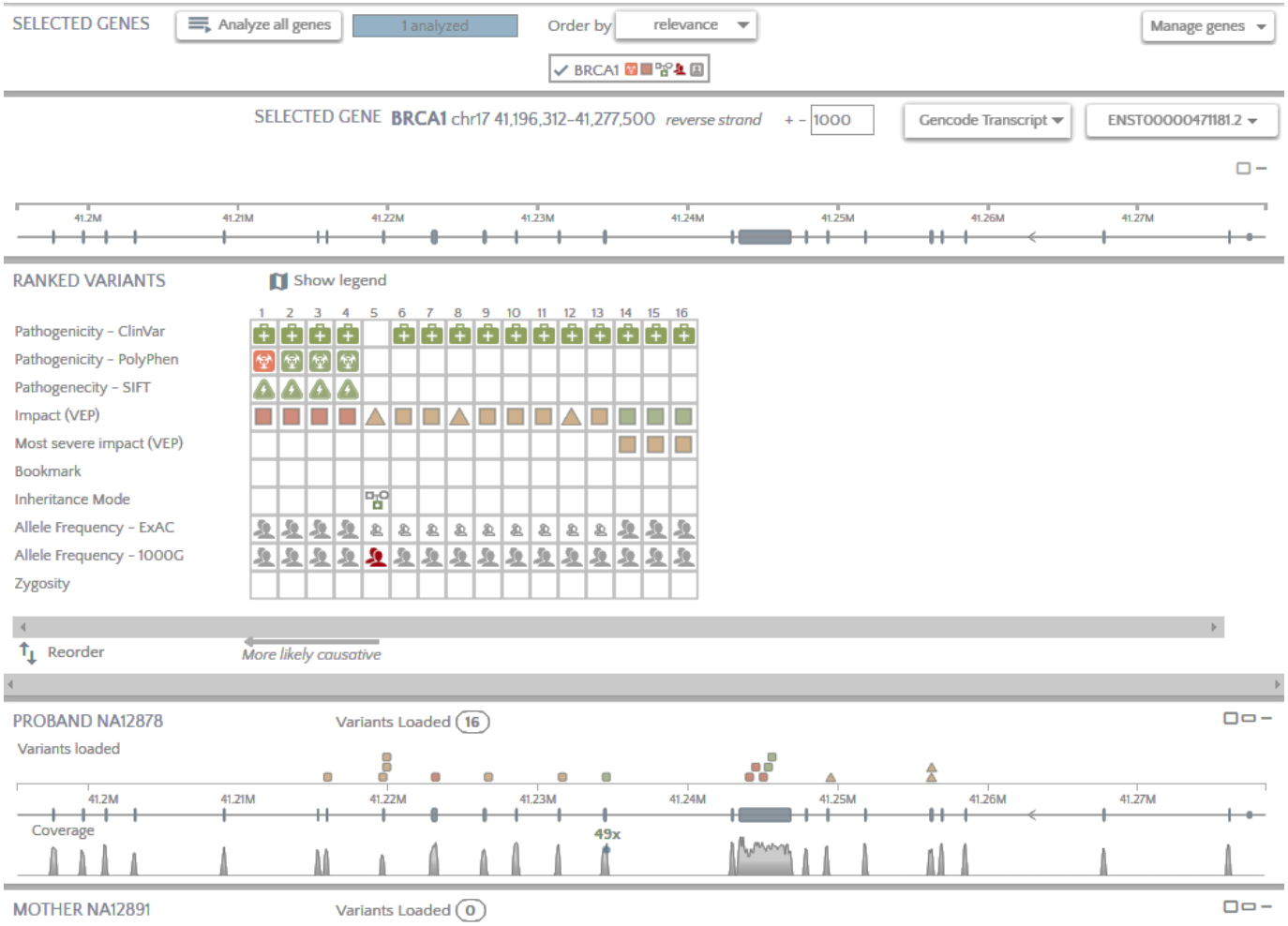
Loss of function variants



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

Problems

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

→ careful when filtering out variants based on dbSNP information

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

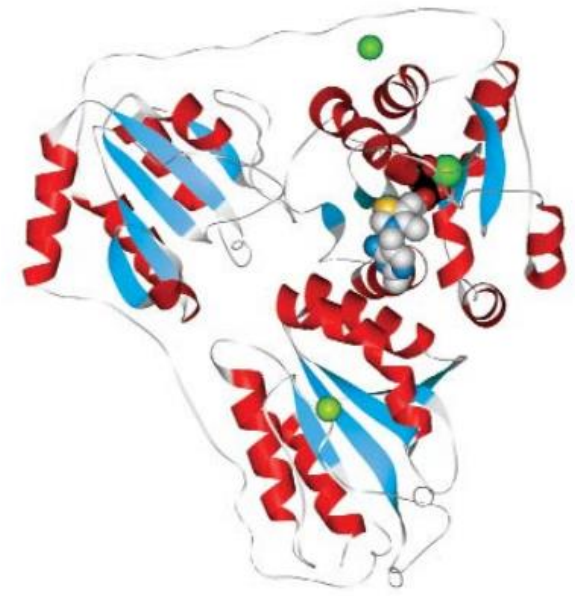
snpEff

- Integrated within Galaxy and GATK.
- SNPs and INDELs

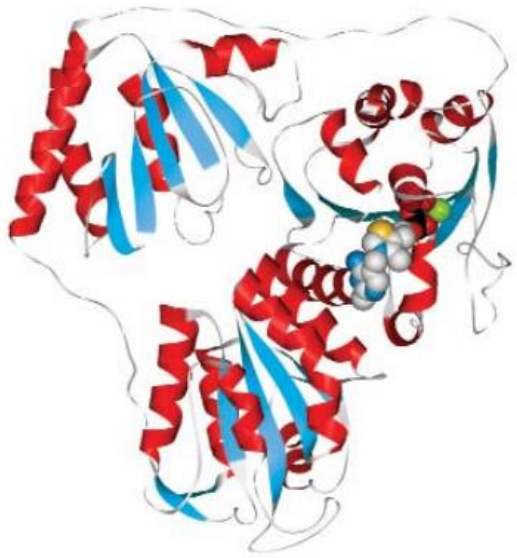
Low seq similarity → same function

Genes separated by hundreds of millions oy years → small sequence identity
→ but remarkable conservation of their structures and functions

