

Applied Bioinformatics

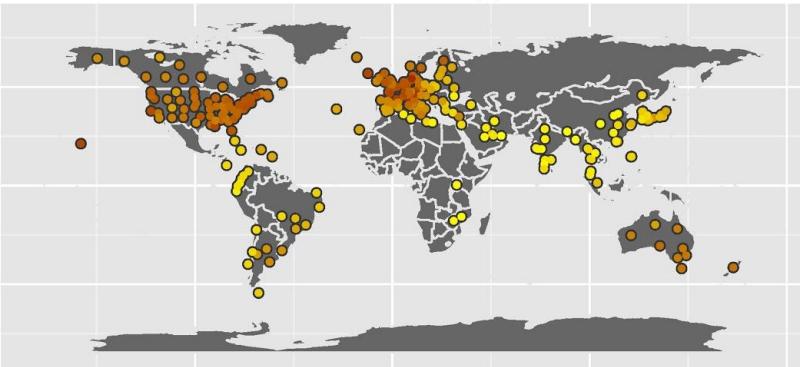
Cancer Genomics II

Dec, 2015

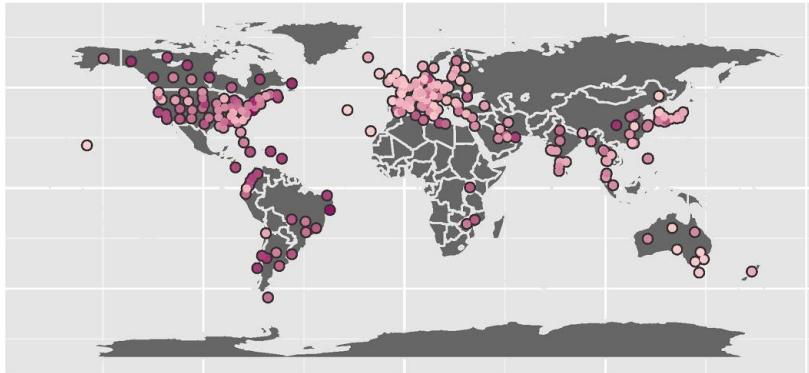
Vladan Arsenijevic, PhD
vladan.arsenijevic@sbgenomics.com

SevenBridges

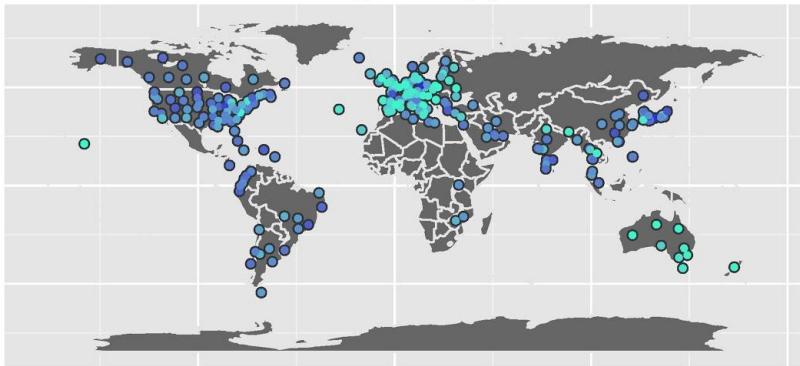
Breast cancer (F)



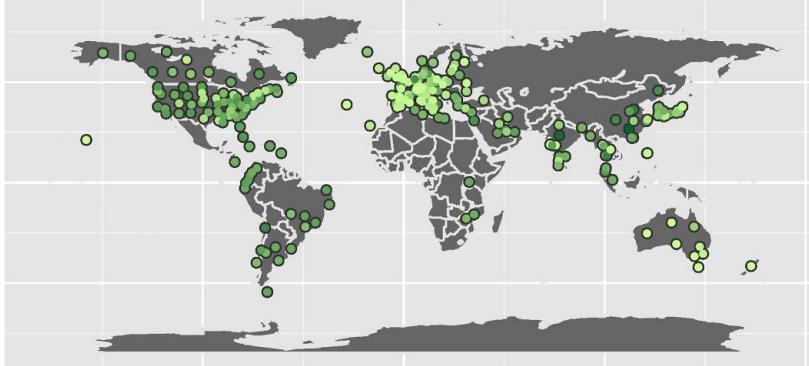
Lung cancer (F)



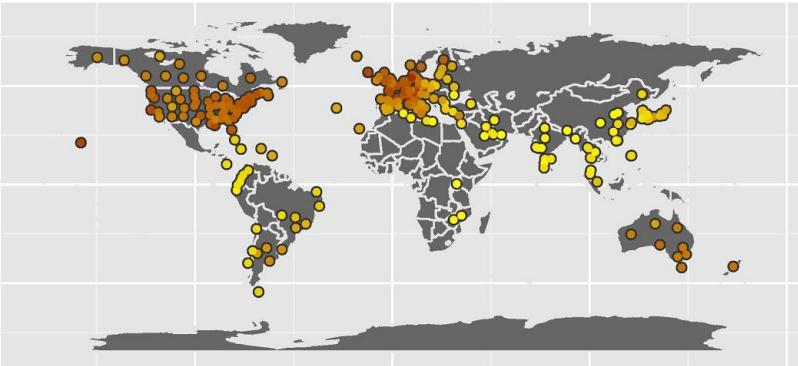
Lung cancer (M)



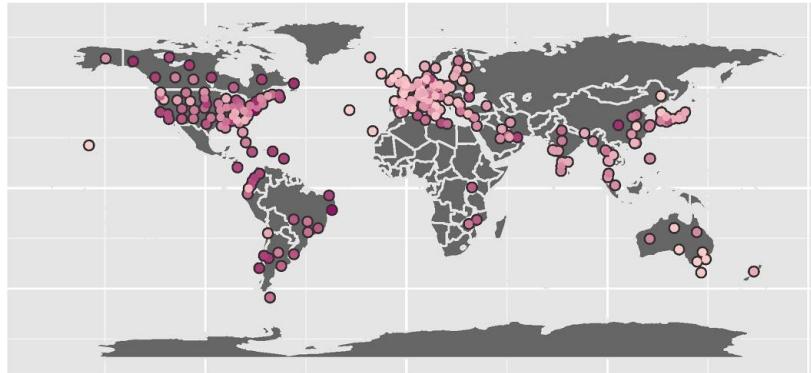
Prostate cancer (M)



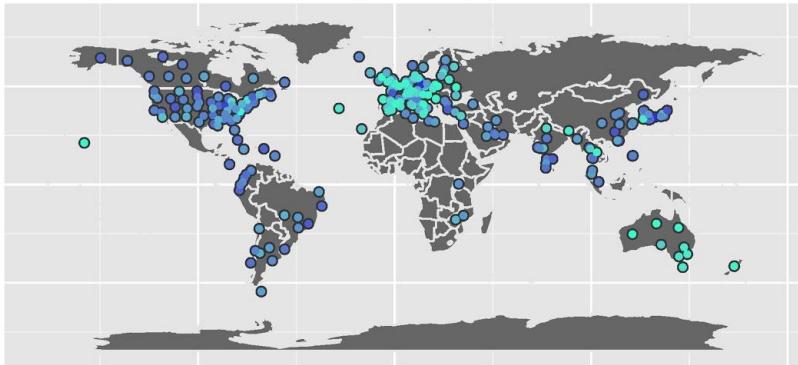
Breast cancer (F)



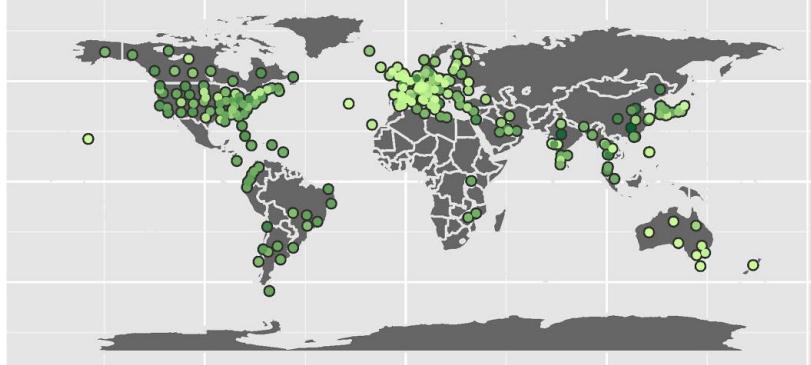
Lung cancer (F)



Lung cancer (M)



Prostate cancer (M)

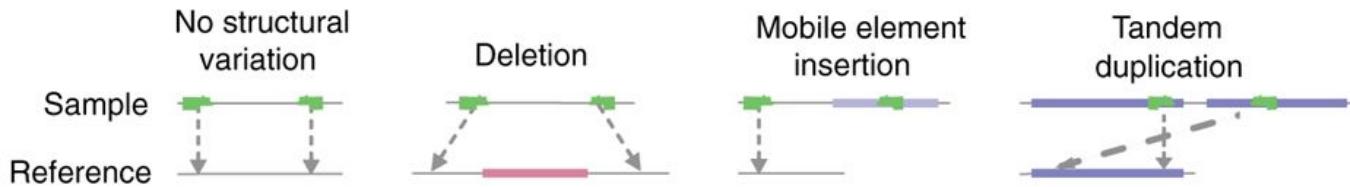


Each year globally, about 14 million people learn they have cancer, and 8 million people die from the disease. Research suggests that one-third of cancer deaths can be prevented, but sometimes services and technologies are not widely available, especially in low- and middle-income countries.

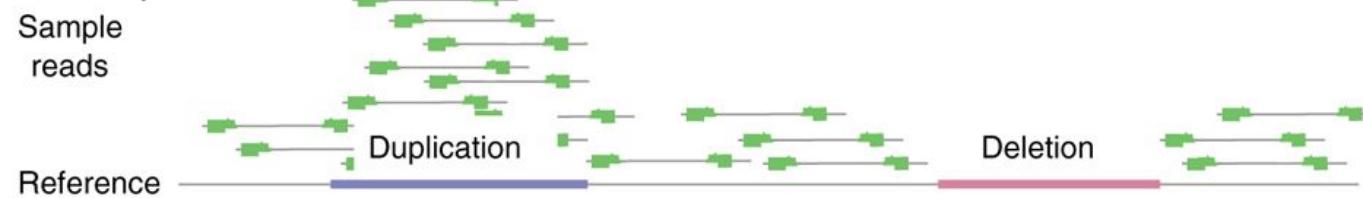
Overview

- Identification of:
 - Copy Number Variations (CNVs)
 - Fusion Genes
- A touch of temporal evolution in tumors
- Gene Expression signature (that can characterize tumors)
- Future directions

