Control-free Tumour Analysis with Galaxy

Saskia Hiltemann (1,2), Hailiang Mei (3) , Mattias de Hollander (4), Peter van der Spek (2), Guido Jenster (1), Andrew Stubbs (2)

1)Dept. of Urology, Josephine Nefkens Institute, Erasmus MC, Rotterdam, The Netherlands; 2) Dept. of Bioinformatics, Erasmus MC, Rotterdam, The Netherlands
3) BioAssist, Netherlands Bioinformatics Center (NBIC), Nijmegen, The Netherlands; 4) Netherlands Institute for Ecology, Wageningen, The Netherlands.

The CTMM TraIT Project

TraIT will develop a long-lasting IT infrastructure for translational medicine that will facilitate the collection, storage, analysis, archiving, sharing and securing of the data generated in the CTMM operational translational research projects.
The project builds on existing expertise to create an IT infrastructure that will help to accelerate the translational research in the Dutch Life Sciences and Health sector.

Galaxy Analysis Tools

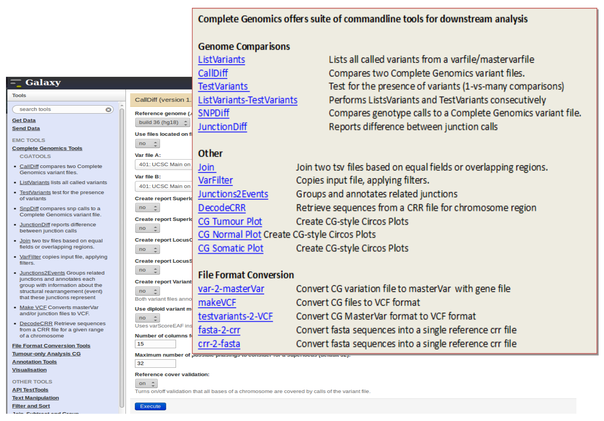
galaxy.nbic.nl

Variant Detection

CGATools

http://cgatools.sourceforge.net/
Complete Genomics Analysis Tools for downstream analysis of Complete Genomics data.

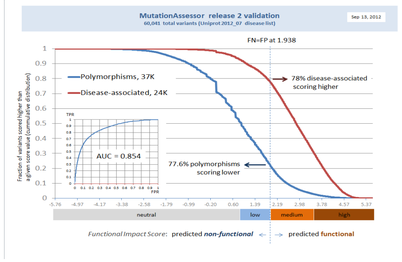
- Genome Comparison Tools
- Variant Filtering Tools
- File Format Conversion Tools
- Circos Plotters



Annotation

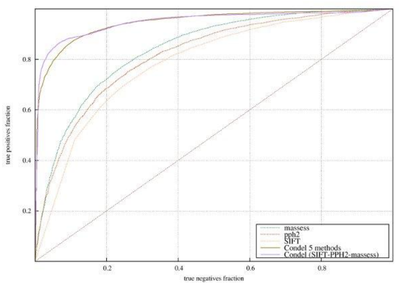
ANNOVAR

www.openbioinformatics.org/annovar/
An efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes.



MutationAssessor

http://mutationassessor.org
Predicts the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms



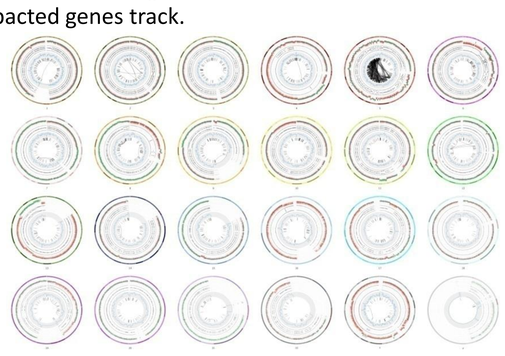
Condel

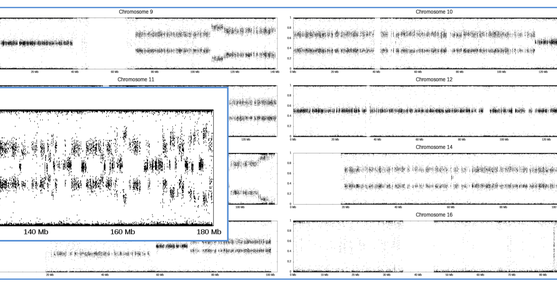
http://bg.upf.edu/condel/home
Consensus **DEL**eteriousness score of non-synonymous single nucleotide variants (SNVs). Integrates the output of computational tools aimed at assessing the impact of non synonymous SNVs on protein function, such as SIFT, PolyPhen2, MutationAssessor

Visualisation

Circos

http://circos.ca
- Plots SVs, SNP density, coverage and B-allele frequency.
- Whole-genome plot, overview of per-genome plots, custom defined regions.
- Impacted genes track.



