Canadian Centre for Computational Genomics

BIOINFORMATICS CANCER ANALYSIS

Mathieu Bourgey, *Ph.D*

MonBUG March meeting

2016-03-09

McGill

Outline

0. C3G presentation

1. Cancer genomics intro

2. Cancer interesting tools

3. GSoC 2016

What is C3G ?

- Genome Canada supported center lead by Dr Guillaume Bourque (McGill University) and Dr Michael Brudno (University of Toronto)
- C3G provides bioinformatics analysis and HPC services and solutions for the life science research community.
- C3G Montreal is divided in 2 sub-unit:
  - Services
  - R&D

C3GMTL - SERVICES

- Whole genome/exome analysis
- Whole transcriptome analysis
- Whole transcriptome assembly
- Genome assembly
- Cancer Analysis
- Small RNA analysis
- Whole genome bisulfite and SeqCap-Epi analysis
- Metagenomics analysis
- ChIP-seq analysis
- High coverage amplicon analysis
- Microarray analysis
- Gene and parthway analysis
- Customized Analysis

Webpage: http://computationalgenomics.ca/

C3GMTL – RESEARCH AND DEVELOPMENT

- Analysis pipelines:
  - https://bitbucket.org/mugqic/mugqic_pipelines
  - Python based framework
  - Reduce micro-management
  - WES/WGS; RNA; RNA assembly; ChipSeq; Pacbio small genome assembly; Illumina runProcessing
- Maintain an up-to-date set of bioinformatics tools (>80) and resources (18) to the community
  - Available through a MUGQIC - Compute Canada partnership
  - Based on the use of CernVM-FS
- Provide and maintain a semi-private local instance of Galaxy using Compute Canada resources
  - Light version of our DNA and RNA pipeline are available
- Integrate new technology and development
  - Single cell
  - Haloplex
  - PacBio assembly for Eukaryote
  - HLA typing
  - Bioinformatics tools: SCoNEs, popSV, BVAtools,…
  - Bioinformatics pipeline development: WGBS, Metagenomics, ***Cancer analysis***

OUTLINE

0. C3G presentation

1. Cancer genomics intro

2. Cancer interesting tools

3. GSoC 2016

Cancer

- Cancer is one of the most common diseases in the developed world:
  - 3 out of 10 deaths are due to cancer in Canada
  - Lung cancer is the most common cancer in men
  - Breast cancer is the most common cancer in women
  - There are over 100 different forms of cancer

What causes cancer?

- Cancer arises from the mutation/alteration of a normal(s) gene(s).

- It is thought that several mutations need to occur to give rise to cancer

- Cells that are old or not functioning properly normally self destruct and are replaced by new cells.

- However, cancerous cells do not self destruct and continue to divide rapidly producing millions of new cancerous cells.

The genomic alterations model

- A wide range of genomic alterations can lead to the development of cancer:
  - Mutations
  - Copy number changes
  - Rearrangements

- Most of these alterations are somatic:
  - They are present in cancer cells but not in a patient's germ line

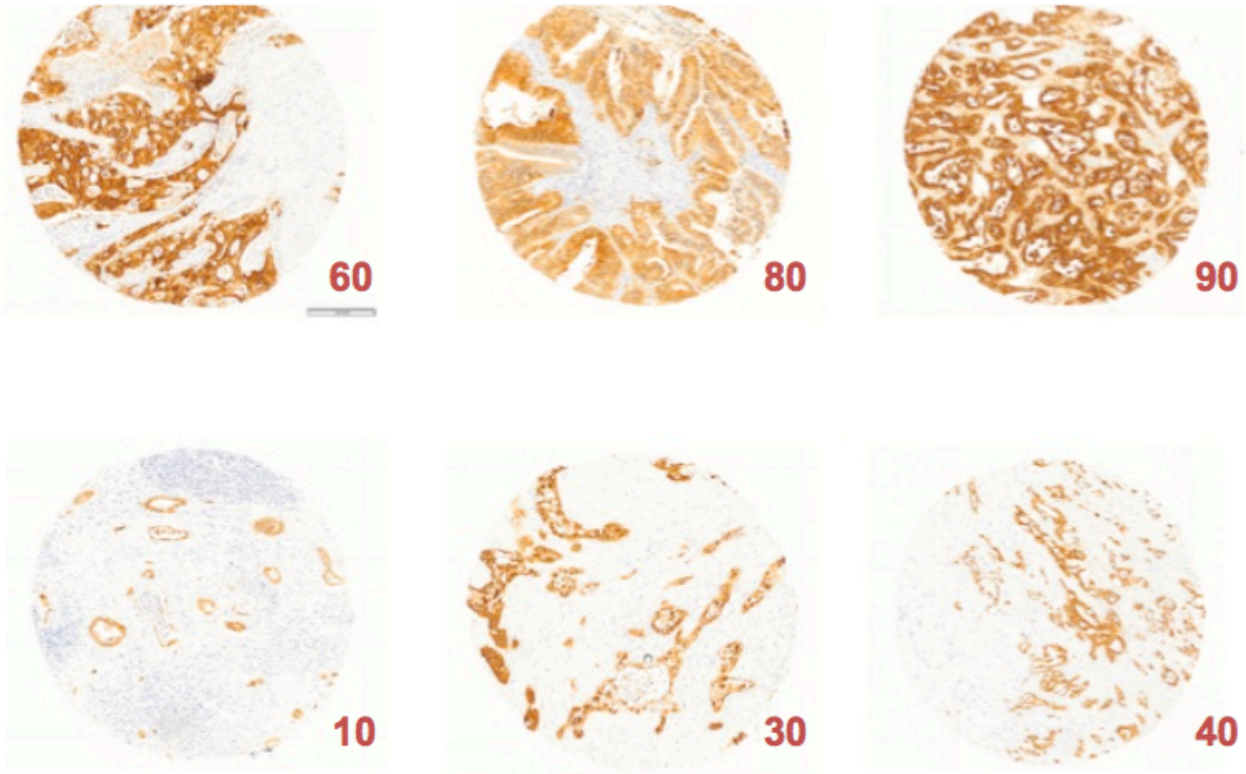
- Somatic genome alterations can provides potential therapies targeted:
  - Treatment with the inhibitors of the epidermal growth factor receptor kinase (EGFR) leads to a significant survival benefit in patients with lung cancer whose tumours carry *EGFR mutations, but no benefit in patients* whose tumours carry wild-type *EGFR*

Rosell, R. *et al. N. Engl. J. Med.*361, 958–967 (2009)

Characteristics of cancer samples

- Lower quantity
  - Quantity of material available is often smaller for cancer sample than for germ line sample
- Lower quality
  - Most cancer biopsy and resection specimens are formalin-fixed and paraffin-embedded (FFPE)
  - Cancer specimens often include substantial fractions of necrotic or apoptotic cells that reduce the average nucleic acid quality
- Lower purity
  - cancer specimen contains a mixture a mixture of cancer and normal genomes (**Cellularity**).
  - The cancers themselves may composed of different clones that have different genomes (heterogeneity)