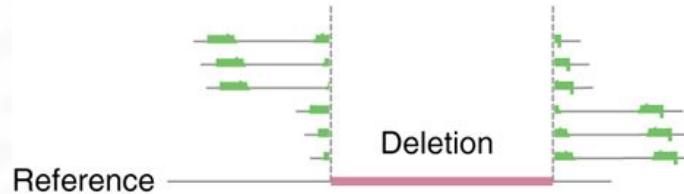
Read pairs



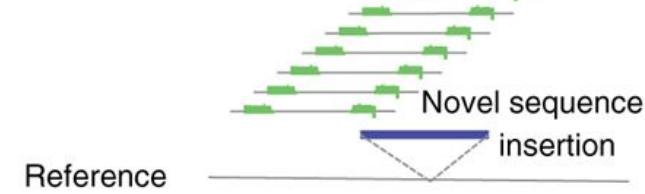
Read depth

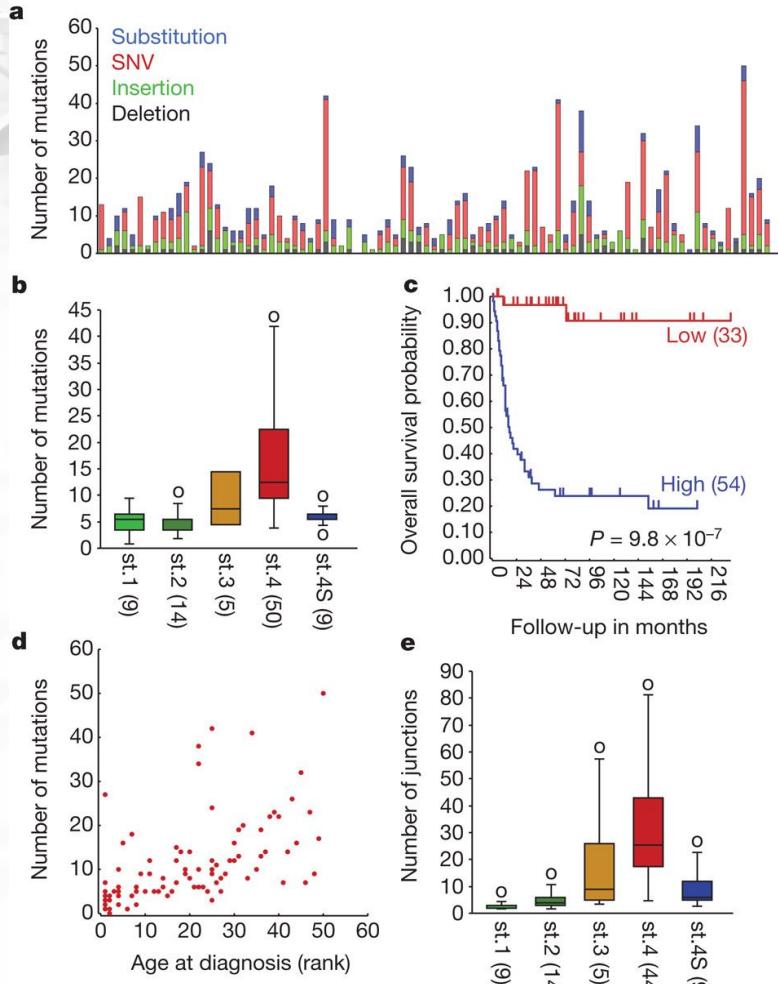


Split reads



Assembly

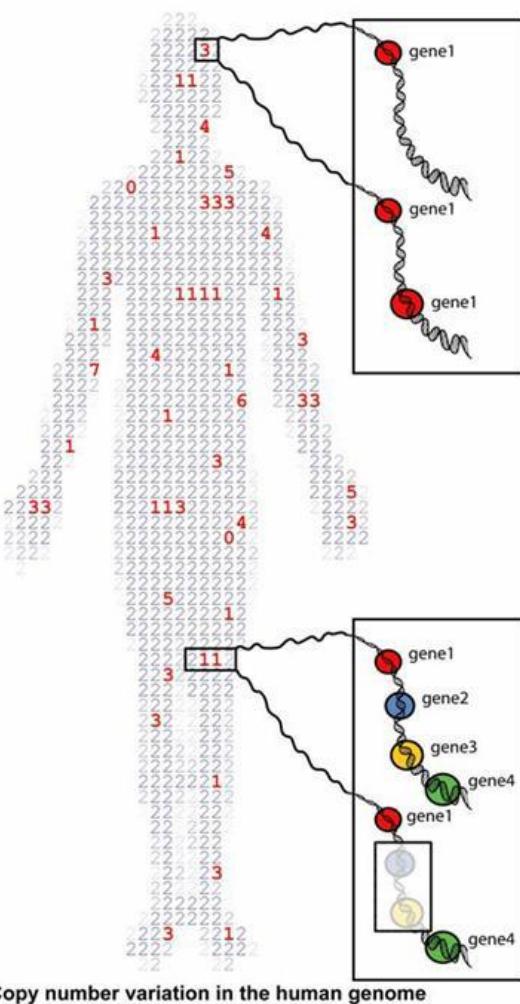




Frequency of somatic mutation (in neuroblastoma) correlates with age, tumor stage, survival

Molenaar et al. Nature (2012)

CNVs



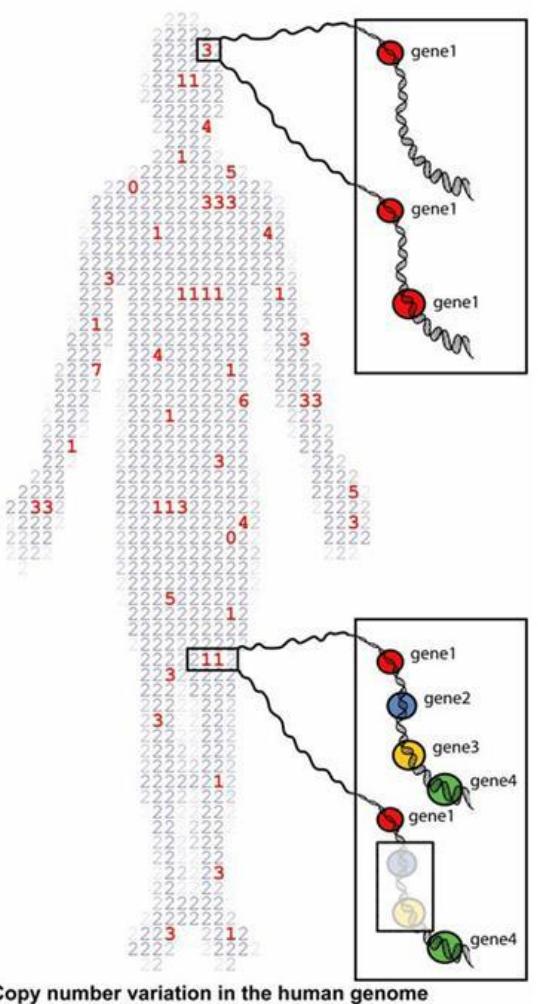
Copy number variation in the human genome

ACGTATATAGT**CTAT**GCTAGCTTAGCA

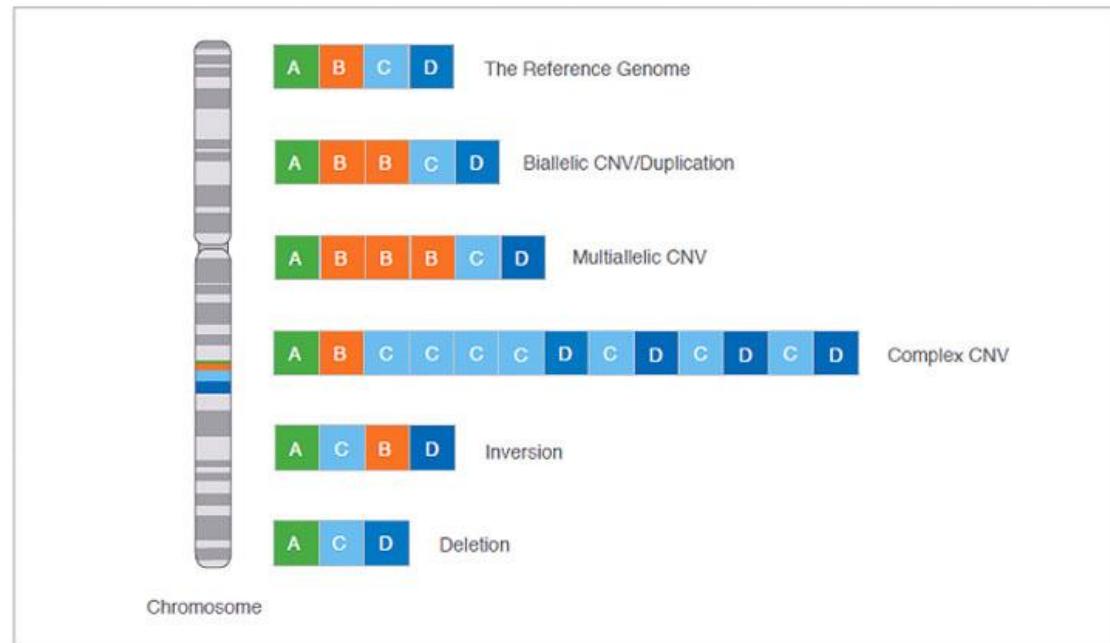
ACGTATATAGT**CTATCTATCTAT**GCTAGCTTAGCA

- A form of structural variation: cells have different number of copies of DNA sections
- CNVs account for roughly 13% of human genomic DNA (most CNVs are stable and inheritable)
- Most CNV gains outnumber losses suggesting they are favoured in evolution: AMY1 is present in 2 copies in our relatives but humans may have as many as 15. This probably comes as an adaptation to starch rich diet
- Some are associated with susceptibility or even resistance to disease: higher copy-number of CCL3L1 increases resistance to HIV infection!
- BUT it's not all good news: gene copy-number is often elevated in cancer cells
- AND ... those genes whose expression and DNA copy-number are increased in cancer are involved in core cancer pathways

CNVs

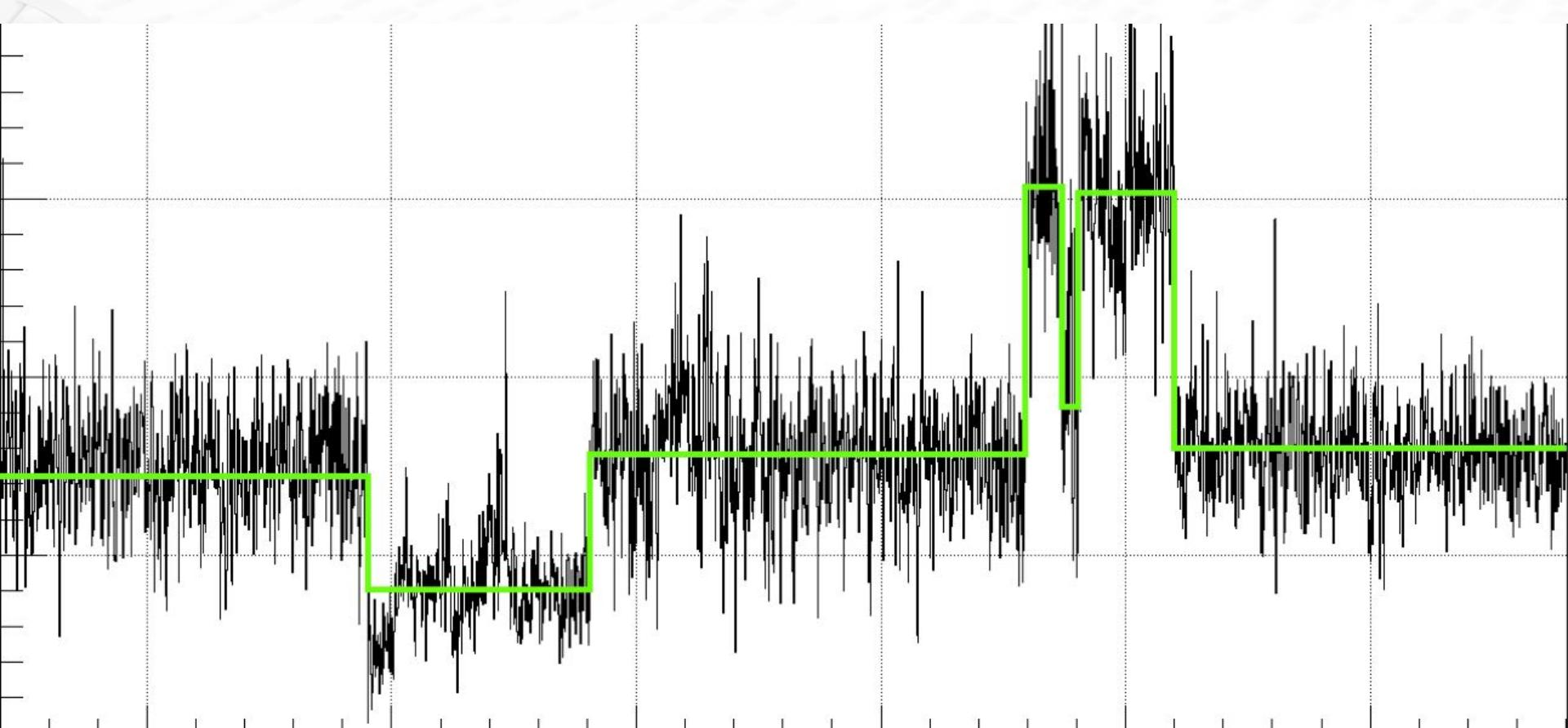


Copy number variation in the human genome



CNV detection: VarScan2, ControlFREEC, TitanCNA, cn.mops,...

