

# 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.

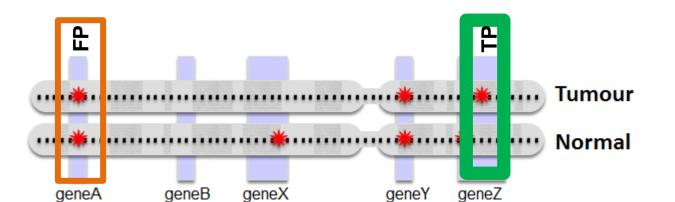
# The CTMM TralT Project

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.



# **Control-free Tumour Analysis**

- For optimal somatic variation detection the tumour genome is compared with a control genome from the same individual



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

lete Genomics offers suite of commandline tools for downstream analysis

### **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

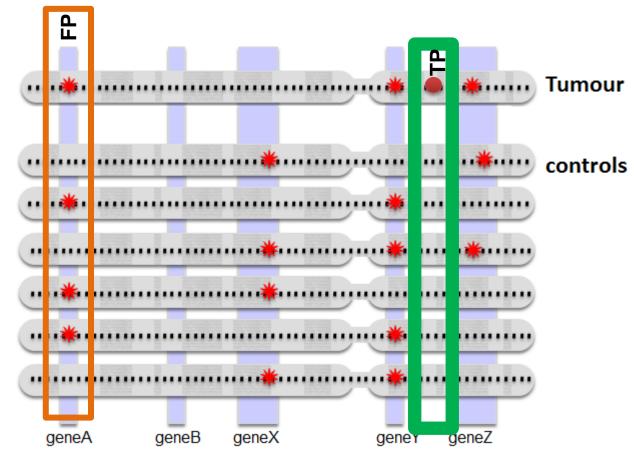
		ListVariants	Lists all called variants from a varfile/mastervarfile
		CallDiff	Compares two Complete Genomics variant files.
Galaxy		TestVariants	Test for the presence of variants (1-vs-many comparisons)
bls	CallDiff (version 1.	ListVariants-TestVariant	
search tools	Reference genome (.	SNPDiff	Compares genotype calls to a Complete Genomics variant f
t Data nd Data	build 36 (hg18) 🛟	JunctionDiff	Reports difference between junction calls
IC TOOLS	Use files located on fi		
mplete Genomics Tools	no 🌲	Other	
CGATOOLS	Var file A:		
CallDiff compares two Complete Genomics variant files.	401: UCSC Main on	Join	Join two tsv files based on equal fields or overlapping regions.
ListVariants lists all called variants	Var file B:	VarFilter	Copies input file, applying filters.
TestVariants test for the presence of variants	401: UCSC Main on Create report Superk	Junctions2Events	Groups and annotates related junctions
<u>SnpDiff</u> compares snp calls to a Complete Genomics variant file.	no 🛟	DecodeCRR	Retrieve sequences from a CRR file for chromosome region
JunctionDiff reports difference between junction calls	Create report Superio	CG Tumour Plot	Create CG-style Circos Plots
Join two tsv files based on equal	Create report LocusC	CG Normal Plot Create	CG-style Circos Plots
fields or overlapping regions.	no 🌲	CG Somatic Plot	Create CG-style Circos Plots
<u>VarFilter</u> copies input file, applying filters.	Create report LocusS		
Junctions2Events Groups related junctions and annotates each	no 🌲	File Format Conversion	
group with information about the	Create report Variant	var-2-masterVar	Convert CG variation file to masterVar with gene file
structural rearrangement (event) that these junctions represent	Both variant files anno	makeVCF	Convert CG files to VCF format
Make VCF Converts masterVar and/or junction files to VCF.	Use diploid variant m		
DecodeCRR Retrieve sequences	no 🛟 Uses varScoreEAF ins	testvariants-2-VCF	Convert CG MasterVar format to VCF format
from a CRR file for a given range of a chromosome	Number of columns fe	fasta-2-crr	Convert fasta sequences into a single reference crr file
e Format Conversion Tools	15	crr-2-fasta	Convert fasta sequences into a single reference crr file
mour-only Analysis CG	Maximum number of		· · ·
notation Tools	32		
sualisation	Reference cover valid	lation:	
HER TOOLS	on 🌲		
<u>I TestTools</u> kt Manipulation	Turns on/off validation t	that all bases of a chromosome are cov	ered by calls of the variant file.
ter and Sort	Execute		
in Subtract and Crown			

ienome Comparisor

- 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**

#### **ANNOVAR**

Annotation

#### www.openbioinformatics.org/annovar/

An efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes.