Example of cellularity

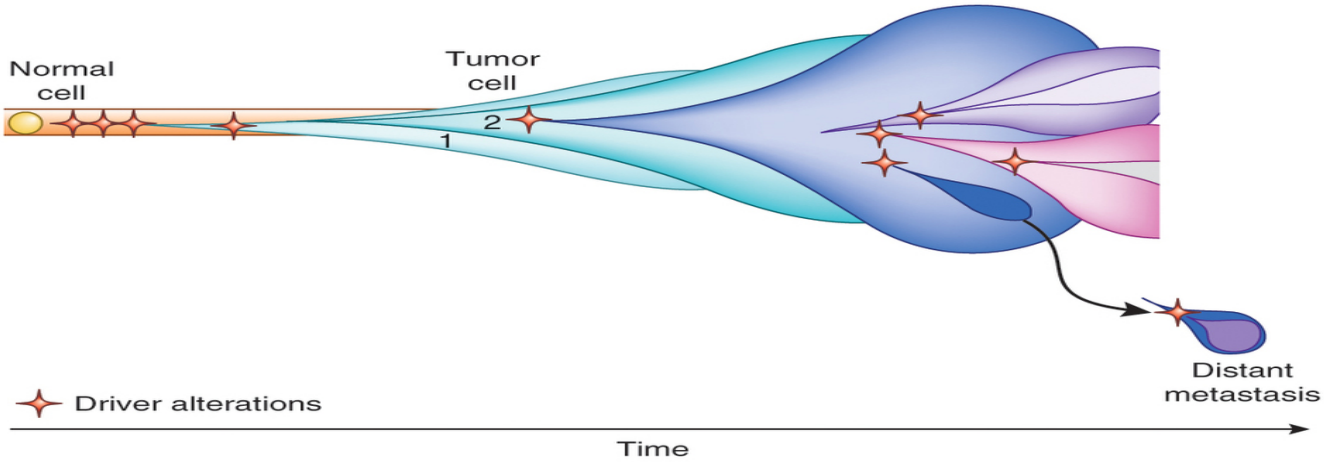


Modified from Dr Mark Cowley

Charateristics of cancer genomes

- Structural variability:
  - Mutation frequency (**Mutational signature**)
  - Global copy number or **Ploidy**
  - Genome structure

- Clonality



From Alizadeh et al. Nature Medicine 21, 846–853 (2015)

Cancer analysis challenge
Driver vs Passengers mutations

OUTLINE

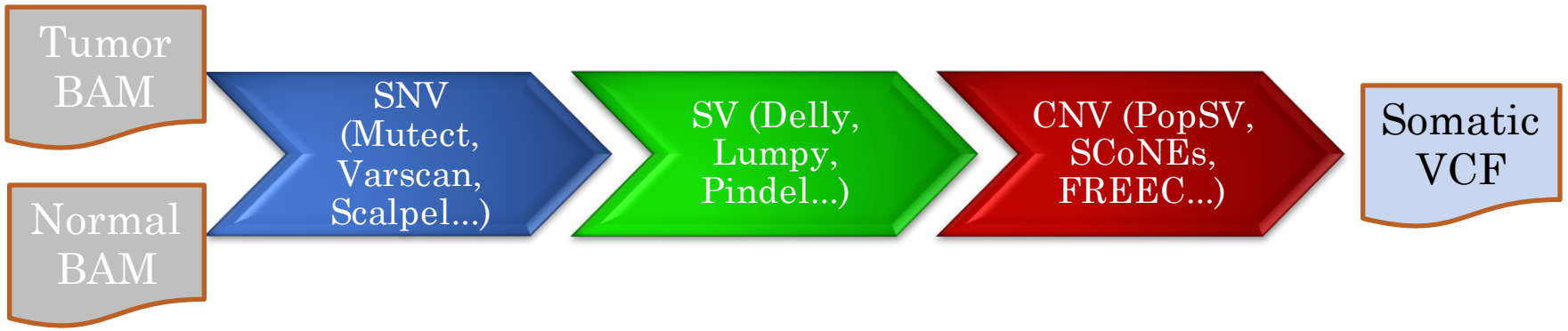
0. C3G presentation

1. Cancer genomics intro

2. Cancer interesting tools

3. GSoC 2016

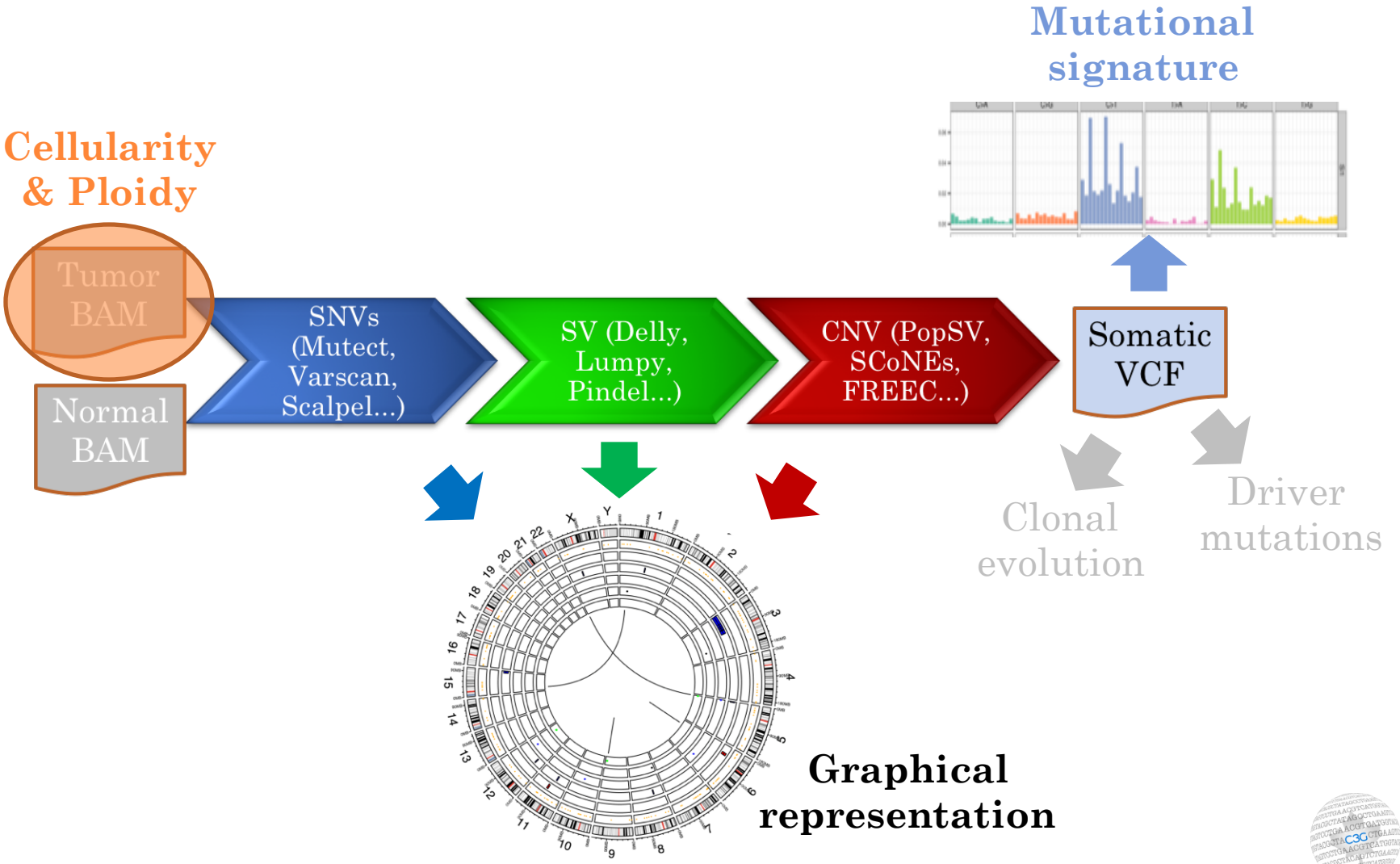
Cancer analysis summary



Lot of studies are conducted each year to define the set of best tools !