- Due to accumulation of high coverage NGS data, read depth approach has gained popularity in recent years
- How read depth method works:
 - Mapping: Control sample works as a reference to which reads are mapped and this is used to calculate the read depth within predefined regions or windows
 - Correction: GC content is taken into account in order to correct read depth calculation
 - Segmentation: Genomic regions with a similar copy number are merged

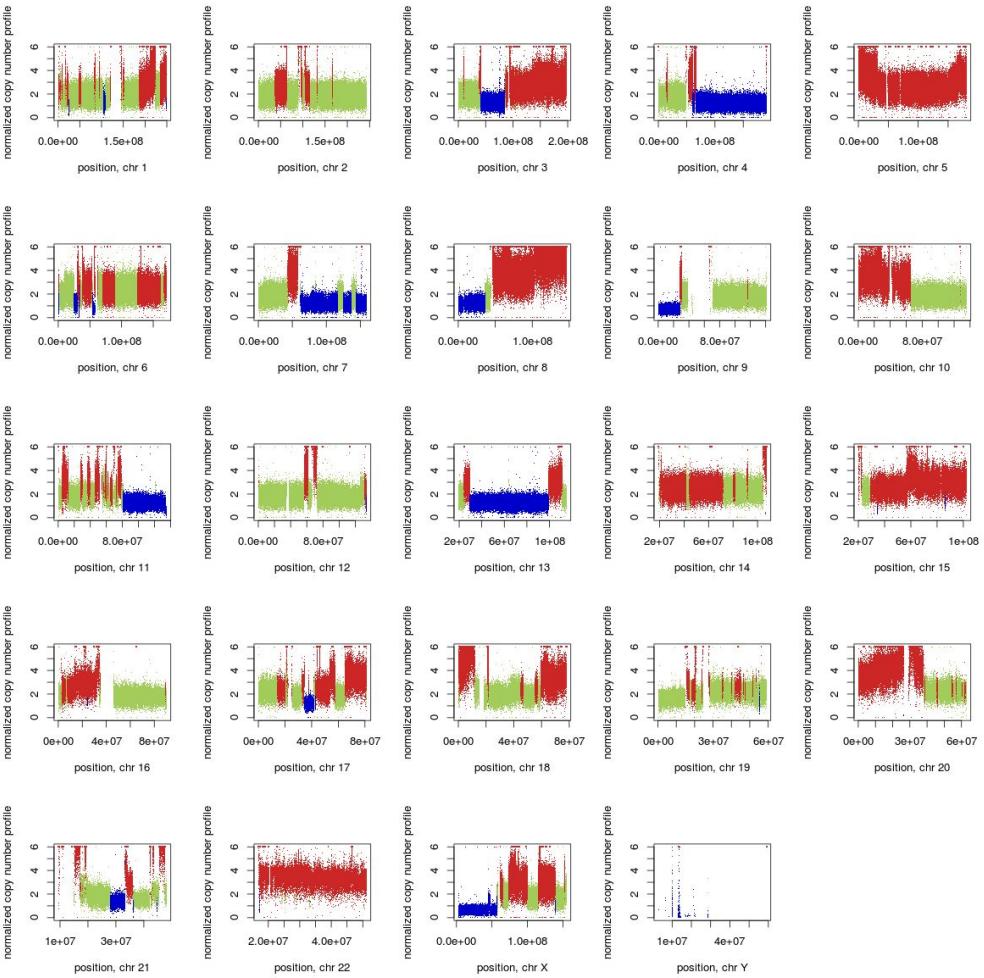


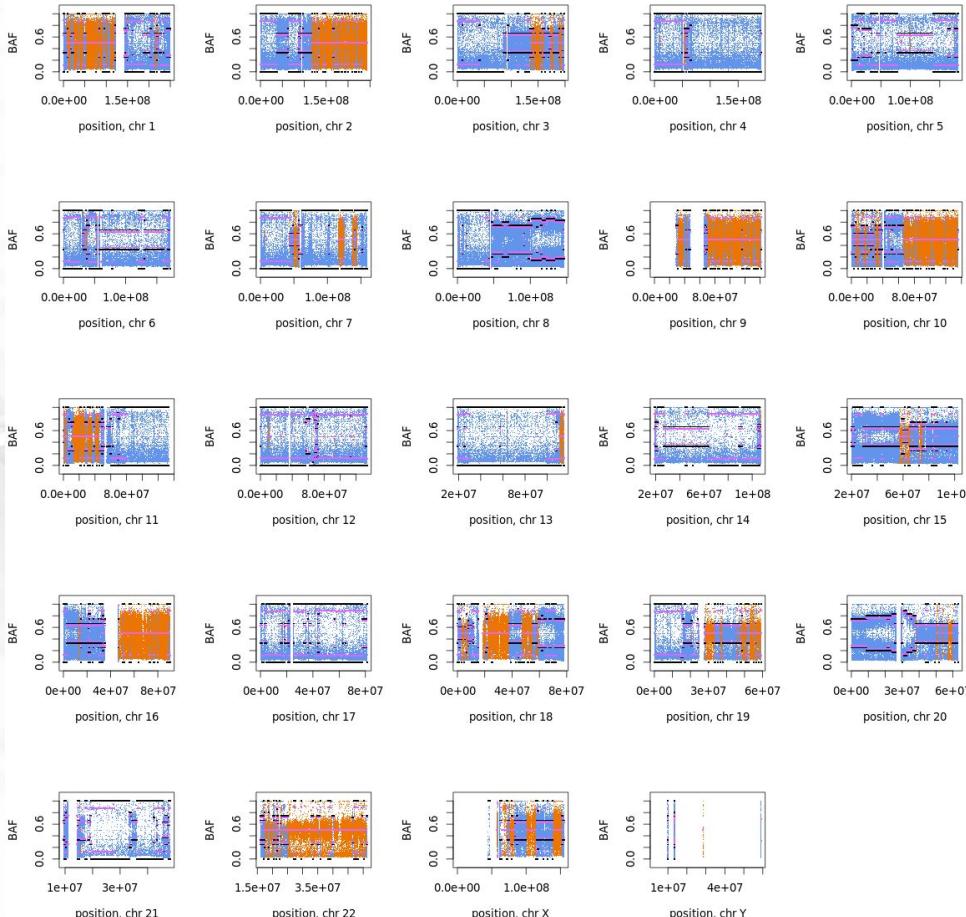
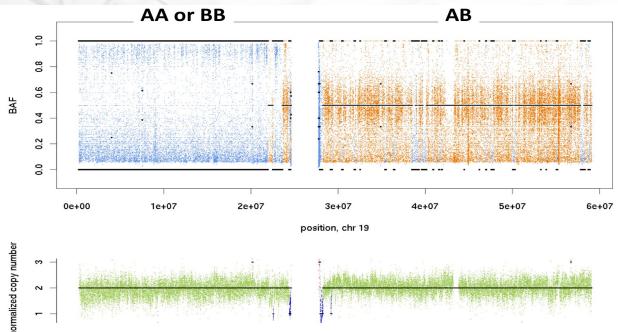
SevenBridges

copy number profiles ->

— normal copy number

— loss — gain





- B allele = alternative variant in dbSNP

0.44 0.5 0.57 0.45 ← B allele frequency (BAF)

$$BAF = \frac{n_B}{n_B + n_A}$$

← Observed nucleotide frequencies

acGatgacgtca**A**atgcttagcgag**G**cacacaaa**T**ac ← Reference genome (A allele)
acCatgacgtca**T**atgcttagcgag**C**cacacaaa**A**ac ← dbSNP (B allele)

A allele: gtcacccatccctc gtgctggtaatcaga
 B allele: gtcacccatccctc gtgctggtaatcaga

$$\text{BAF} = \frac{n_B}{n_B + n_A}$$

(A) Normal cell (FM)

#SNP	F	M	BAF
1	A	-	0
2	B	-	1/2
3	A	-	1/2
4	B	-	1

(B) Copy neutral LOH (FF)

#SNP	F	F	BAF
1	A	-	0
2	B	-	1
3	A	-	0
4	B	-	1

(C) Four copies (FFMM)

#SNP	F	F	M	M	BAF
1	A	A	-	-	0
2	B	B	-	-	1/2
3	A	A	-	-	1/2
4	B	B	-	-	1

(D) Single loss (F)

#SNP	F	BAF
1	A	0
2	B	1
3	A	0
4	B	1

(E) Three copies (FFM)

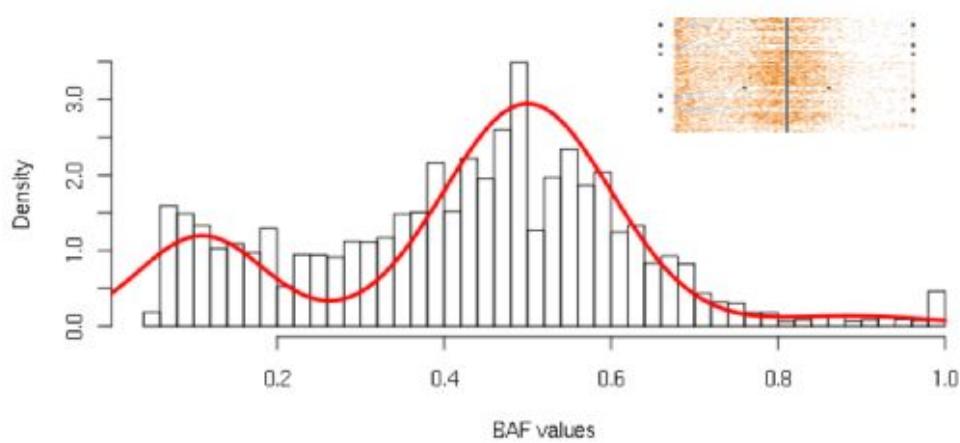
#SNP	F	F	M	BAF
1	A	A	-	0
2	B	B	-	2/3
3	A	A	-	1/3
4	B	B	-	1

(F) Four copies (FFFFM)