BFD



PDC



Benzoylformate decarboxylase (BFD) and pyruvate decarboxylase (PDC) **share a common fold** and overall biochemical function, but they recognize different substrates and have **low (21%) sequence identity**

http://www3.uah.es/farmamol/New_Science_Press/nsp-protein-4.pdf

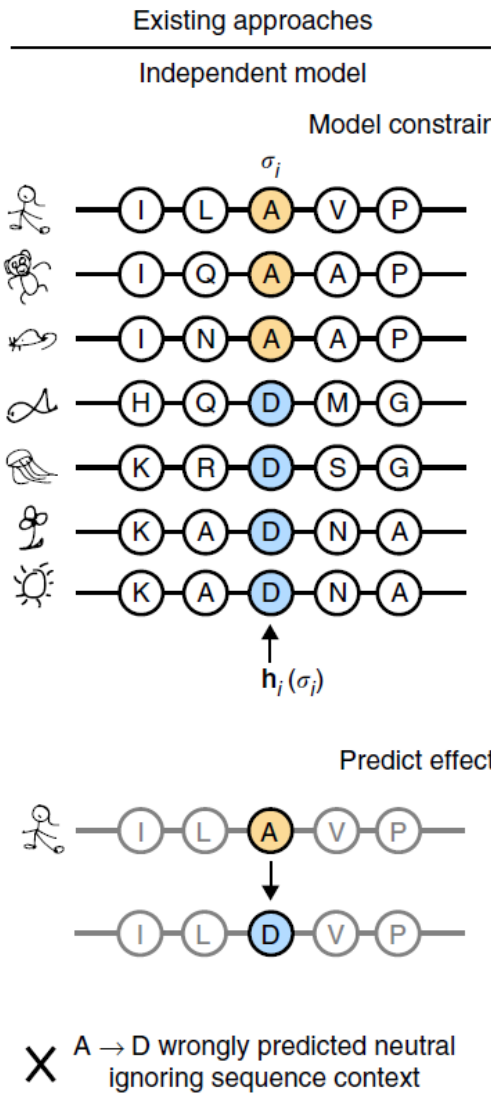
Existing vs new approach

Existing

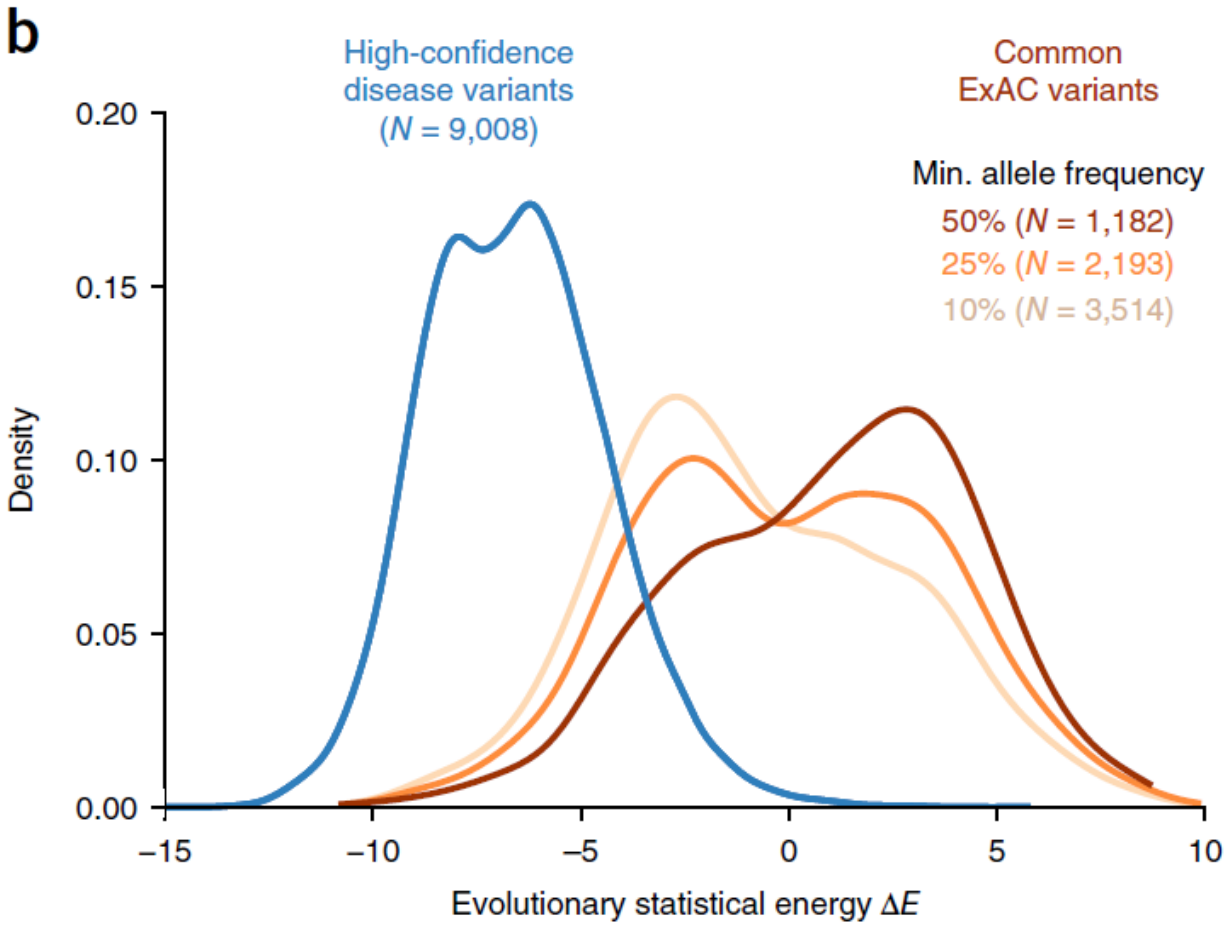
- SIFT, PolyPhen-2, CADD, …
- do not consider epistatis → non-independence of the effects of mutations

EVmutation

- explicitly models interactions between all the pairs of residues in proteins



Distinguish disease variants



Δ*E* distinguish human disease-associated variants from common alleles in the population

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.