BUT IS IT ENOUGH ?

Extended cancer analysis



Canadian Centre for Computational Genomics

Celluarity and ploidy

Cellularity and ploidy

- As shown in introduction cancer analysis suffer from specificity of cancer sample (Cellularity) and of cancer genome (ploidy)
- Few cancer analysis tools take these parameters into account while it does variant calling. (SNV or SV)
- One method exists for CNV analysis using SNParray data:
  - ASCAT
- NGS methods:
  - ABSOLUTE
  - absCN-seq
  - *PurBayes*
  - *PurityEst*
  - *Titan*
- 2 methods have tried to extend ASCAT to NGS data:
  - Battenberg algorithm (from ASCAT co-author; alpha version)
  - ***Sequenza (published, good R vignettes)***

Annals of Oncology 26: 64–70, 2015
doi:10.1093/annonc/mdu479
Published online 15 October 2014

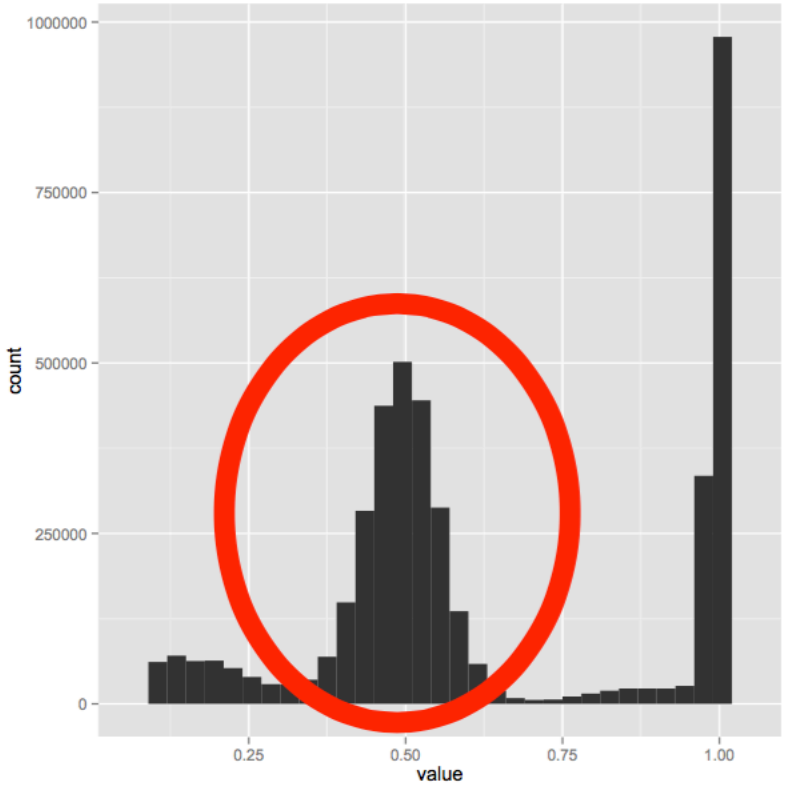
Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data

F. Favero[1], T. Joshi[1], A. M. Marquard[1], N. J. Birkbak[1], M. Krzystanek[1], Q. Li[1,2], Z. Szallasi[1,3] & A. C. Eklund[1*]

[1]*Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark;* [2]*Medical School, Xiamen University, Xiamen, China;* [3]*Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology (CHIP@HST), Harvard Medical School, Boston, USA*

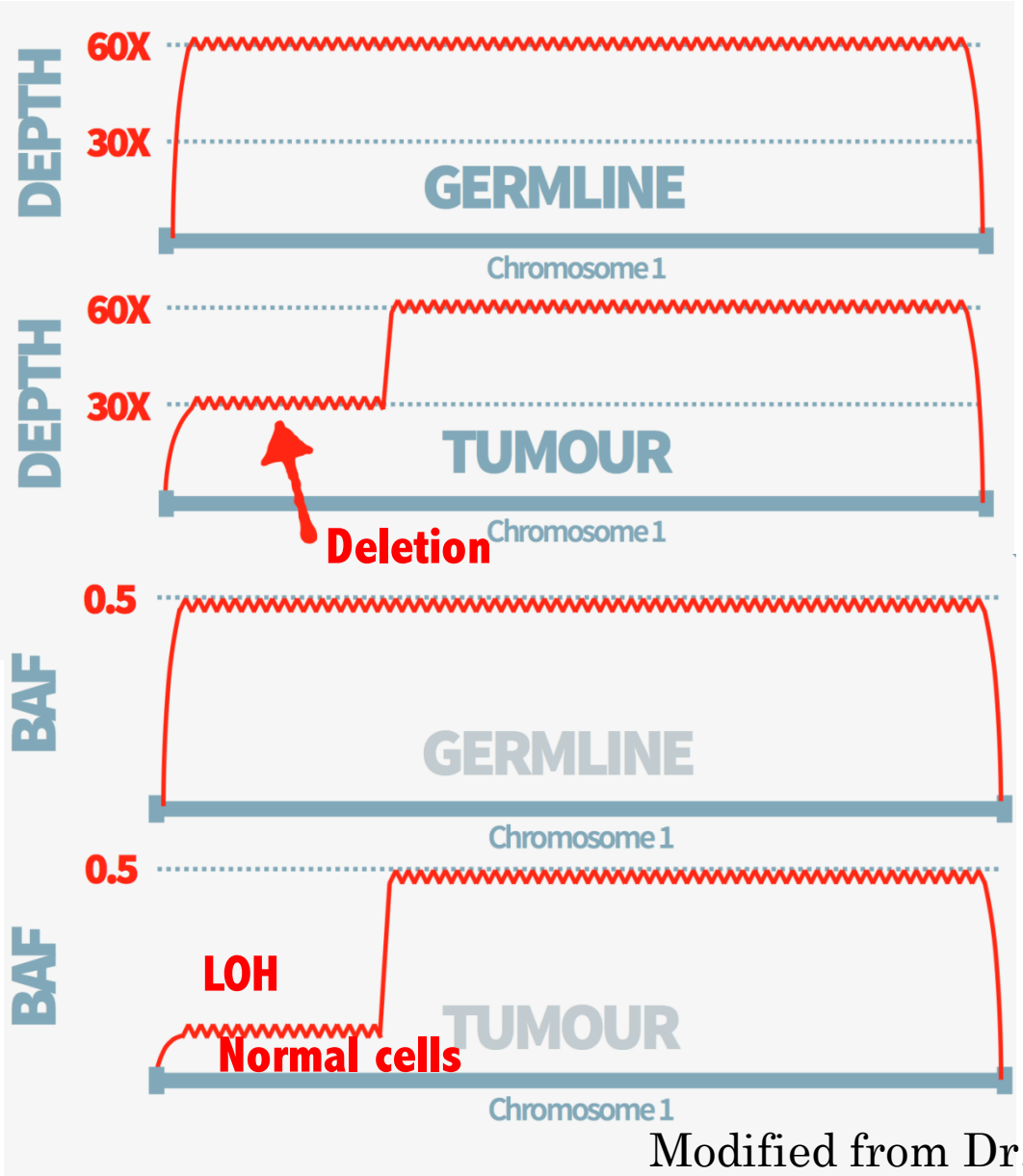
Received 7 April 2014; revised 8 October 2014; accepted 9 October 2014