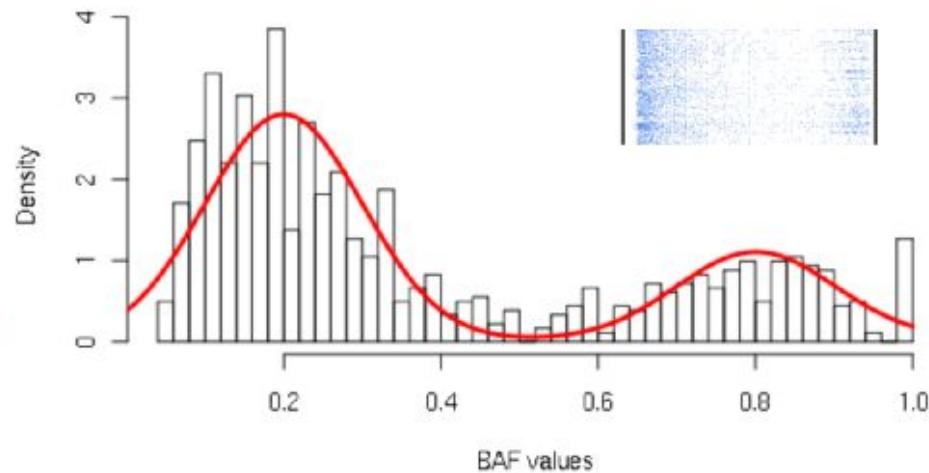
#SNP	F	F	F	M	BAF
1	A	A	A	-	0
2	B	B	B	-	3/4
3	A	A	A	-	1/4
4	B	B	B	-	1

Gaussian Mixture Model fit of BAFs to infer the genotype status of a region

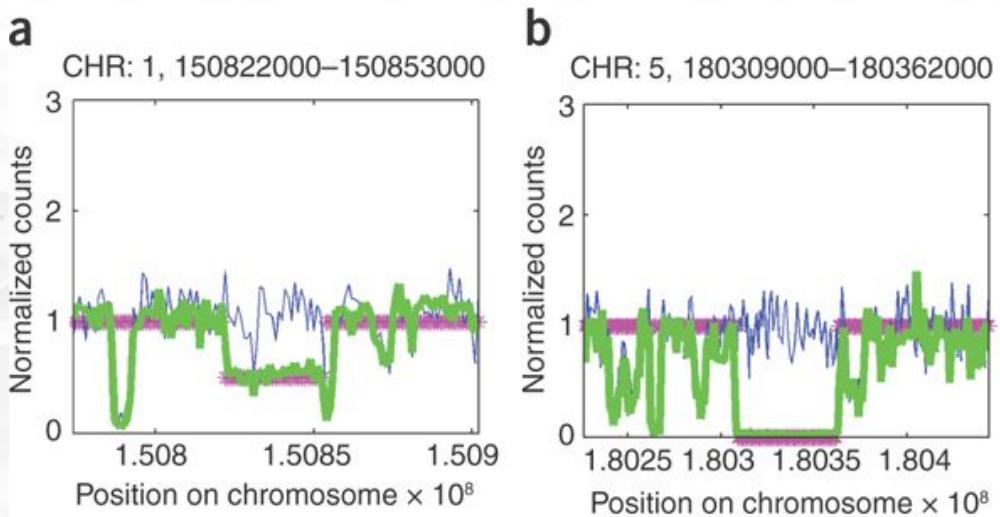
The fit indicates that the genotype = **AB**



The fit indicates that the genotype = **AA/BB**
with **40%** contamination by normal ("AB") cells



- purity measures the contamination of tumor sample by normal stromal cells
- ploidy is the number of sets of chromosomes in the nucleus of a cell, so 2 for normal samples, but normally ranges between 1.5-5 in the tumor samples; in this context can be obtained as the average of copy number of the entire tumor genome
- there are tools that estimate these quantities, but need to make a whole run to get a number that should be returned as an input of the same run (!?)



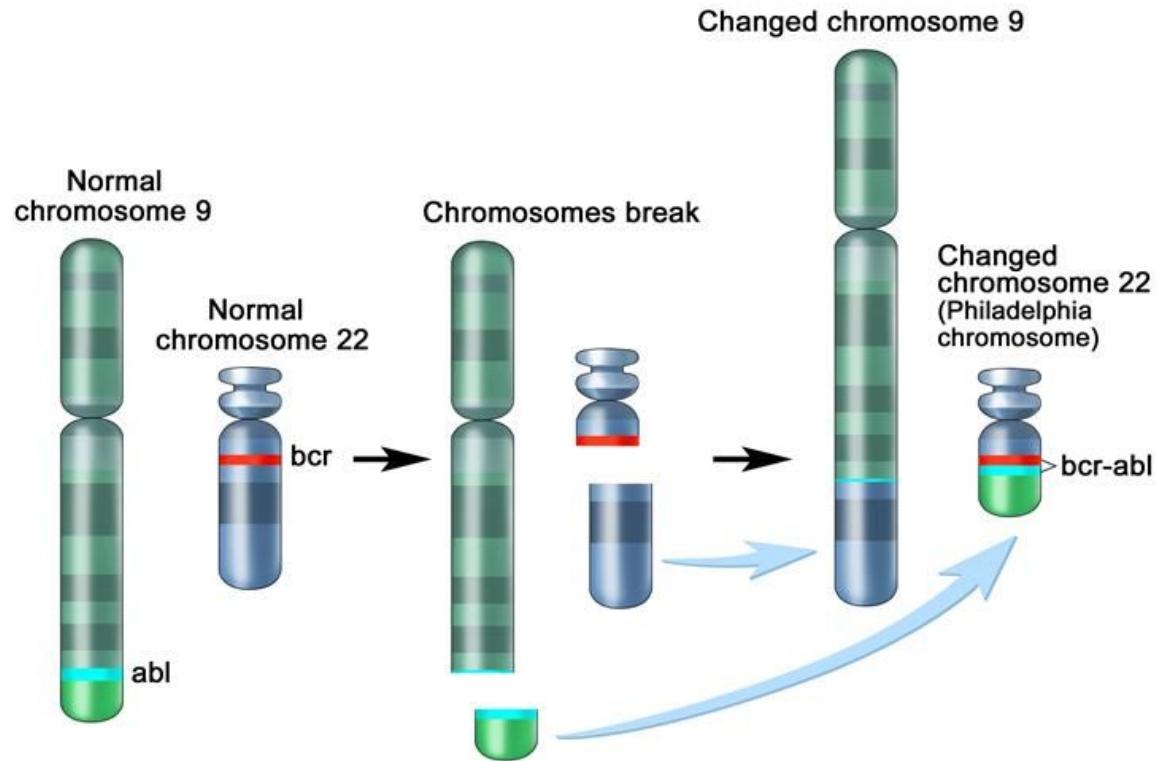
degeneracy: the same copy number can be explained by multiple combinations of tumor ploidy and purity
(e.g., a homozygous deletion combined with 30% tumor purity can also be explained as a heterozygous deletion combined with 60% tumor purity)

BAFs cluster differently in these two cases

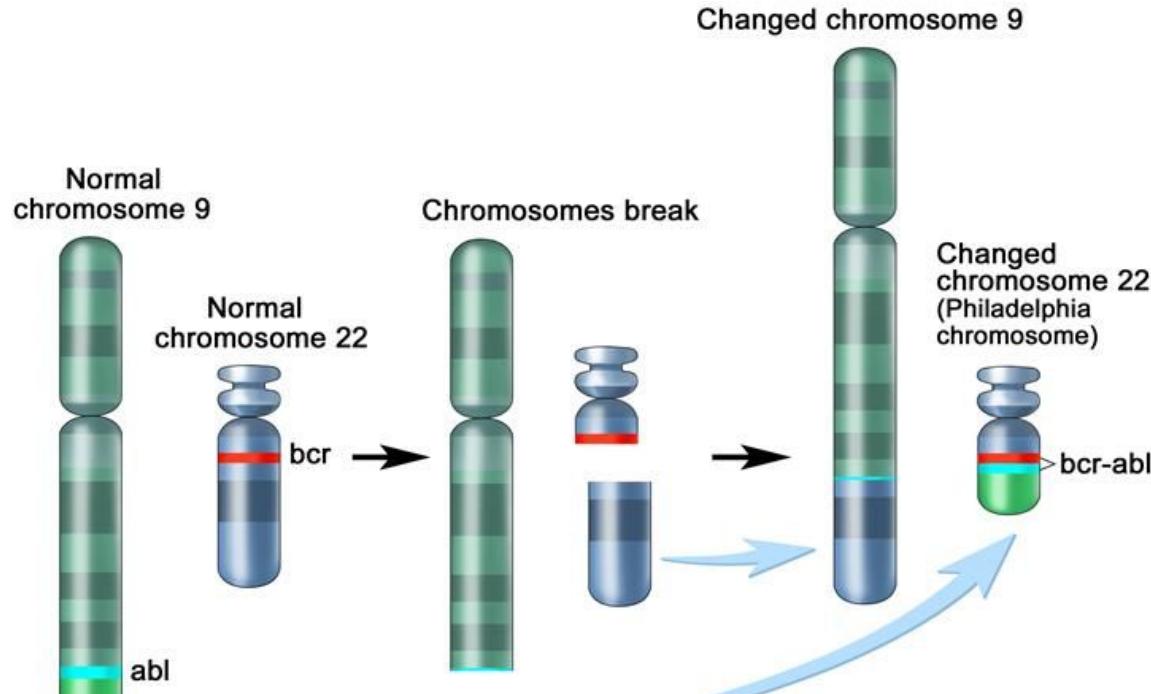
Pushkarev et al. Nature Biotechnology (2009)

Fusion Genes

- chromosomal translocation, interstitial deletion or chromosomal inversion
- fusion genes may lead to a gene product with a new or different function from the two fusion partners
- fusions are highly unusual in normal tissue but are found at higher frequency in cancer tissue
- predict the oncogenic potential of fusion genes in order to identify those fusion sequences with higher probability of being driver of cancer
- finding the exact point of fusion helps in the better characterization of the disease



Biomarkers



The BCR-ABL1 fusion represents an example of interchromosomal rearrangement (commonly found in chronic myelogenous leukemia), while the prostate cancer-associated TMPRSS2-ERG occurs as a result of an intrachromosomal translocation.

Two types of fusion can be distinguished: ones that alter regulatory regions and thus change gene expression without affecting the gene product (breakpoints within the non-coding regions); and those that give rise to a new gene product which is independent of either partner (breakpoints occur within the coding regions of one or both genes)

A

Chromosome 3, gene A



Chromosome 3, gene B

Fusion product
from intergenic splicing**B**

Chromosome 3, gene A



Chromosome 4, gene B

Fusion product
from transgenic splicing

cgap.nci.nih.gov/Chromosomes/Mitelman

Most Visited ▾ Linux Mint Community Forums Blog News ▾

National Cancer Institute U.S. National Institutes of Health | www.cancer.gov

CANCER GENOME ANATOMY PROJECT

CGAP How To Genes Chromosomes Tissues SAGE Genie RNAi Pathways Tools

 Chromosomes

Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer

Searching the Database

The information in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer relates chromosomal aberrations to tumor characteristics, based either on individual cases or associations. All the data have been manually culled from the literature by Felix Mitelman, Berit Johansson, and Fredrik Mertens.

CGAP has developed six web search tools to help you analyze the information within the Mitelman Database:

- The Cases Quick Searcher allows you to query the individual patient cases using the four major fields: aberration, breakpoint, morphology, and topography.
- The Cases Full Searcher permits a more detailed search of the same individual patient cases as above, by including more cytogenetic field choices and adding search fields for patient characteristics and references.
- The Molecular Biology Associations Searcher does not search any of the individual patient cases. It searches studies pertaining to gene rearrangements as a consequence of cytogenetic aberrations.
- The Clinical Associations Searcher does not search any of the individual patient cases. It searches studies pertaining to clinical associations of cytogenetic aberrations and/or gene rearrangements.
- The Recurrent Chromosome Aberrations Searcher provides a way to search for structural and numerical abnormalities that are recurrent, i.e., present in two or more cases with the same morphology and topography.
- The Reference Searcher queries only the references themselves, i.e., the references from the individual cases and the molecular biology and clinical associations.