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

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

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
- …

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

Visualization

Visualization

Genome browsers - most widely-used tools

Read

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

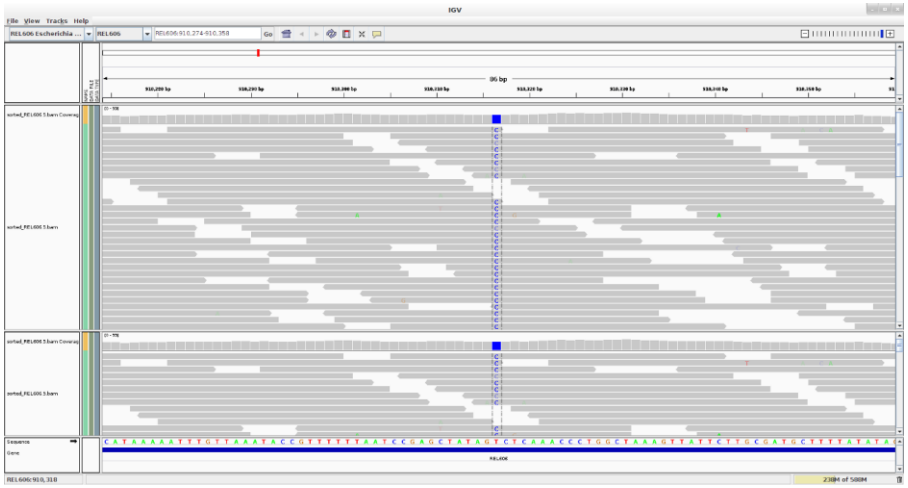
Able to

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

Genome Browsers

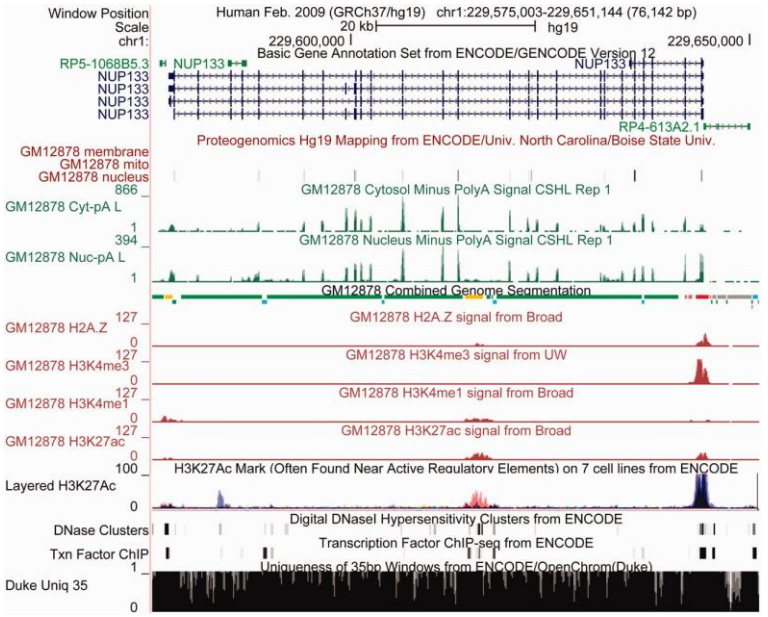
IGV (Integrative Genomics Viewer)

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



UCSC

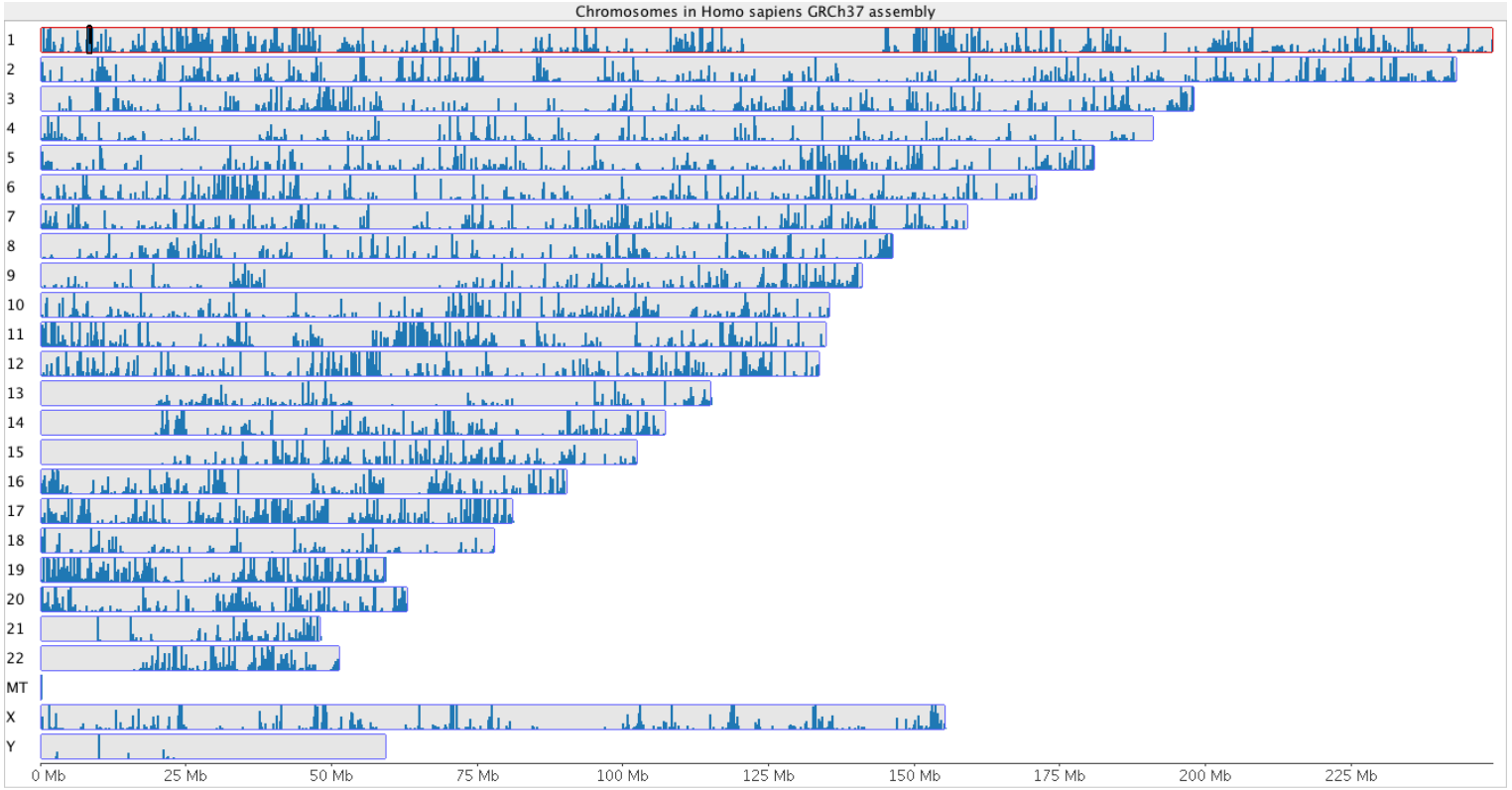
- Web based tool
- Offers many different annotation tracks
- Needs some configuration to display own data