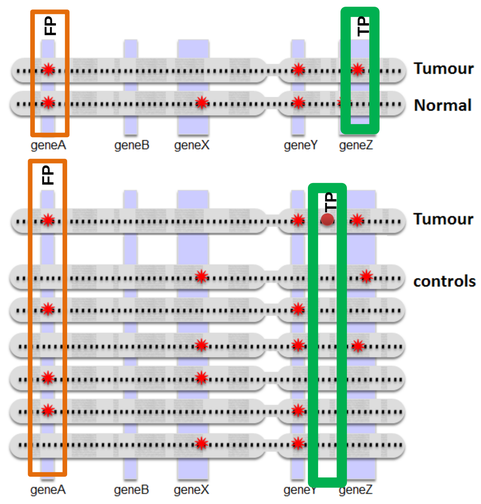


Gnuplot

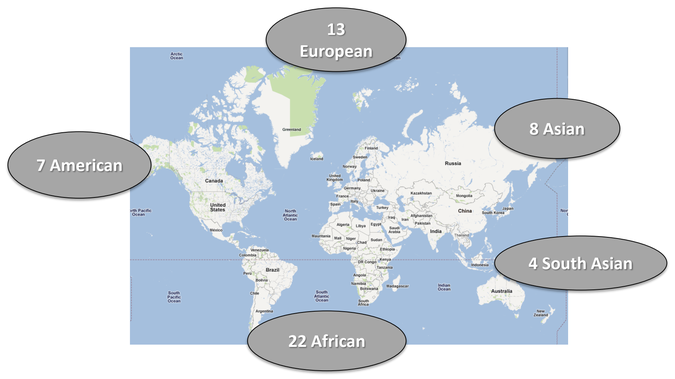
http://www.gnuplot.info
- B-allelefrequency form Complete Genomics masterVar files.
- Generic genomic data plotter `<chr - position - value>`.
- Output: Single-chromosome plots, all chromosomes in single image, custom defined regions.

Control-free Tumour Analysis

- For **optimal somatic variation** detection the tumour genome is compared with a control genome from the same individual
- In the **absence of associated normal** tissue then a set of control genomes are used as a virtual normal sample.
- Limitation of **virtual normal** compute resources, terabytes of reference data and permanent storage
- We have implemented a workflow to detect **tumour only somatic variation** in Galaxy using open source applications
- The analysis has been evaluated for **sensitivity** and **specificity** for whole genome sequencing of publically available tumour-normal pairs.

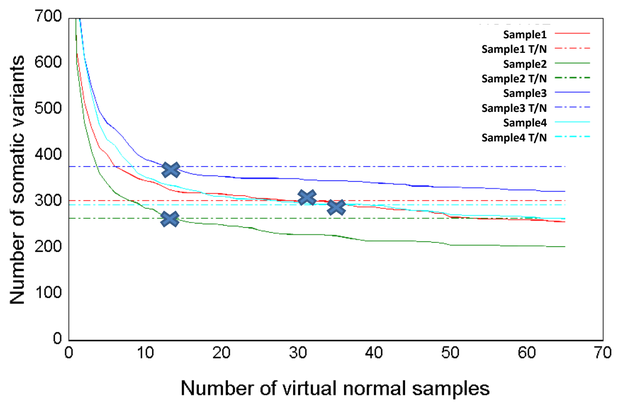


Virtual Normal Set



54 public samples of healthy, unrelated individuals, sequenced by Complete Genomics

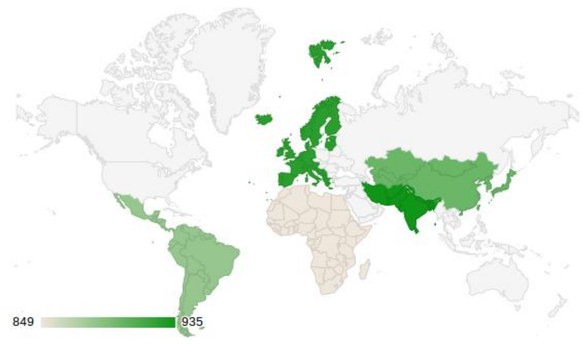
Results



Using 10-40 *Virtual Normal* samples, the same amount of structural variants can be filtered as with an associated normal sample.

Increasing the number of *Virtual Normal* samples further allows for the filtering of increasingly rare polymorphisms.

For optimal filtering use set of genetically diverse genomes as a virtual normal



Galaxy Tools

For comparison of SVs and Small Variants to *Virtual Normal* will be available from NBIC Galaxy.

galaxy.nbic.nl

TraIT Galaxy HPC CLOUD Architecture

The rapid evolution of NGS technologies together with decreasing cost are creating a challenge to store and analyze the vast amount of sequencing data that are generated by experimental biologists. Configuring suitable data analysis software and having access to readily available computation and storage are the two major bottlenecks faced by many research groups

Advantages of Cloud Systems
- Rapid elasticity (scale up/down dynamically)
- Full adminstrational rights
- Perfect for project-based research
- Access to powerful compute systems

BiGGrid & SARA HPC Cloud (Calligo)
- 19 x Intel Xeon 32 core processors (608 cores)
- 19 x 256GB RAM (7.75 TB RAM)
- 400TB shared storage