Database last updated on November 11, 2015
Total number of cases = **66,234**
Total number of gene fusions = **10,180**

Need help! To learn about the Mitelman Database and how to search it, please visit:

- All about the Mitelman Database, which provides background information about the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.
- Mitelman Database Search Help, which contains information on how to use the search tools.
- ISCN Abbreviated Terms and Symbols, which provides a list of terms and symbols used to describe chromosome abnormalities.

Citation of the Database

To cite the use of the Mitelman Database from this CGAP Website in a publication, please quote the following: "Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2015). Mitelman F, Johansson B and Mertens F (Eds.), <http://cgap.nci.nih.gov/Chromosomes/Mitelman>"

The Mitelman Database is supported by the Swedish Cancer Society and the Swedish Childhood Cancer Foundation. The database is updated quarterly in February, May, August, and November.

SevenBridges

cgap.nci.nih.gov/Chromosomes/Mitelman

Most Visited Linux Mint Community Forums Blog News

National Cancer Institute U.S. National Institutes of Health | www.cancer.gov

CANCER GENOME ANATOMY PROJECT

[CGAP How To](#) [Genes](#) [Chromosomes](#) [Tissues](#) [SAGE Genie](#) [RNAi](#) [Pathways](#) [Tools](#)

 Chromosomes

Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer

Searching the Database

The information in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer relates chromosomal aberrations to tumor characteristics, Johansson, and Fredrik Mertens.

CGAP has developed six web search tools to help you analyze the information within the Mitelman Database:

- The Cases Quick Searcher allows you to query the individual patient cases using the four major fields: aberration, breakpoint, morphology, and topography.
- The Cases Full Searcher permits a more detailed search of the same individual patient cases as above, by including more cytogenetic field choices and adding search fields for patient characteristics and references.
- The Molecular Biology Associations Searcher does not search any of the individual patient cases. It searches studies pertaining to gene rearrangements as a consequence of cytogenetic aberrations.
- The Clinical Associations Searcher does not search any of the individual patient cases. It searches studies pertaining to clinical associations of cytogenetic aberrations and/or gene rearrangements.
- The Recurrent Chromosome Aberrations Searcher provides a way to search for structural and numerical abnormalities that are recurrent, i.e., present in two or more cases with the same morphology and topography.
- The Reference Searcher queries only the references themselves, i.e., the references from the individual cases and the molecular biology and clinical associations.

Database last updated on November 11, 2015
 Total number of cases = **66,234**
 Total number of gene fusions = **10,180**

Need help? To learn about the Mitelman Database and how to search it, please visit:

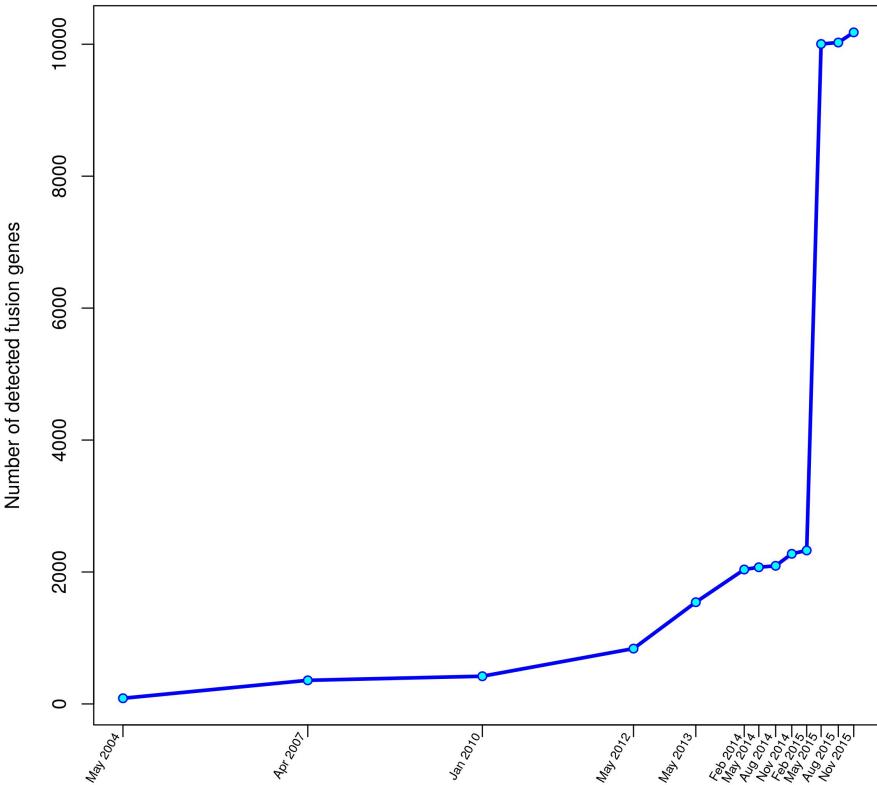
- All about the Mitelman Database, which provides background information about the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.
- Mitelman Database Search Help, which contains information on how to use the search tools.
- ISCN Abbreviated Terms and Symbols, which provides a list of terms and symbols used to describe chromosome abnormalities.

Citation of the Database

To cite the use of the Mitelman Database from this CGAP Website in a publication, please quote the following: "Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer".

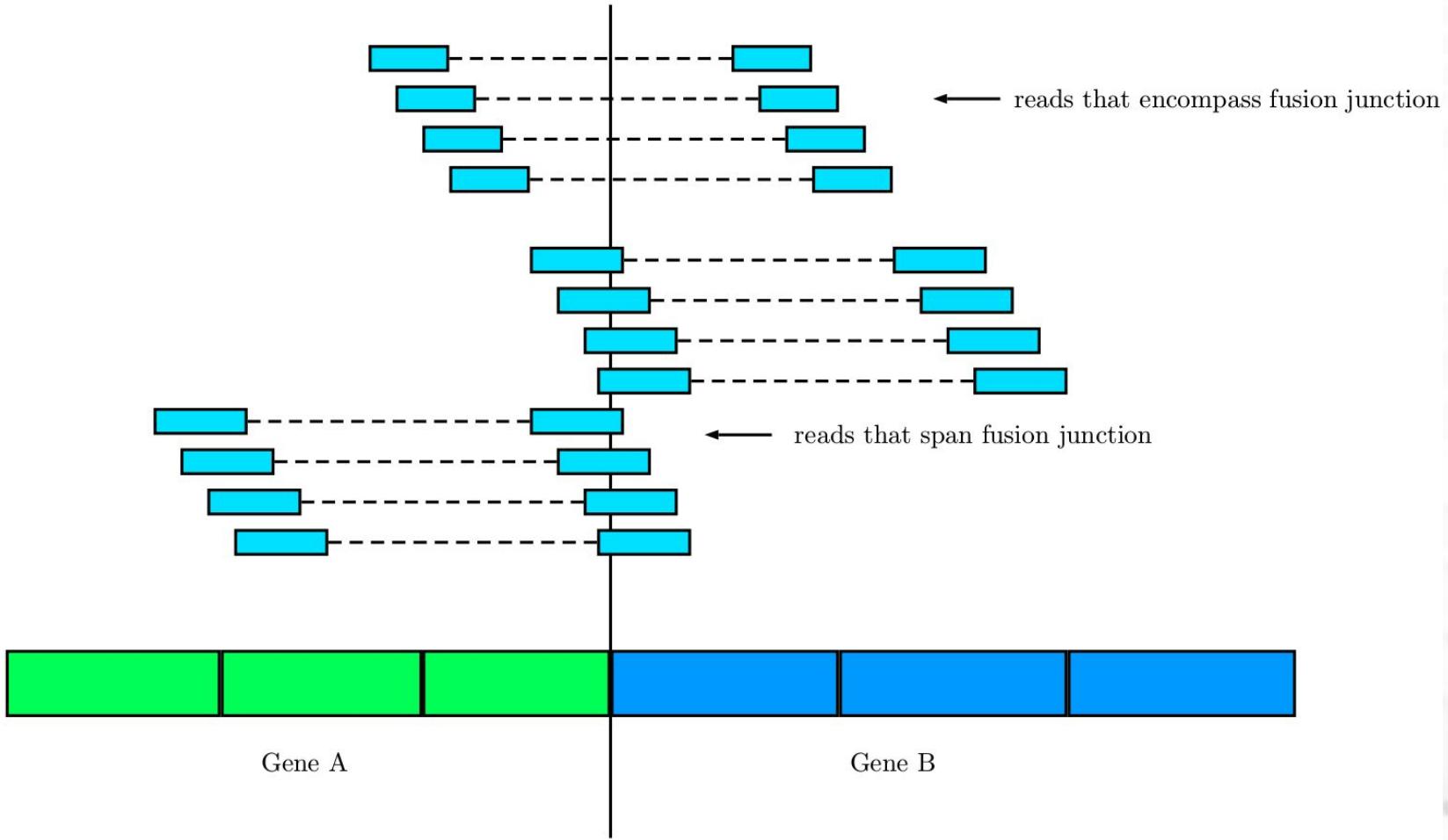
The Mitelman Database is supported by the Swedish Cancer Society and the Swedish Childhood Cancer Foundation. The database is updated quarterly.