Coverage visualization

Coverage histogram for chromosomes

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

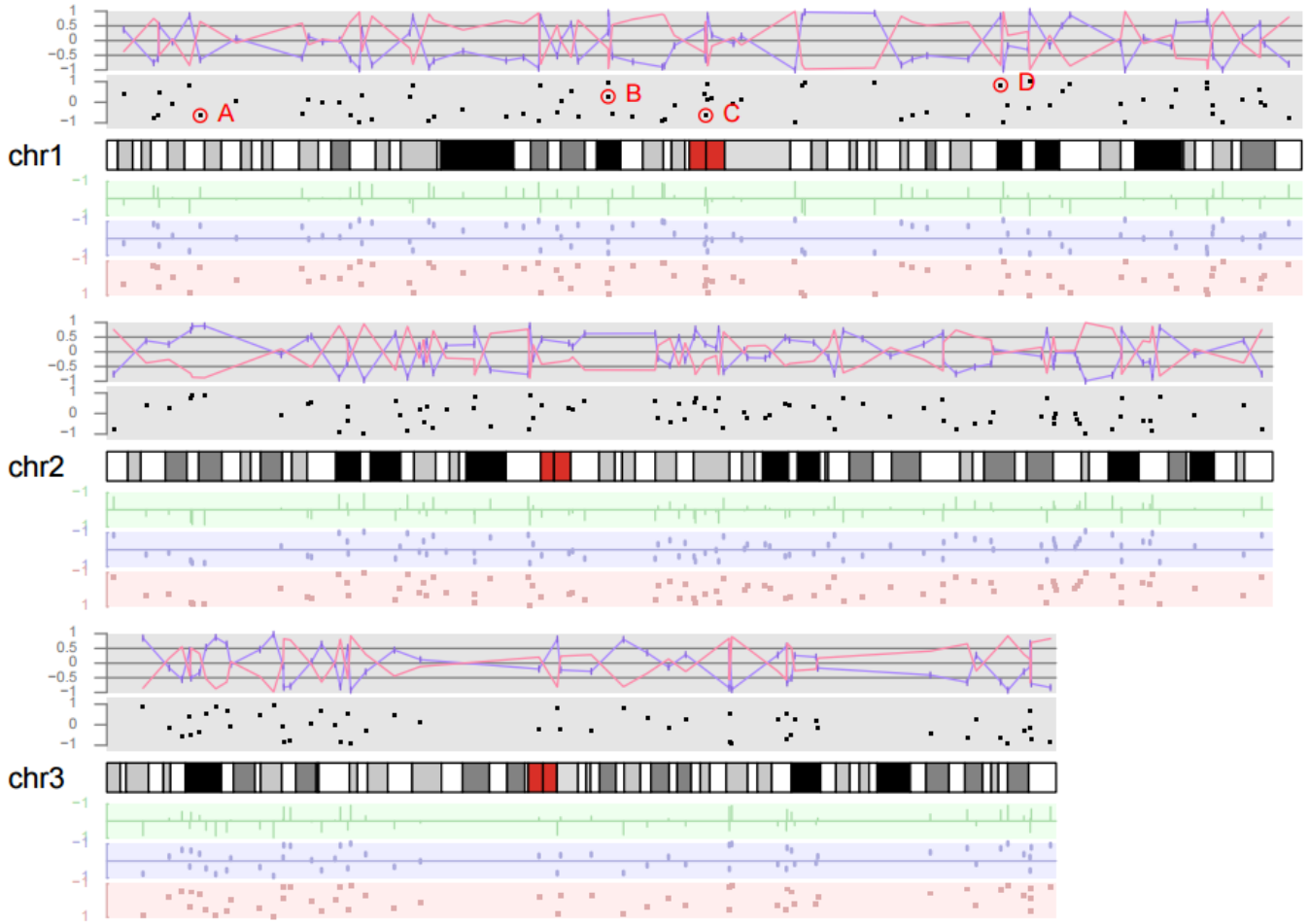


http://seqanswers.com/forums/attachment.php?attachmentid=2118&d=1364889859

Genome-wide visualization

karyoploteR

- Customizable karyotypes with arbitrary data

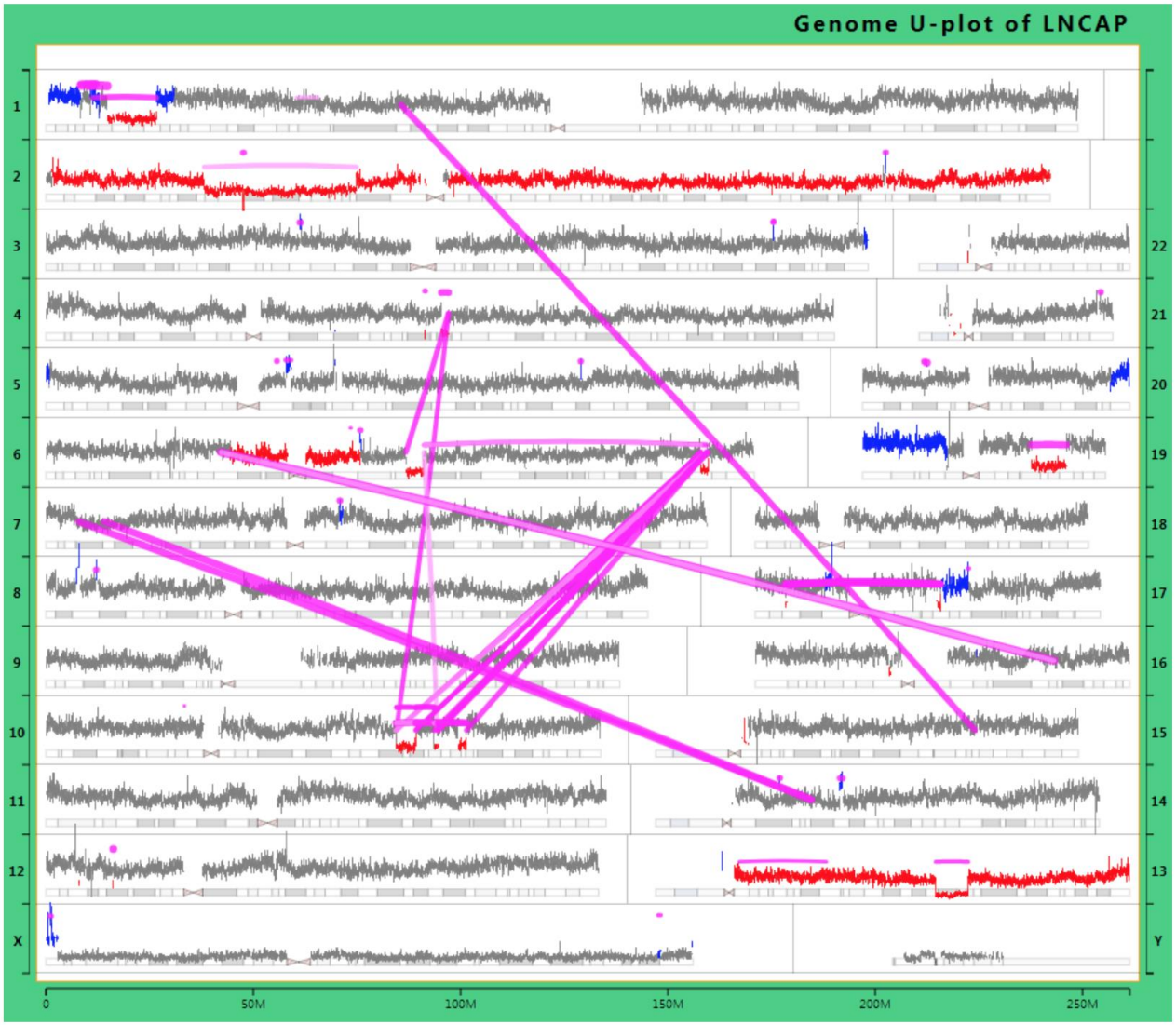


CNV Visulization

GenomeUPlot

Visualize Chromosomal abnormalities

https://github.com/gaitat/GenomeUPlot

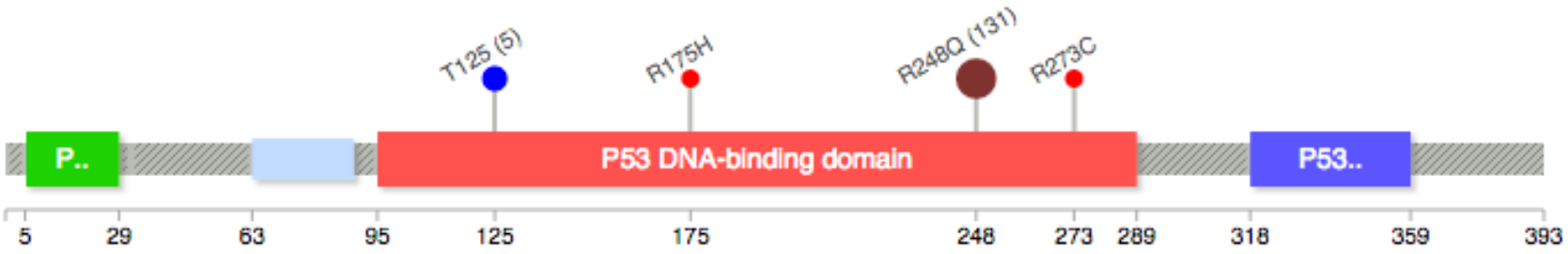


Lollipop

Lollipop-style mutation diagrams for annotating genetic variations

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

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