- Use read depth and BAF to estimate cellularity and ploidy.
- Use cellularity estimate to improve variants and LOH calls.



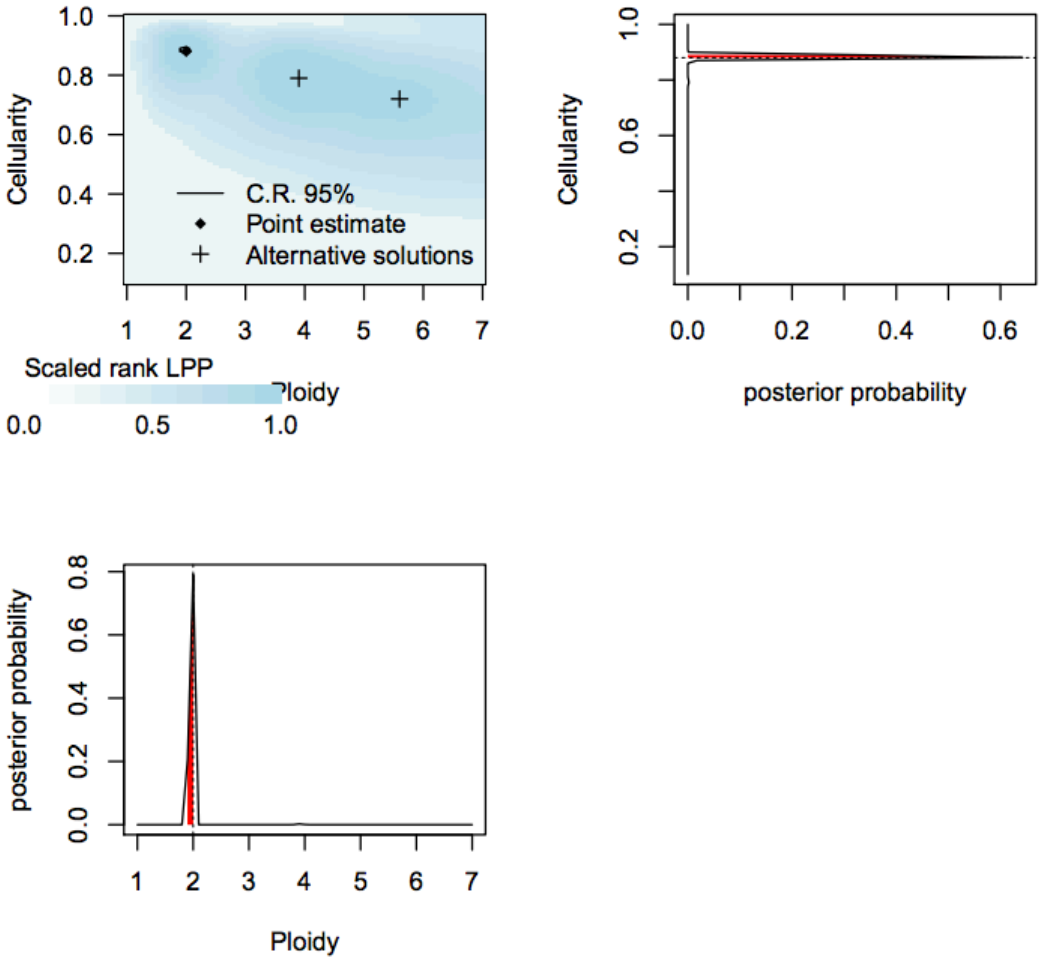
Modified from Dr. Velimir Gayevskiy

Cellularity estimation principle

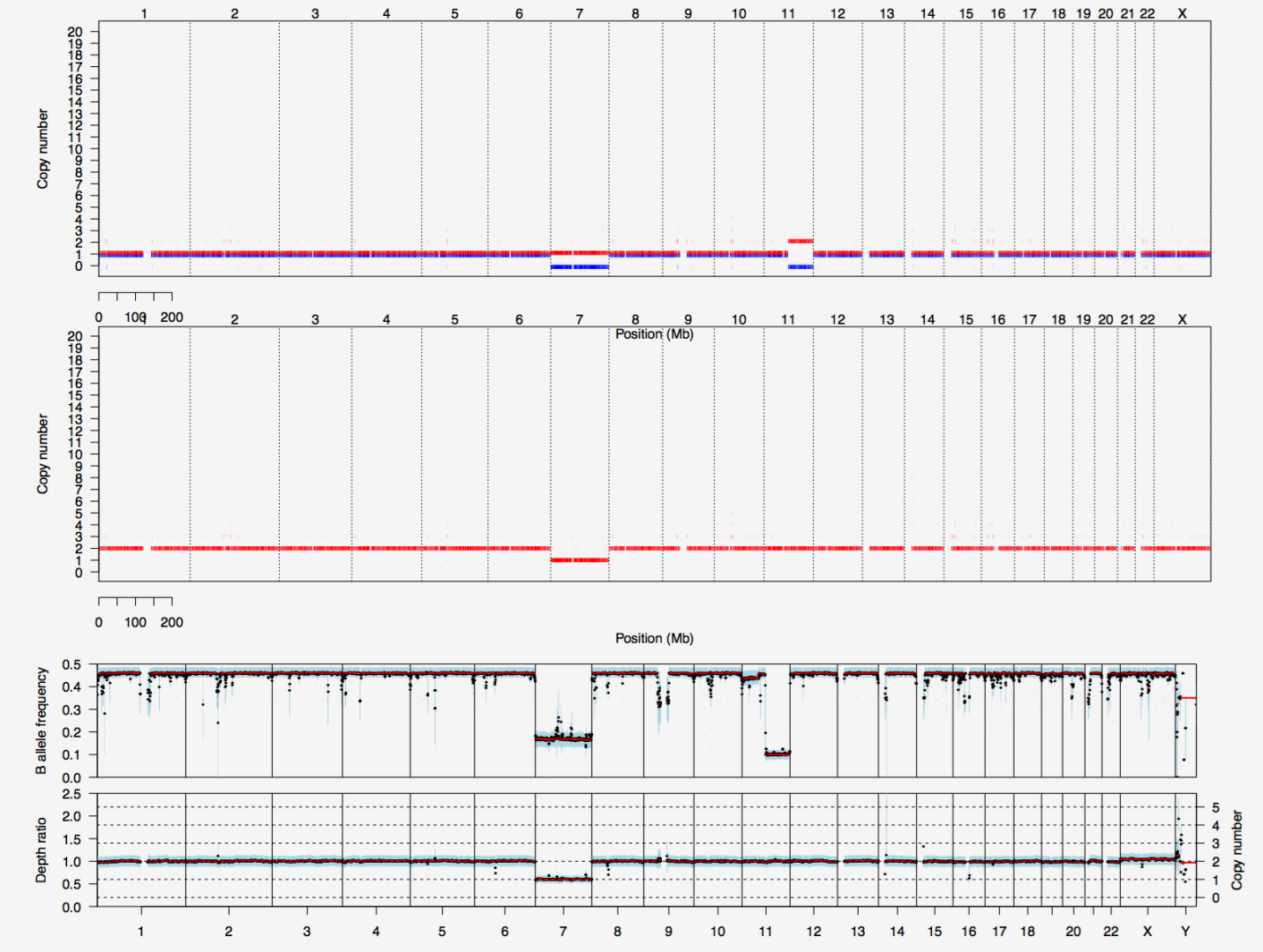


Modified from Dr. Velimir Gayevskiy

Outputs – parameters estimates

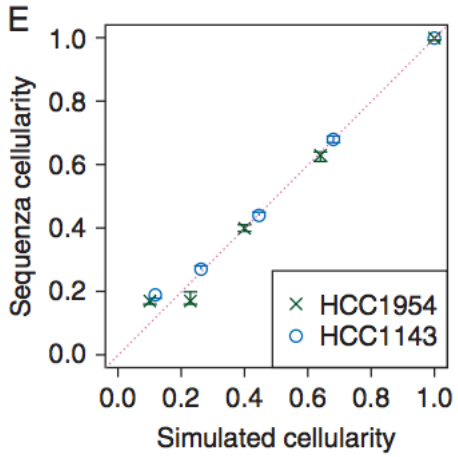


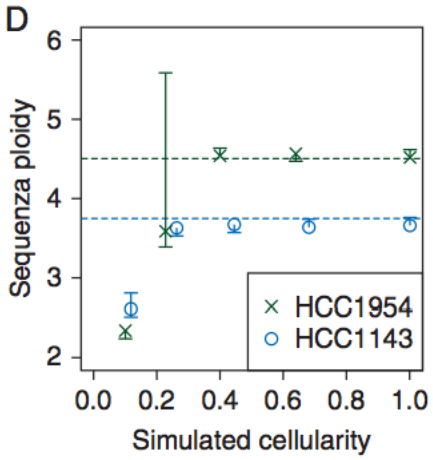
Outputs – Genome view



VALIDATION – SIMULATED DATA

- WGS, aligned to the hg19 genome at ×30 of coverage, of two cell lines HCC1143 and HCC1954, (matching normal blood)
- Simulated admixtures at tumor cellularity of 20%, 40%, 60%, and 80%





Cellularity estimates are goods and ploidy estimates seems accurate when cellularity values >= 30 %

VALIDATION – COMPARISON WITH OTHERS

- Run Sequenza, ABSOLUTE and absCN-seq on 30 WES data from TCGA and compare to ASCAT

Algorithm	r_ρ	r_ψ	$F_{\Delta CN=0}$
Sequenza	0.90 (0.91)	0.42 (0.94)	0.69
ABSOLUTE	0.19 (0.61)	0.13 (0.50)	0.08
absCN-seq	0.46 (0.65)	−0.26 (0.46)	0.02

Sequenza show an higher accuracy than its competitors

Application to real data

- By the authors:
  - Among the 29 renal cancer samples sequenza call 17 3p deletion (know to affect 70-80% of this cancer patients)
- By myself:
  - Applying the methods to 74 cagekid non-conventional paired WES sample.
  - Trying to understand:
    - Why some sample are hypermutated
    - Some interesting somatic mutations

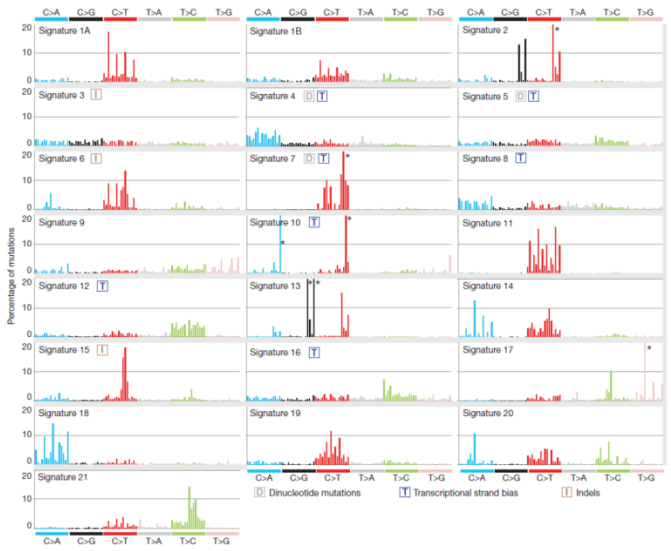
chromoson	position	ref_allele	alt_alleles	gene_name	sequenza cellularity
12	133202816	C	T	POLE	
Patients with this POLE mutation					
	Depth	Allele Frequency	Variant normal	Variant tumor ratio	
K2150038	172 , 293	C:102 A:0 T:71 N:0 G:0,C:173 A:0 T:120 N:0 G:0	0,410404624	0,409556314	0.48
LR265	215 , 207	C:1 A:0 T:214 N:0 G:0,C:39 A:0 T:170 N:0 G:0	0,995348837	0,813397129	0.24
LR368	149 , 147	C:58 A:0 T:91 N:0 G:0,C:60 A:0 T:89 N:0 G:0	0,610738255	0,597315436	0.87
LR412	72 , 264	C:42 A:0 T:31 N:0 G:0,C:125 A:0 T:140 N:0 G:0	0,424657534	0,528301887	0.66
K2110054	299 , 205	C:147 A:0 T:152 N:0 G:0,C:110 A:0 T:95 N:0 G:0	0,508361204	0,463414634	0.49

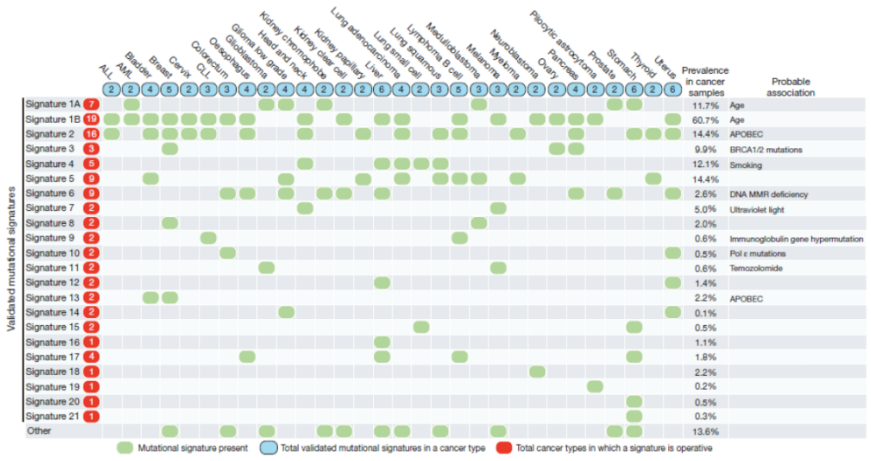
Canadian Centre for Computational Genomics

Mutational signature

What is somatic signature?