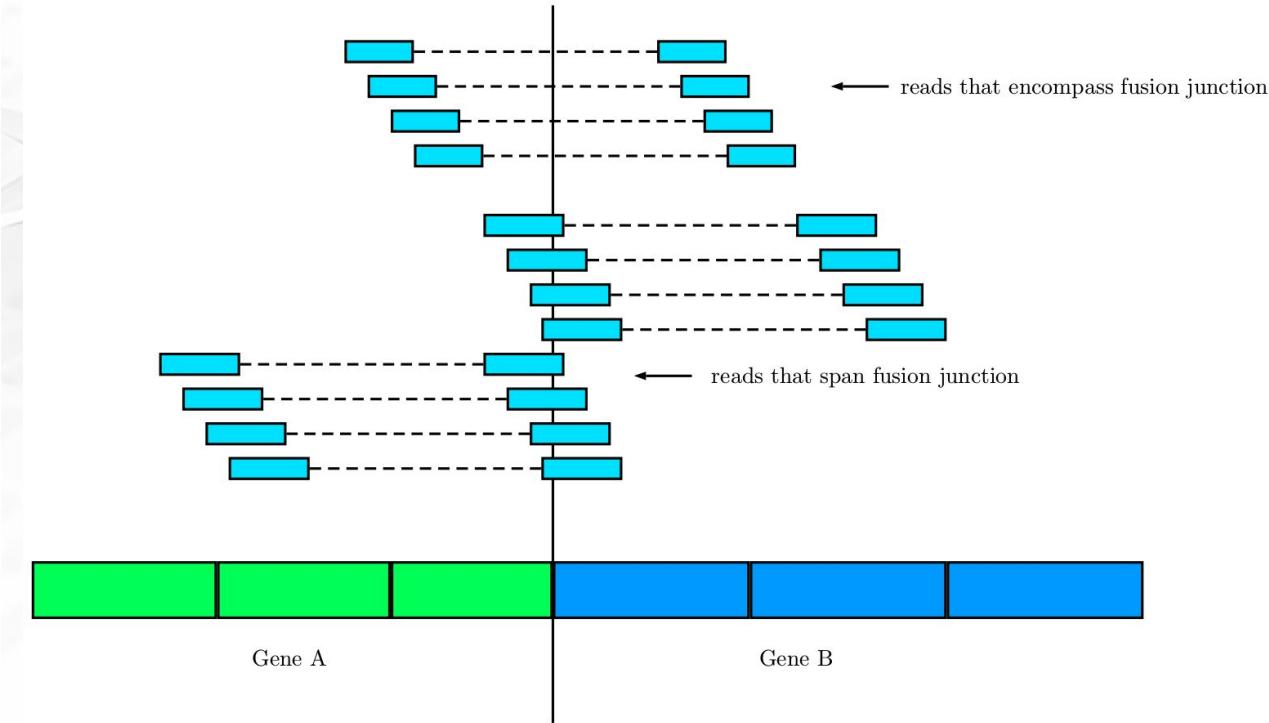
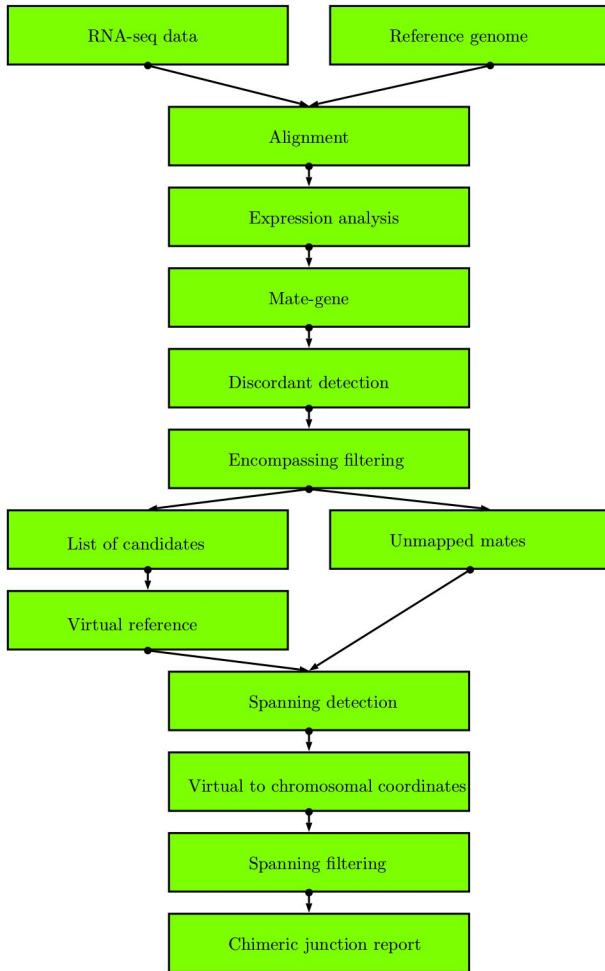
Fusion detections over time



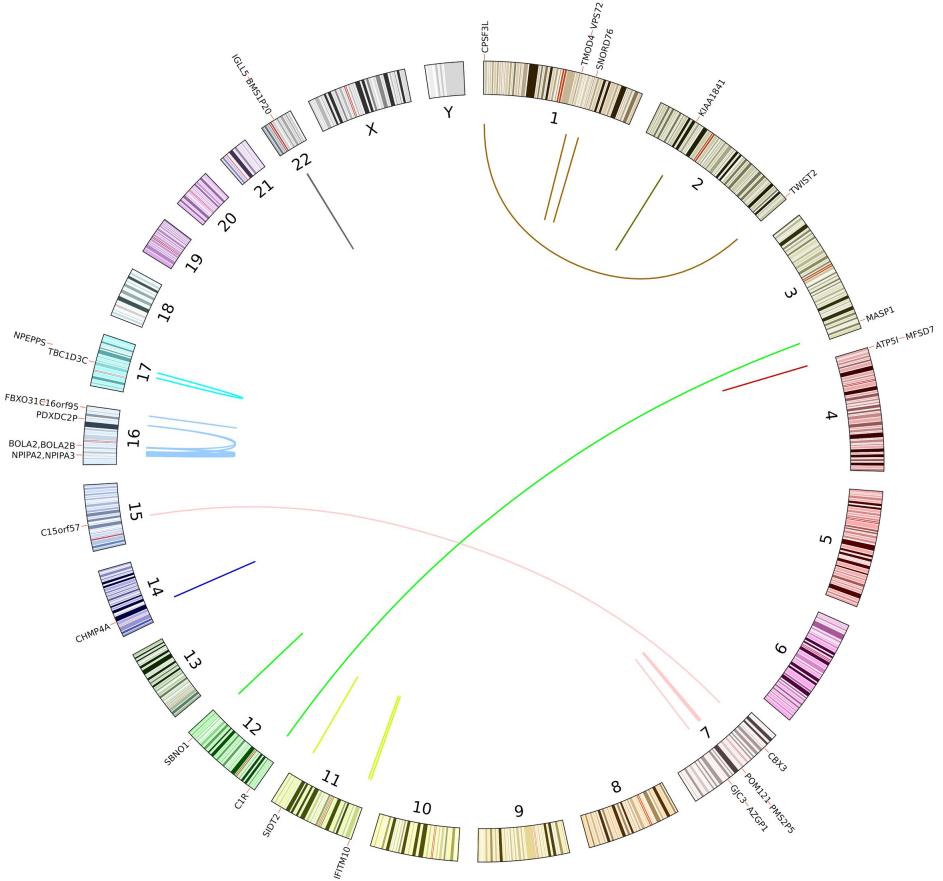
SevenBridges

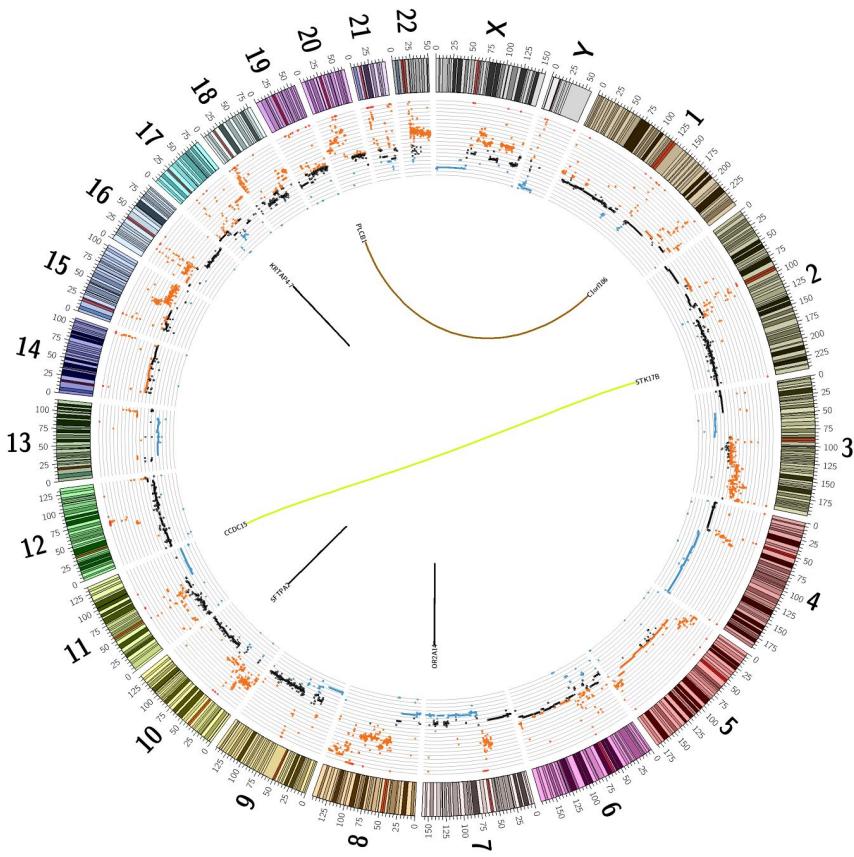
Application	Tool Name	Source Code	Year	Reference
Detection	STAR	https://code.google.com/p/rna-star/	2012	[18]
Detection/Identification	Bellerophontes	http://eda.polito.it/bellerophontes/	2012	[19]
Detection/Identification	BreakFusion	http://bioinformatics.mdanderson.org/main/BreakFusion	2012	[20]
Detection/Identification	ChimeraScan	http://chimerascan.googlecode.com	2011	[21]
Detection/Identification	deFuse	http://sourceforge.net/projects/defuse/	2011	[22]
Detection/Identification	Dissect	http://dissect-trans.sourceforge.net/Home	2012	[23]
Detection/Identification	FusionCatcher	https://code.google.com/p/fusioncatcher/	2014	[24]
Detection/Identification	FusionFinder	http://code.google.com/p/fusioncatcher/	2012	[25]
Detection/Identification	FusionHunter	http://bioen-comphbo.bioen.illinois.edu/FusionHunter/	2011	[26]
Detection/Identification	FusionMap	http://www.arrayserver.com/wiki/index.php?title=FusionMap	2011	[27]
Detection/Identification	FusionQ	https://sites.google.com/site/fusionql/home/	2013	[28]
Detection/Identification	FusionSeq	http://archive.gersteinlab.org/proj/naseq/fusionseq/	2010	[29]
Detection/Identification	JAFFA	https://code.google.com/p/jaffa-project/	2015	[30]
Detection/Identification	MOJO	https://github.com/cband/MOJO	2014	[31]
Detection/Identification	shortFuse	https://bitbucket.org/mckinse/shortfuse/overview	2011	[32]
Detection/Identification	SnowShoes-FTD	http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm	2011	[33]
Detection/Identification	SOAPfuse	http://soap.genomics.org.cn/soapfuse.html	2013	[34]
Detection/Identification	SOAPfusion	http://soap.genomics.org.cn/SOAPfusion.html	2013	[35]
Detection/Identification	TopHat-Fusion	http://tophat-fusion.sourceforge.net/	2011	[36]
Detection/Identification	TRUP	https://github.com/tuping/TRUP	2015	[37]
Detection/Identification - focus on viral events	ViraFusionSeq	http://sourceforge.net/projects/virafusionseq/	2013	[38]
Detection/Identification combined RNA & DNA	Comrad	https://code.google.com/p/comrad/	2011	[39]
Detection/Identification combined RNA & DNA	nFuse	https://code.google.com/p/nfuse/	2012	[40]
Detection/Identification/Visualization	FusionAnalyzer	http://www.ngsbicocca.org/html/fusion_analyser.html	2012	[41]
Filtering & Prioritization	Chimera	http://www.bioconductor.org/packages/release/bioc/html/chimera.html	2014	[42]
Filtering & Prioritization	Oncofuse	http://www.unav.es/genetica/oncofuse.html	2013	[43]
Filtering & Prioritization	Pegasus	http://sourceforge.net/projects/pegasus-fus/	2014	[44]