Analysis pipelines and workflow systems

Pipelines

Bcbio pipeline

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

Crossbow

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

HiPipe

- DNA and RNA analysis
- BWA, GATK

ngs_backbone

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

Pipelines

SeqWare

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

Firehose

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

Nextflow

- Data-driven computational pipelines
- nf-core → a collection of high quality Nextflow pipelines

Galaxy

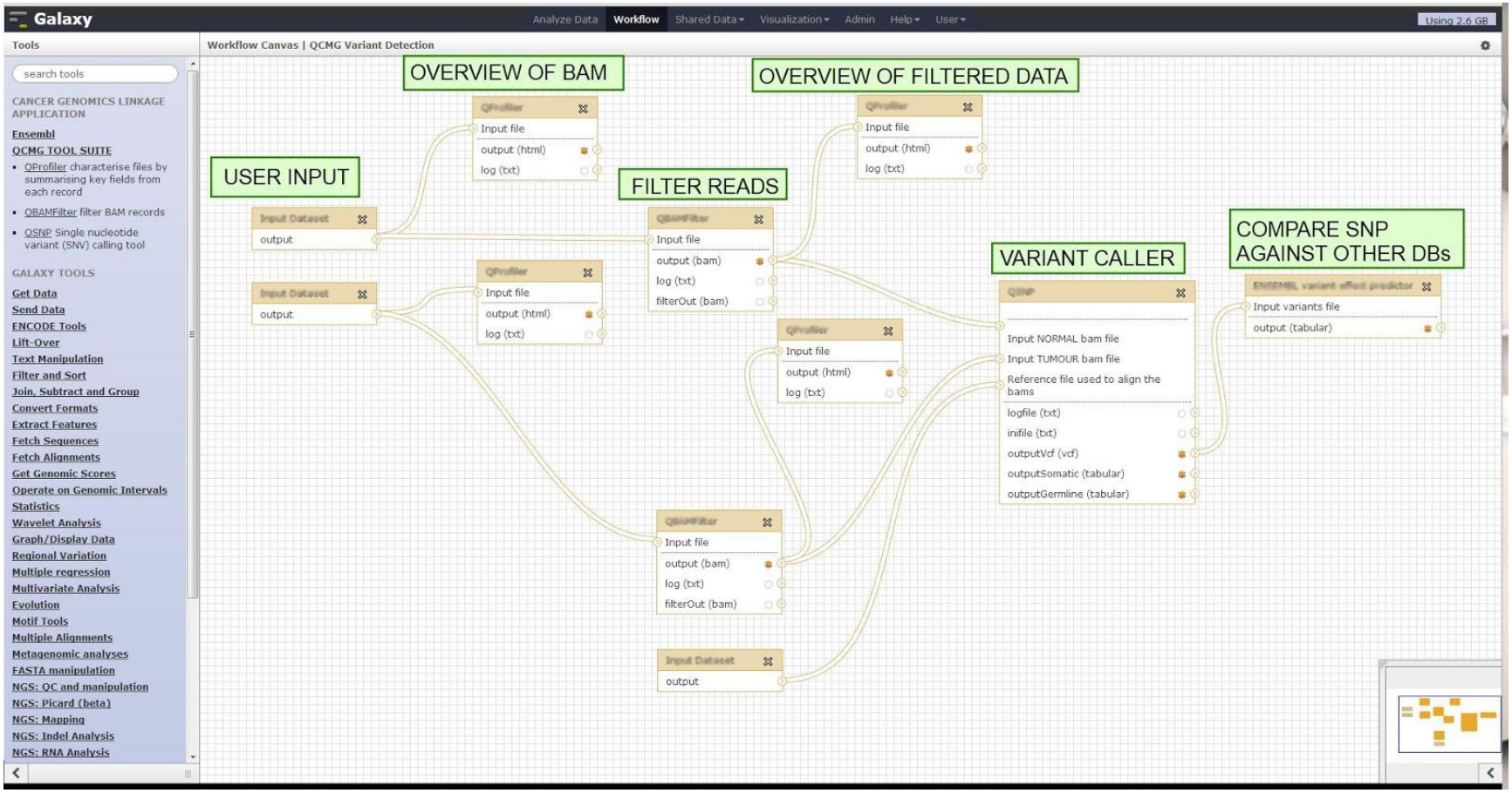
"Galaxy is an open, web-based platform for data intensive biomedical research"