- The most common genetic model for cancer development is the accumulation of DNA mutations.
- Recently Alexandrov et al. (Nature 2013) developped a method to ***cluster cancer sample together by the type of the mutation and also what the neighbouring bases are.***
- Common mutational processes that are regularly identified in cancer sequencing have been linked to specific signature:





Analyzing somatic signature

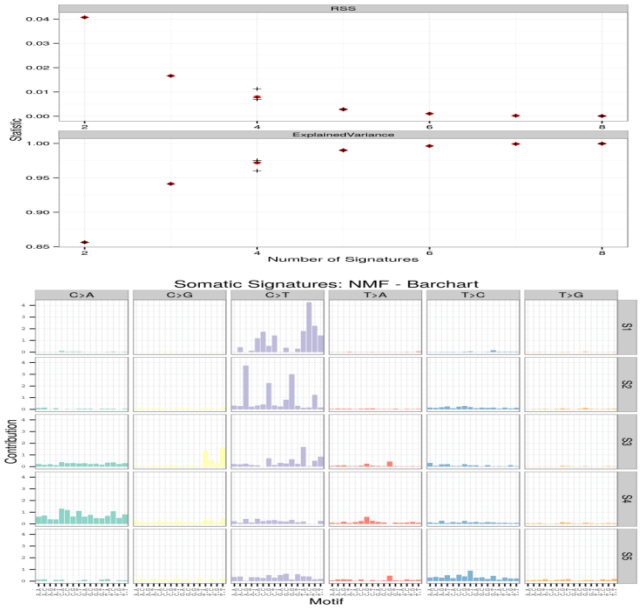
- Extraction of somatic signature
  - Alexandrov's method
    - WTSI Mutational Signature Framework
    - Matlab script
    - Not user friendly
    - Proprietary software
  - **J. Gehring's implementation *(Bioinformatics 2015)***
    - SomaticSignature R package
    - In Bioconductor
    - User friendly
    - With a good vignette

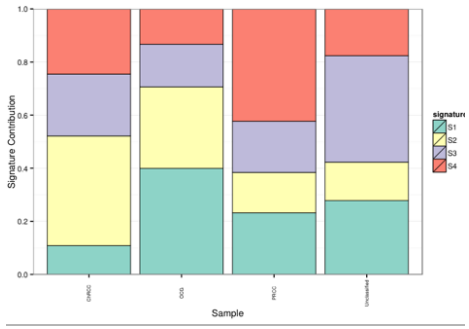
SOMATICSIGNATURE PRINCIPLE

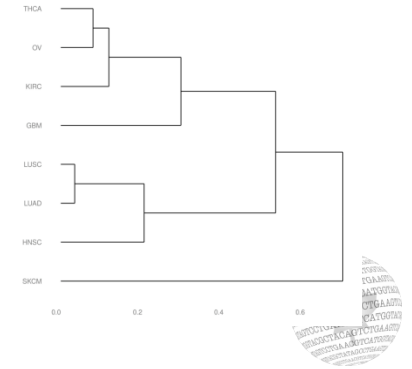
- The somatic motifs for each variant are retrieved from the reference sequence
  - mutationContext function and
- They are converted to a matrix representation
  - motifMatrix function.
- Somatic signatures are estimated with a method of choice
  - nmfDecomposition (alexandrov)
  - pcaDecomposition (large data set)
- The somatic signatures and their representation in the samples are assessed with a set of accessor and plotting functions.

How it works ?

- Once you generated the mutation context matrix
  - Find out how many signatures you have in the data
    - assessNumberSignatures
  - Identify the corresponding signature
    - identifySignatures
  - Determine signature contribution
    - plotSamples
  - Samples clustering
    - clusterSpectrum



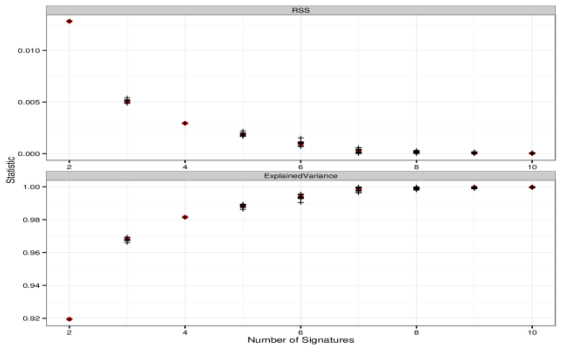


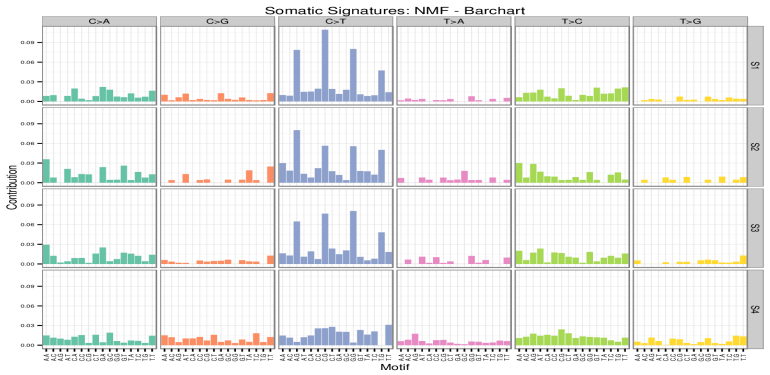


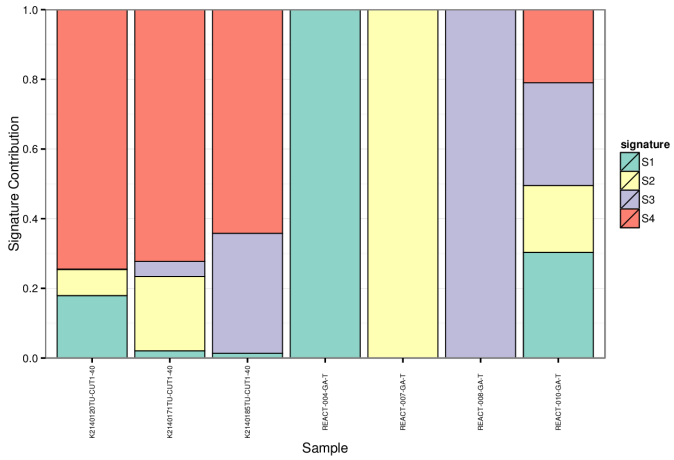
Canadian Centre for Computational Genomics