Circos representation



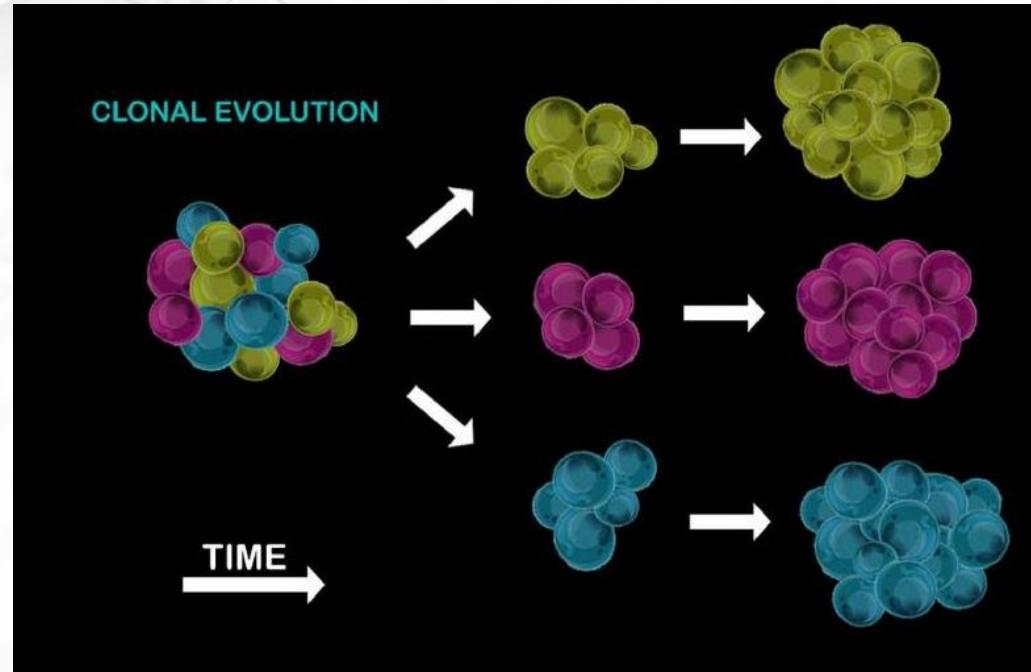


SevenBridges

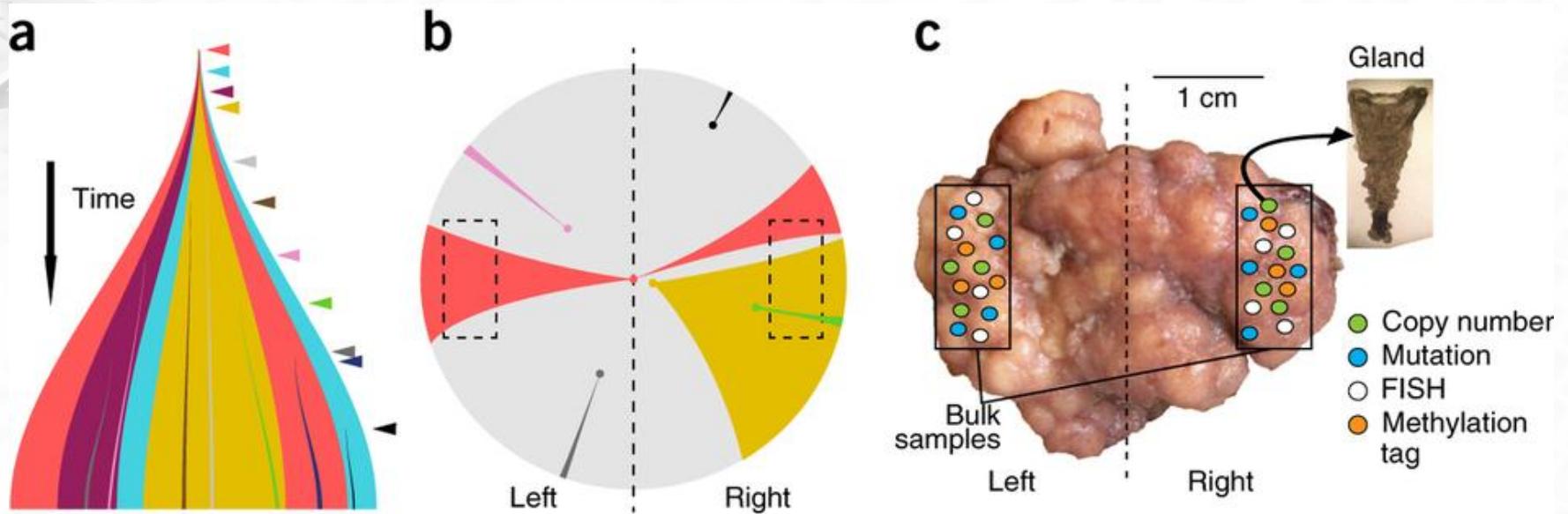
How do mutations accumulate over time to cause a cancer

traditional model: clonal expansion

- sequential stepwise accumulation of alterations
- acquisition of new somatic mutations followed by selective sweeps and large clonal expansion

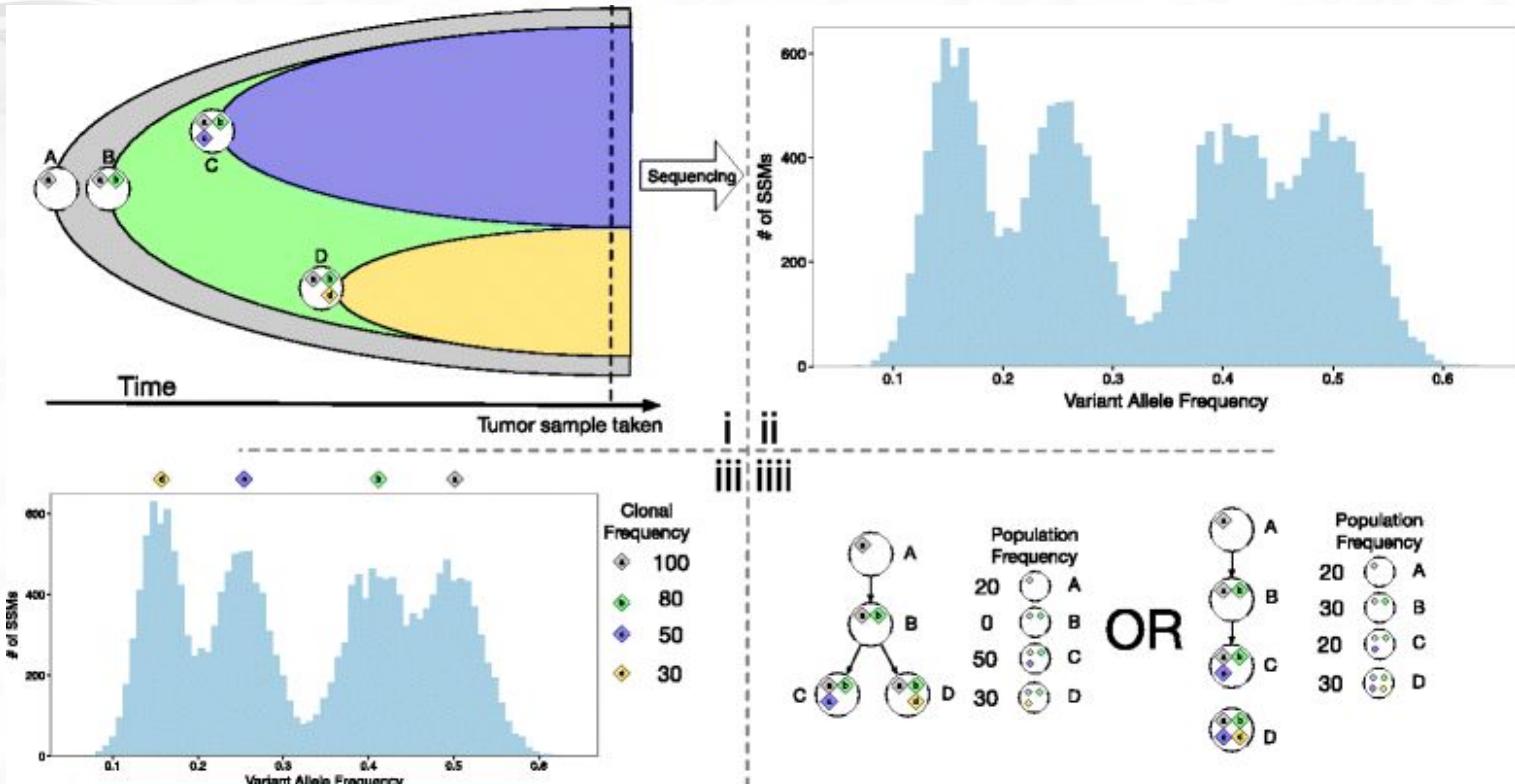


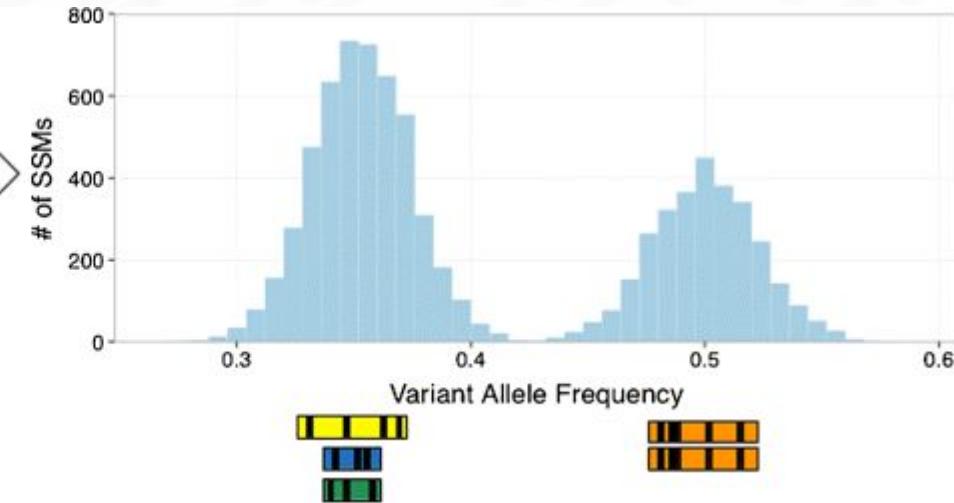
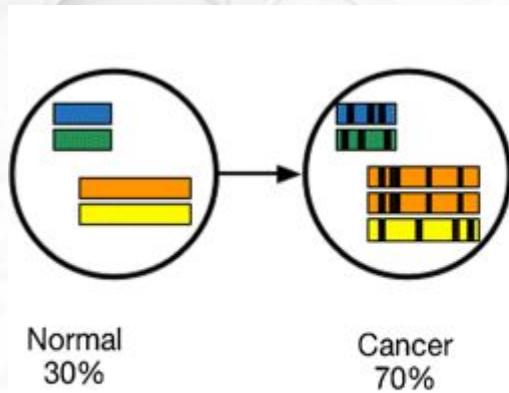
- observed extensive intratumoral heterogeneity (ITH) -- exhibits as distinct cellular morphology, gene expression, proliferation, metastatic potential,... not compatible with traditional model
- absence of selective sweeps (due to rapidly expanding population and spatial constraints)
- private (subclonal) alterations, e.g. CNAs and point mutations, that occur early should be pervasive in the final tumor