- 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

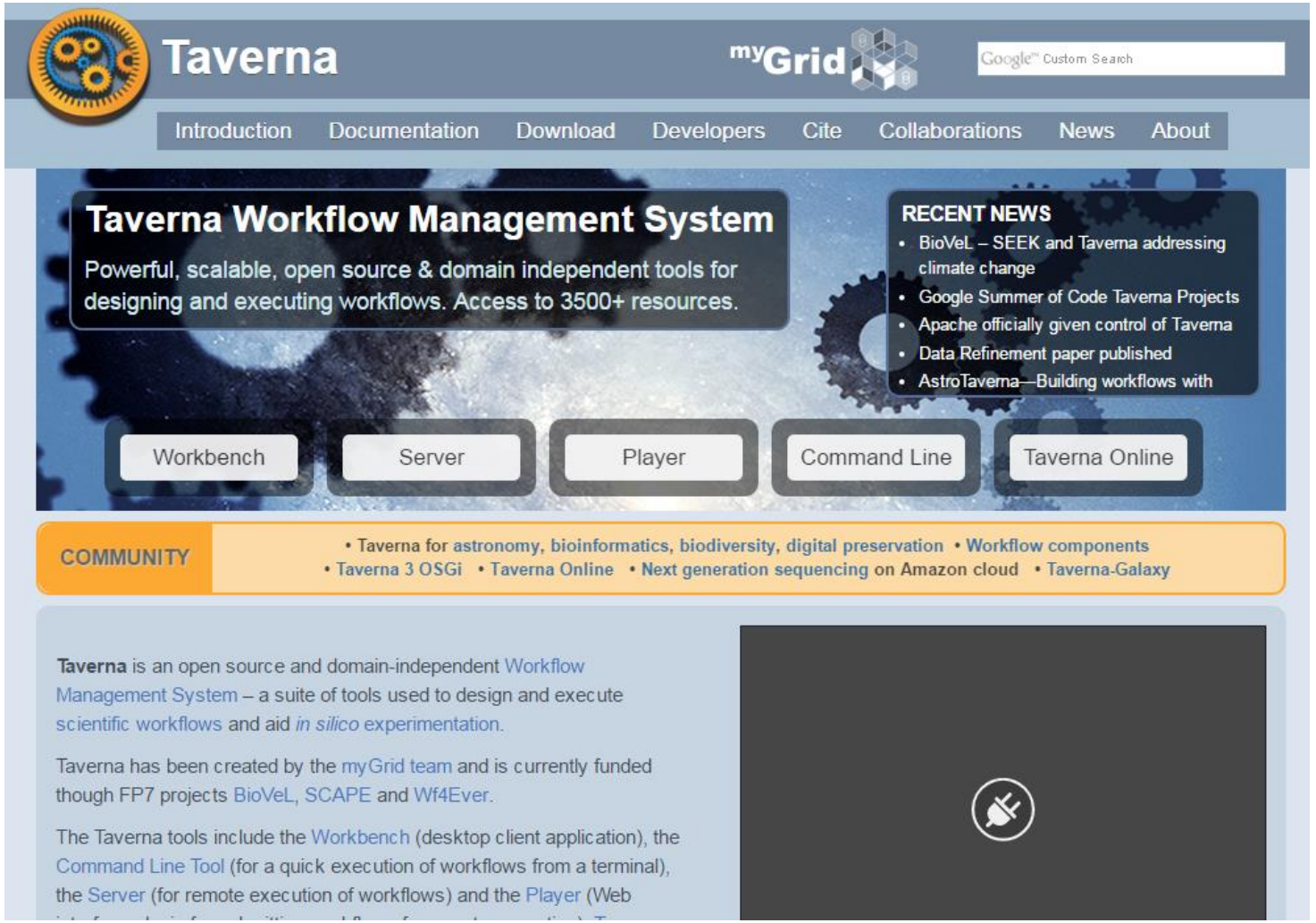
Galaxy



„The Cancer Genomics Linkage Application”