#### **MutationAssessor**

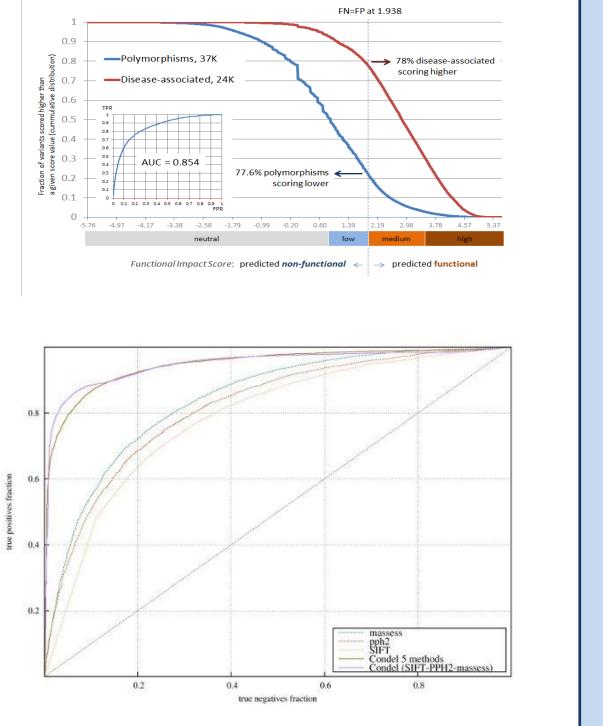
#### http://mutationassessor.org

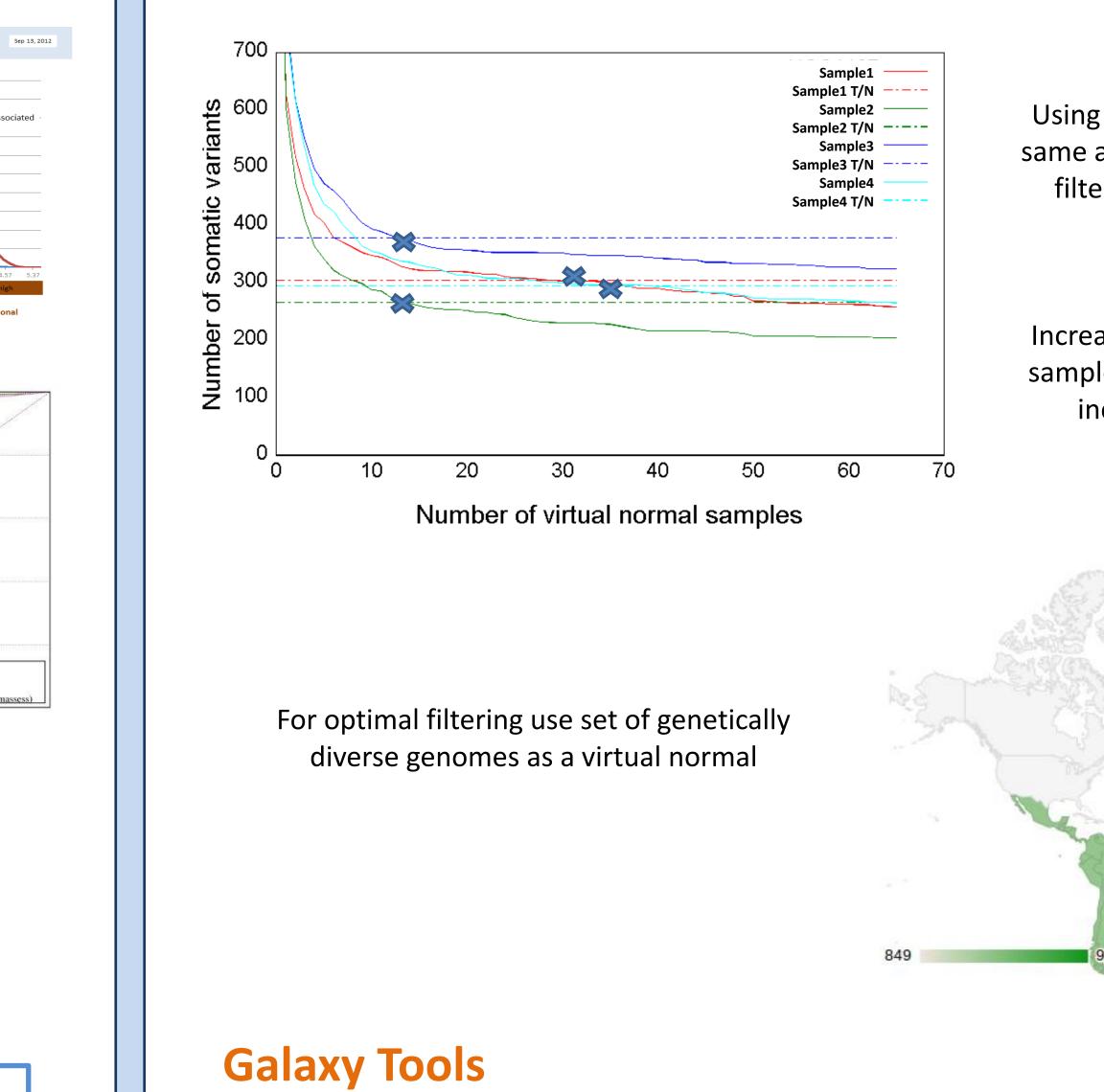
Predicts the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms

### Condel

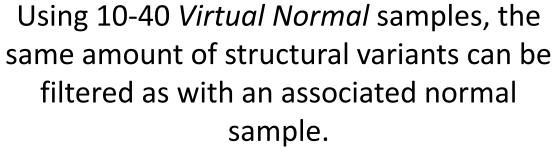
#### http://bg.upf.edu/condel/home

**Con**sensus **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





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



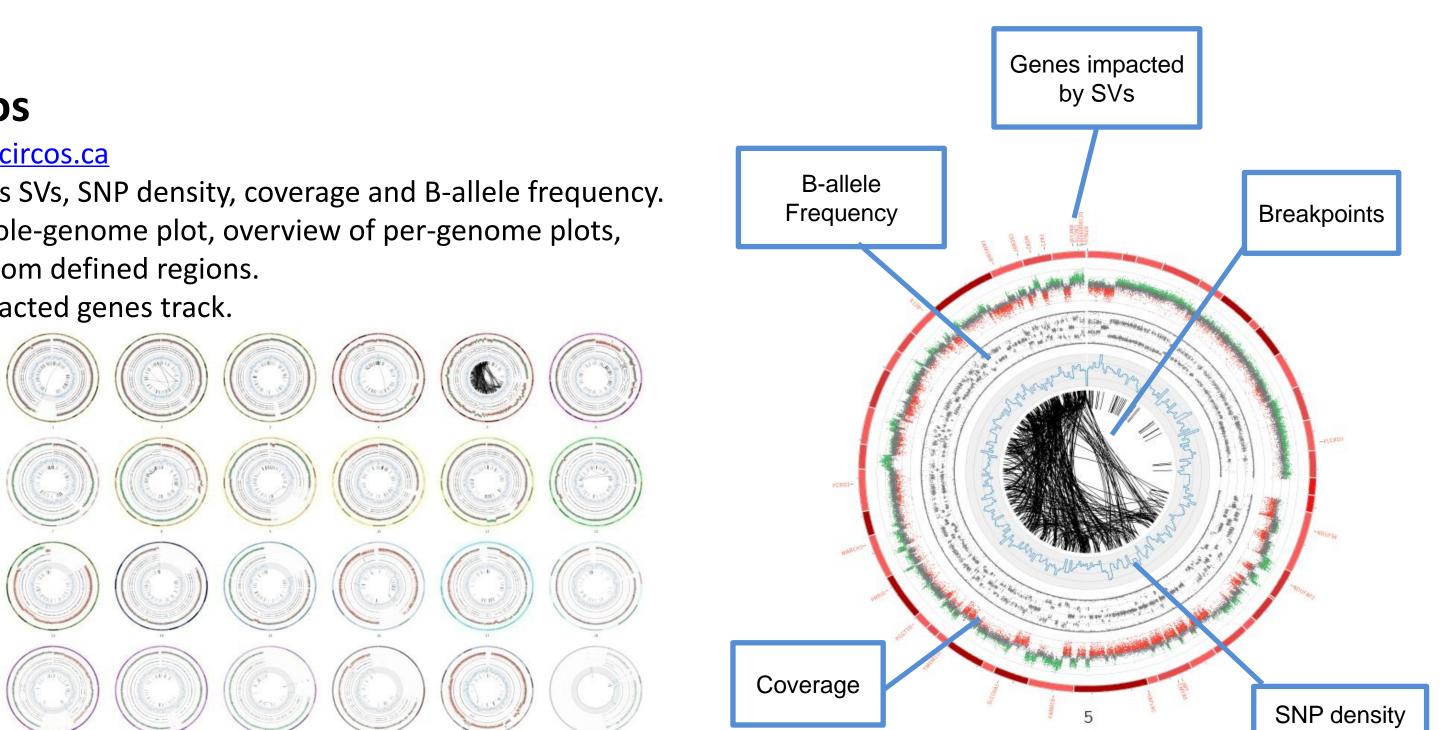
Increasing the number of *Virtual Normal* samples further allows for the filtering of increasingly rare polymorphisms.