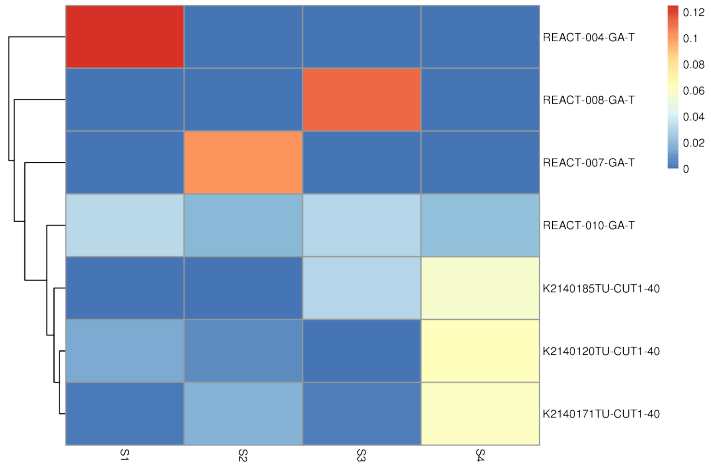
Real exemple

- Mixing 3 samples from kidney cancer with 4 samples from Typhoid cancer







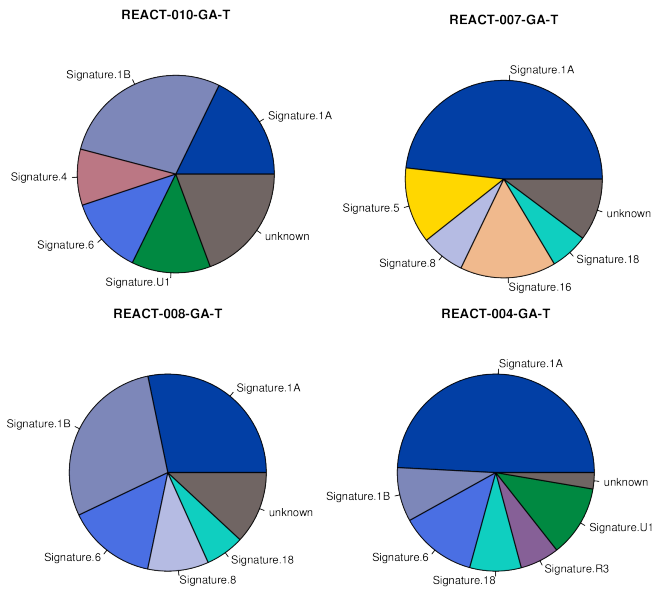


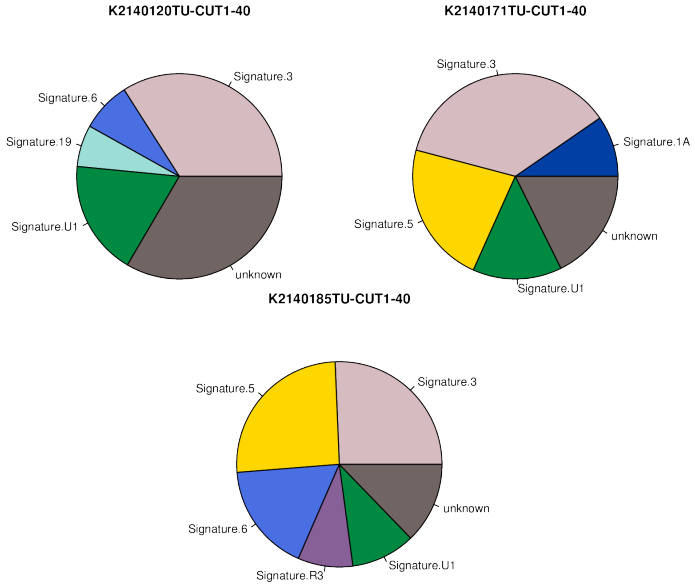
What is missing ?

- The limitation of this method is to not provide a way to link identified signature with a set of predefined signature
  - Is my signatures correspond to one of those detected by Alexendrov (or by a competitor) ?
- Riku et al. (Nat Gen 2015) proposed to use of the mean Kullback-Leibler divergence.
  - Available from R package entropy
  - It suffer from overparmetrization
  - Most of the time results are ambigous
    - All signature point to same one
- 2 weeks ago Rosenthal et al. (Genome Biology 2016) release deconstructSigs
  - Cran R package
  - I just start to test it
  - It is very promising

Testing DeconstructSig on real exemple

- Typhoid cancer
  - Mostly drive by signature 1.A and 1.B
    - Age
- Kidney
  - Mostly drive by signature 3 and Unknown
    - BRCA1/2 mutation



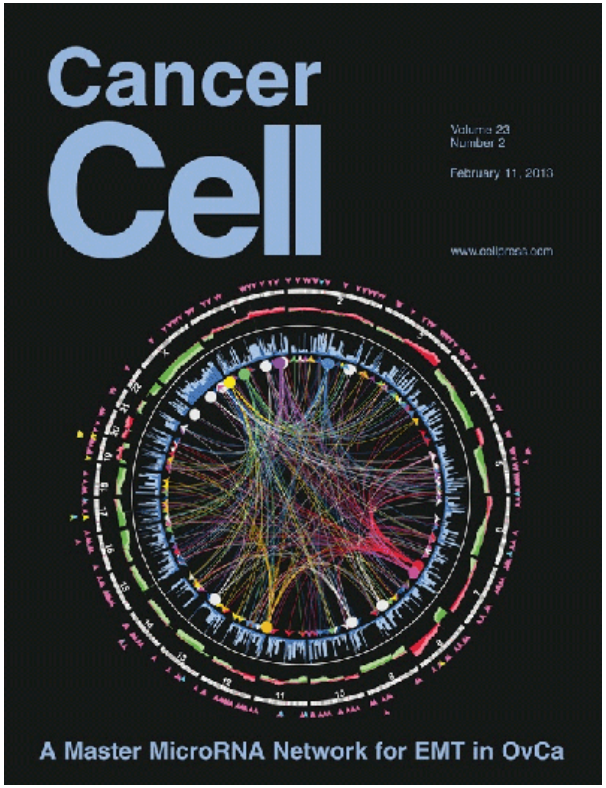


Canadian Centre for Computational Genomics

Visualization of cancer variants

Cancer data

- A major goal of cancer genomics analysis is to identify genetic changes which lead to the tumor evolution.
- In a more technical view, we ends-up with different types of genomics changes Point mutations Copy number variations Structural variants (insertion, inversion, deletion, duplication, translocation)
- A common visualization of these data is to give an overview of change for the 23 chromosome of the genome in circular view.