Taverna

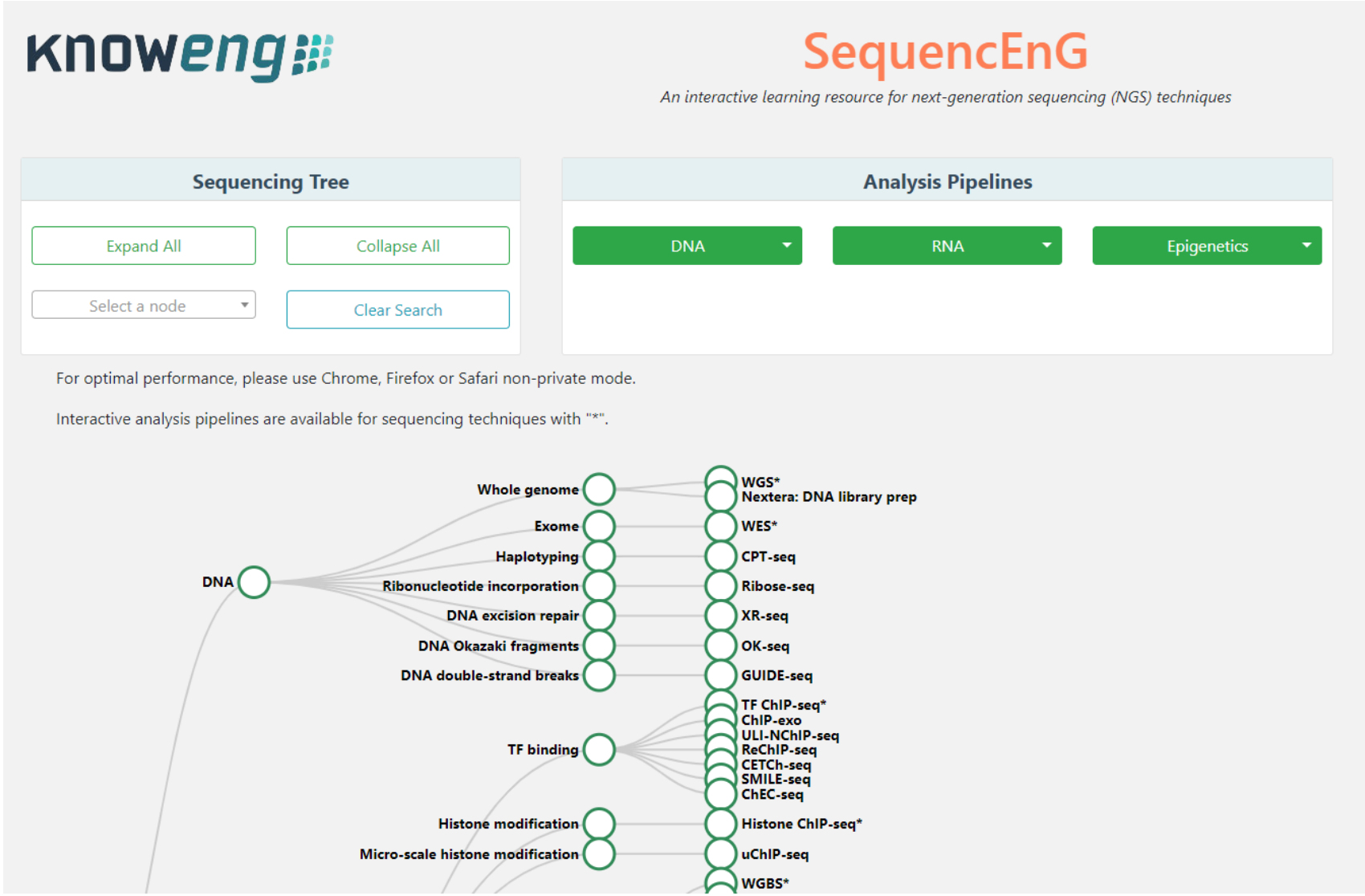
- Workflow management system



Other useful information

SequencEnG

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



Genome in a Bottle (GIAB)

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

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

IGSR: International Genome Sample Resource

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

IGSR and the 1000 Genomes Project

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

Recommendations

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

Where can you get help and information?

Biostar

- A high quality question & answer Web site.

SEQanswers

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

http://omictools.com/

- An informative directory for multi-omic data analysis

Rosalind (http://rosalind.info/)

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

Collection of helps

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

Useful information

List of one liners

https://github.com/stephenturner/oneliners

Basic awk & sed

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

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

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

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

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

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

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

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

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

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

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

SAM and BAM filtering oneliners

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

bamfilter_oneliners.md Raw

SAM and BAM filtering one-liners

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

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

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

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

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

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

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

Count with flagstat for additional information:

```
samtools flagstat filename.bam
```

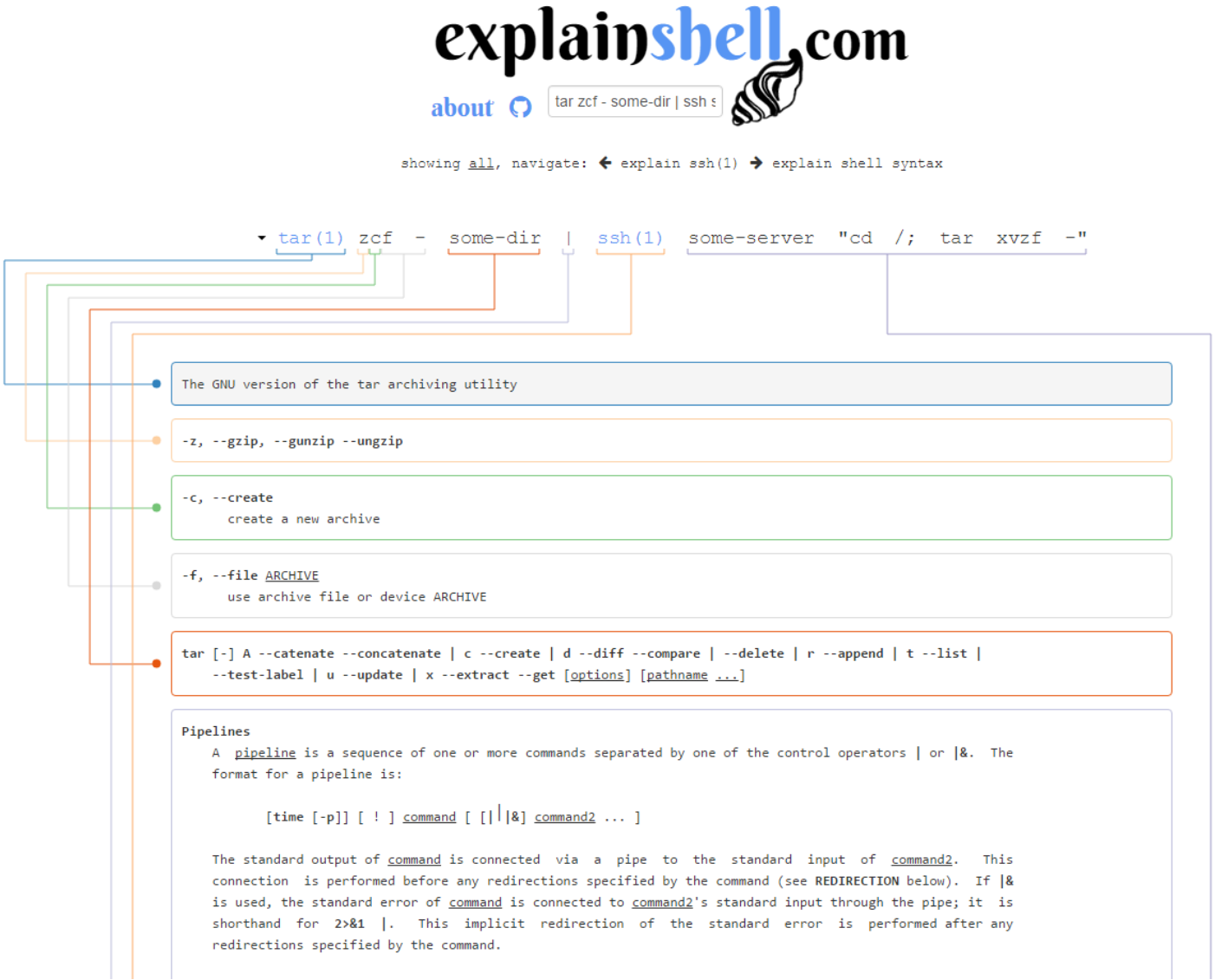
Count the number of alignments (reads mapping to multiple locations counted multiple times)