#### Circos

#### http://circos.ca

Visualisation

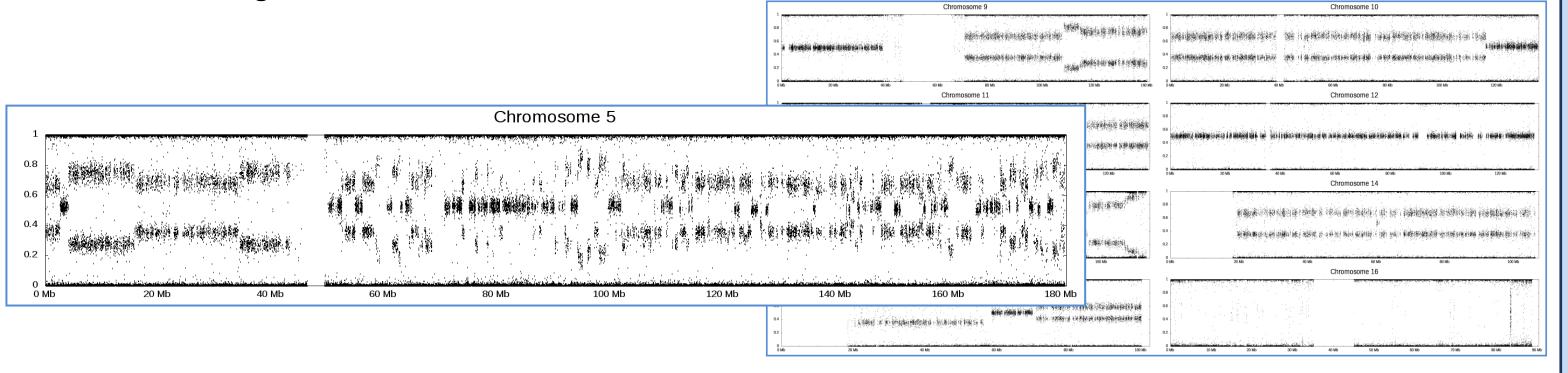
- Plots SVs, SNP density, coverage and B-allele frequency.
- Whole-genome plot, overview of per-genome plots, custom defined regions.
- Impacted genes track.



#### 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.





# **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