CIRCOS PLOT

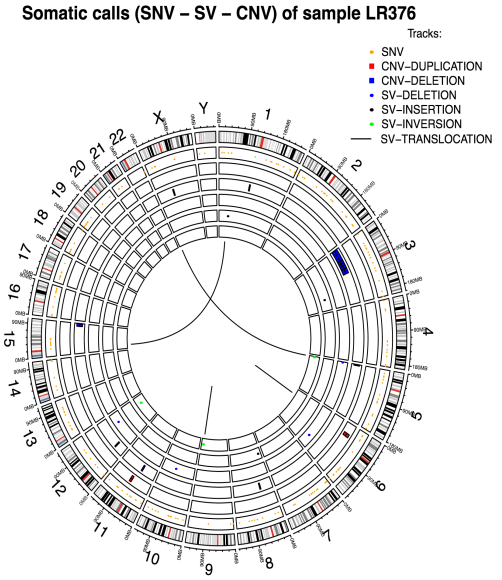
- What are the different options to draw it ?
  - Circos (www.circos.ca)
    - Not user friendly
    - Only recommended for very complex plot
  - Rcircos (R package, *Zhang H et al. Bioinformatics 2013*)
    - Limited vignettes
    - A little bit more complicated than circlize but a nicer rendering
  - circlize (R package, *Gu Z et al. Bioinformatics 2014*)
    - User friendly
    - Use bed format
    - Good vignettes
    - For genomic data: A set of methods and a specific vignette
    - Generic methods and vignettes

CIRCLIZE GENERIC WORKFLOW

- Set-up the generic graphical parameters
  - circos.par function
- Draw reference ideograms
  - circos.initializeWithIdeogram function
- draw 1 track for each type of variants
  - circos.genomicTrackPlotRegion function +
    - circos.genomicPoints function
    - circos.genomicRect function
    - circos.genomicLines function
- Draw positional links to represent transclocations
  - circos.link function
- A good graph needs title and legends
  - Use generic R methods

Real data example

- See details at: https://github.com/mbourgey/EBI_cancer_workshop_visualization/tree/Montreal_R_UserGroup
- Starting from 3 files
  - Breakdancer tsv for SVs
  - Mutect vcf for SNVs
  - SCoNEs tsv for CNVs



OUTLINE

0. C3G presentation

1. Cancer genomics intro

2. Cancer interesting tools

3. GSoC 2016

Google Summer of Code (GSOC)

- Student gets $5000 for writing open source code for 3 months.
- Timeline:
  - Feb - Admins for open source organizations e.g. R, Bioconductor apply to Google.
  - Mar - Mentors suggest projects for each org. Students submit project proposals to Google. Google gives funding for n students to an org.
  - Apr - The top n students get $500 and begin coding.
  - Jul - Midterm evaluation, pass = $2250 (for student).
  - Aug - Final evaluation, pass = $2250 (for student).
  - Nov - Orgs get $500/student mentored

What makes a good GSOC project?

- Coding projects should:
  - Result in free/open-source software.
  - Be 3 months of full time work for a student.
  - Could include writing documentation and tests.
  - Should not include original research.

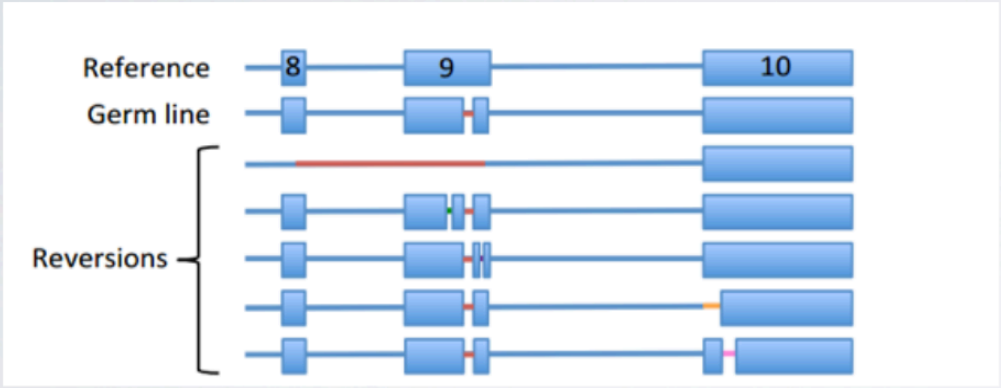
C3G IS AN ACCEPTED ORGANISATION FOR GSOC 2016

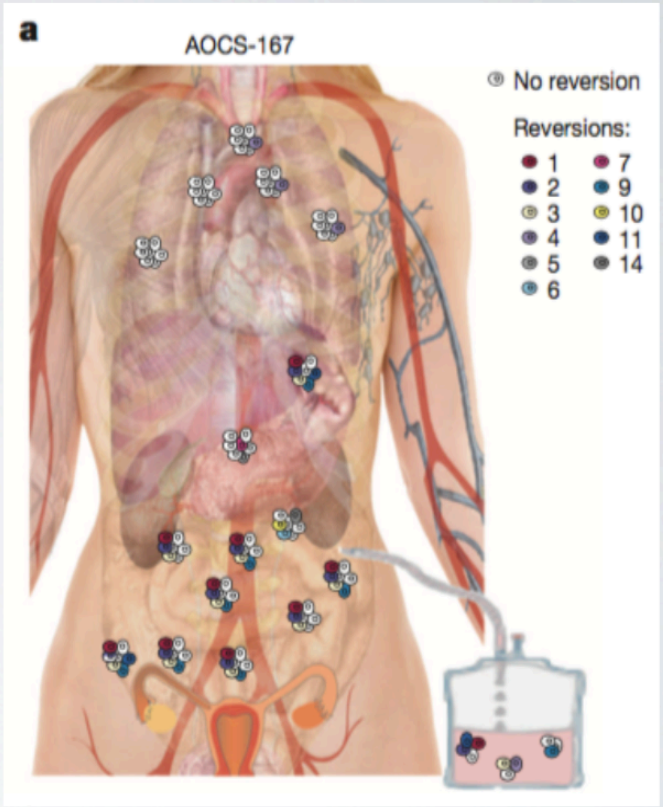
- https://summerofcode.withgoogle.com/organizations/5649665110310912/
- Details at: https://bitbucket.org/mugqic/gsoc2016
- List of proposed project so far :
  - Flowchart creator for MUGQIC Pipelines
  - Implement base modification analysis in the pacbio_assembly pipeline from MUGQIC Pipelines
  - Integrate structural variants calls in the tumor_pair pipeline from MUGQIC Pipeline
  - Improve SegAnnDB interactive genomic segmentation web app
  - Develop a noise reduction engine for SCoNEs
  - Develop add-ons for SCoNEs
  - EGA Data Submission database (mEGAdata)
  - Development of an HTML dynamic matrix to represent datasets with multiple dimensions
  - Population genetics simulation and modelling

Canadian Centre for Computational Genomics

BONUS – SURPRIZING CANCER FEATURE

Ovarian Cancer - somatic reversions of inherited BRCA2 mutations





- germline 5bp deletion in BRCA2
- patient acquired resistance to PARP inhibitor olaparib and to carboplatin
- In 18 samples from same patient, 15/19 somatic mutations identified restored the reading frame

From Patch et al. Nature 2015

Acknowledgment

- C3G-Montreal
  - Guillaume Bourque
  - Robert Everleigh
  - Toby Hocking
  - François Lefebvre
  - Jean Monlong
  - Edouard Henrion
- Riazalhosseini's lab
  - Yasser Riazalhosseini
  - Madeleine Arseneault
- External:
  - Velimir Gayevskiy *(Garvan Insitut)*
  - Ann-Marie Patch *(Queensland Centre for Medical Genomics)*
  - Louis Letourneau *(sportlogIQ)*