Collection of published “guides” for bioinformaticians

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

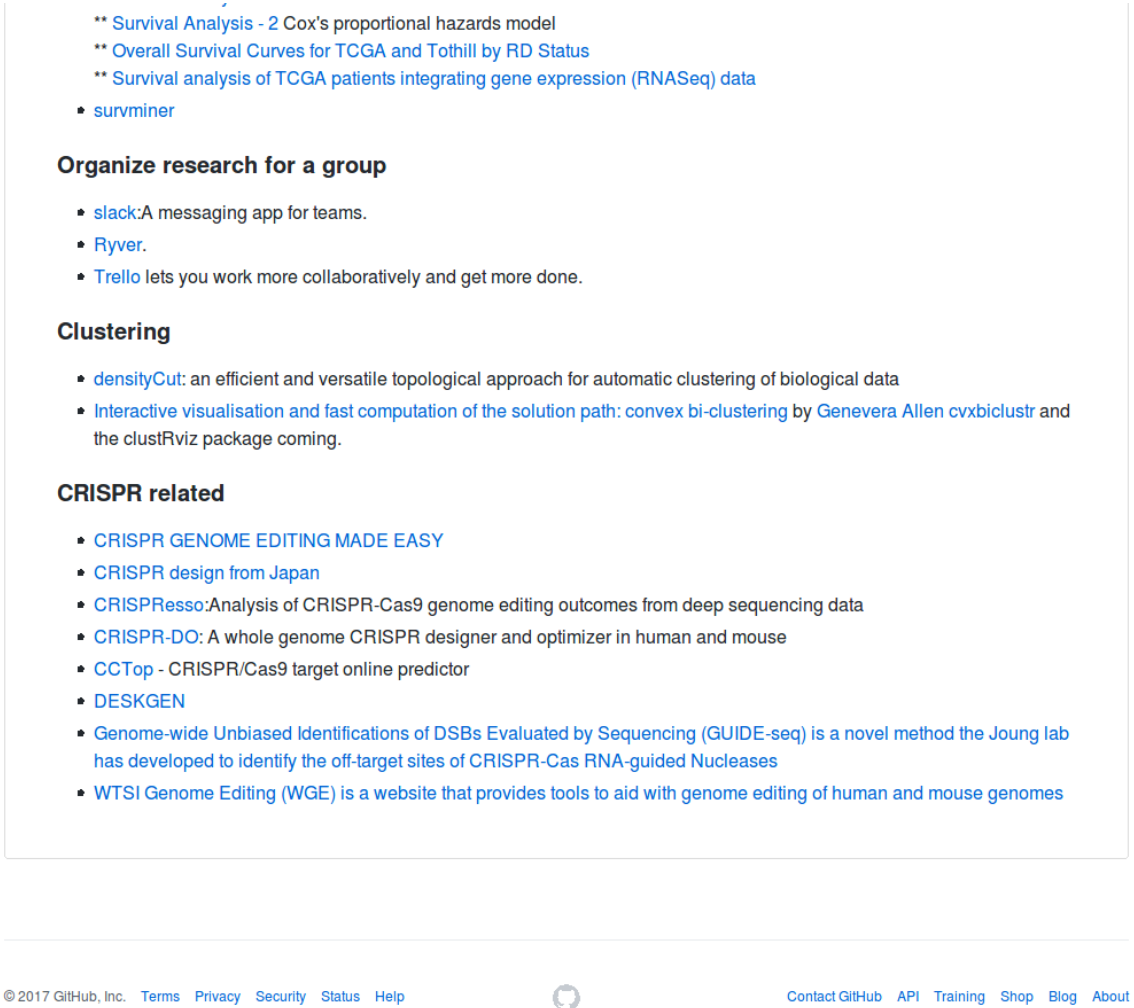
1. Loman N and Watson M (2013) So you want to be a computational biologist? *Nature Biotech* **31(11)**:996-998 [link]
2. Corpas M, Fatumo S, Schneider R. (2012) How not to be a bioinformatician. *Source Code Biol Med.* **7(1)**:3. [link]
3. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy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

Explain Shell - https://explainshell.com/



Huge resource

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



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

- survminer

Organize research for a group

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

Clustering

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

CRISPR related

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

© 2017 GitHub, Inc. Terms Privacy Security Status Help Contact GitHub API Training Shop Blog About

Hints for work

Hints for work

- 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)

Analysis platforms

Illumina basespace -- https://basespace.illumina.com/home/index
Moser, Basílio, Tolios, Ruge

Ion Reporter -- https://ionreporter.lifetechnologies.com/ir/
Tomaselli, Bindeus, Gollobich

Galaxy -- https://main.g2.bx.psu.edu/
Gutsohn, Buchner, Lehrach

Chipster -- http://chipster.csc.fi/
Brauneis, Juno, Beganovic, Majewski

GenePattern -- https://genepattern.broadinstitute.org/gp
Schwarz, Lang, Spielvogel

Practicals – Day 4