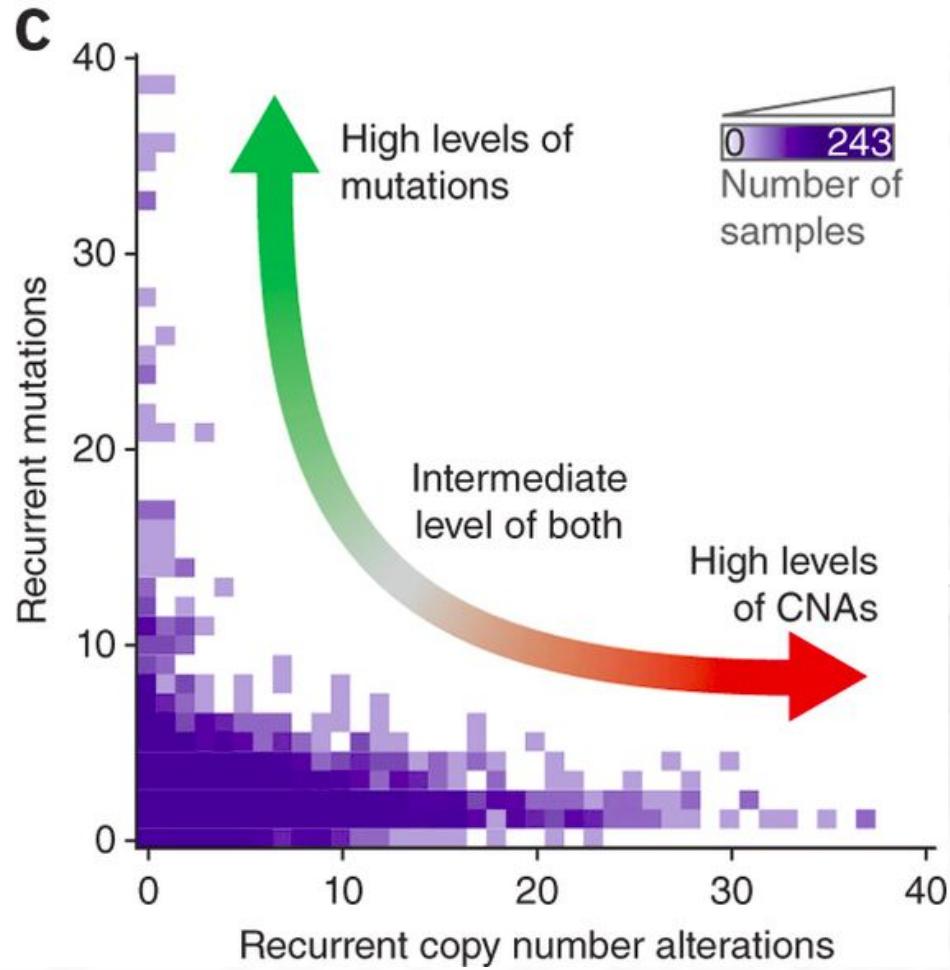


- genetic variegation (subclone mixing in the early tumor followed by scattering)
- detectable private alterations occur early; late alterations only present in small regions (undetectable)

Sottoriva et al. Nature Genetics (2015)







Ciriello et al. Nature Genetics (2013)

* The Cancer Genome Atlas (TCGA)

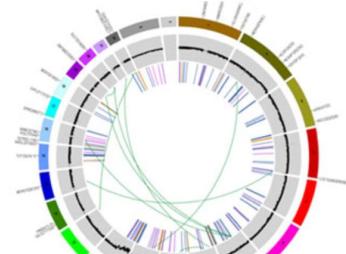
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Launch Data Portal | Contact Us | For the Media

Search

Home About Cancer Genomics Cancers Selected for Study Research Highlights Publications News and Events About TCGA

Program Overview
Explore how The Cancer Genome Atlas works, the components of the TCGA Research Network and TCGA's place in the cancer genomics field in the Program Overview.
[Learn More ▶](#)



 Profile of Researcher Dr. Hui Shen  TCGA's Study of Prostate Cancer  Cancers Selected for Study  About TCGA

Research Briefs 

October 2015
Researchers Use TCGA Data for First Pan-Cancer Analyses of RNA-Editing

September 2015
DNA Methylation Inhibitor Triggers Anti-Viral Immune Response in Cancer

April 2015
Using TCGA Data to Find a Novel Target for Triple-Negative Breast Cancer

News and Announcements 

November 05, 2015
TCGA study identifies seven distinct subtypes of prostate cancer
A comprehensive analysis of 333 prostate cancers identified key genetic alterations that may help improve classification and treatment of the disease.

November 04, 2015
Improved understanding of the genetic drivers of papillary renal cell carcinoma
A comprehensive genomic analysis of 161 tumors from people with papillary renal cell carcinoma (PRCC) – the second most common form of kidney

Launch Data Portal 

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA.

Questions About Cancer

Visit www.cancer.gov
Call 1-800-4-CANCER
Use [LiveHelp Online Chat](#)

Multimedia Library

 Images
 Videos and Animations
 Podcasts
 Interactive

Stay Connected

 Sign up for email updates
 RSS newsfeeds

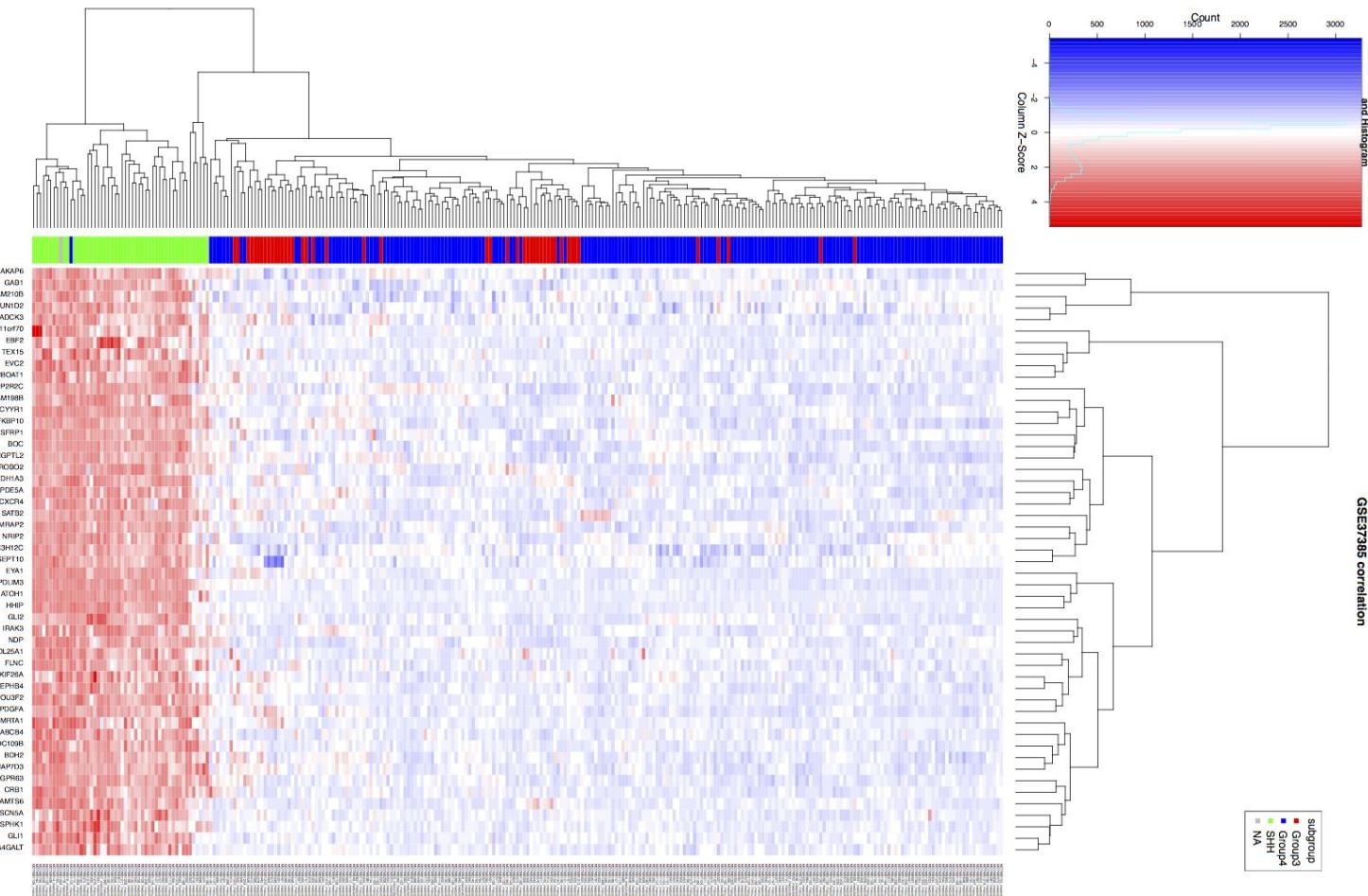
Integrative view of cancer

- Genomics (WGS, WEX)
- RNA--Seq
- Epigenomics/ChIP--Seq
- DNA--Methylation
- Single--cell
- Proteomics
- Metabolomics

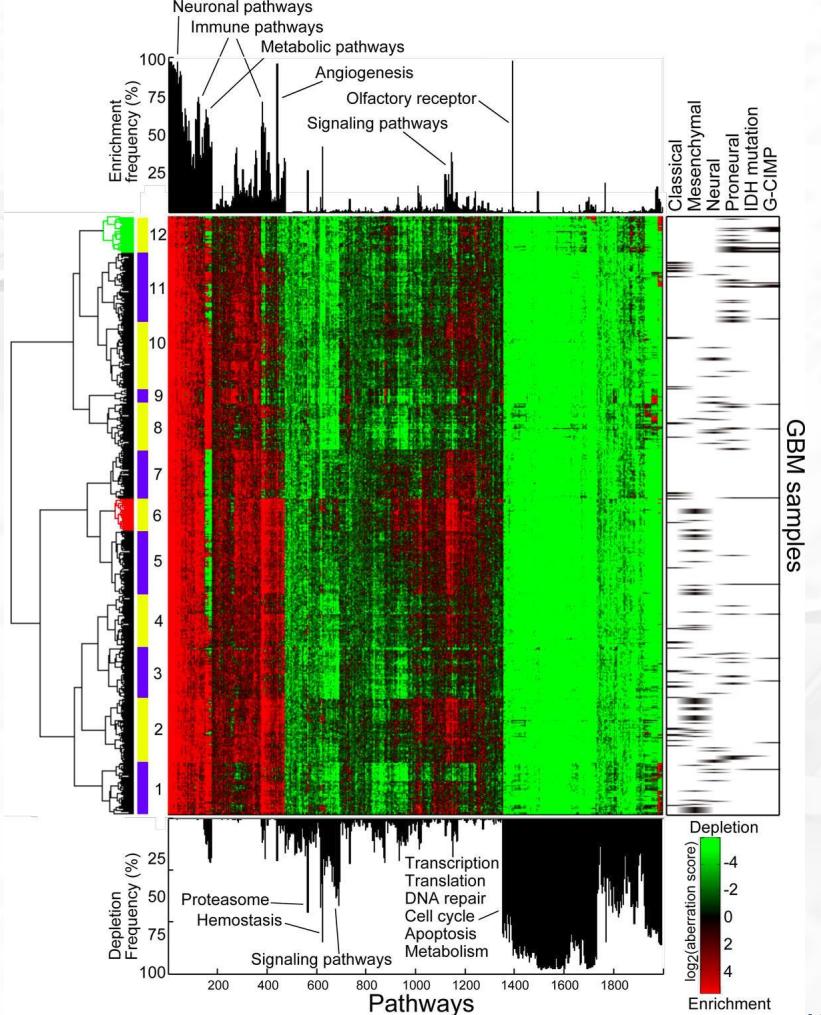
Integrative view of cancer

- Genomics (WGS, WEX)
- RNA--Seq
- Epigenomics/ChIP--Seq
- DNA--Methylation
- Single--cell
- Proteomics
- Metabolomics

Combine all of these to identify the driving causes for each type of tumor (evaluate frequency of gene mutations in known genes)



SevenBridges



SevenBridges

Networks & Pathways

Comparison and combination of these type of complex data



Wu et al. (2010)

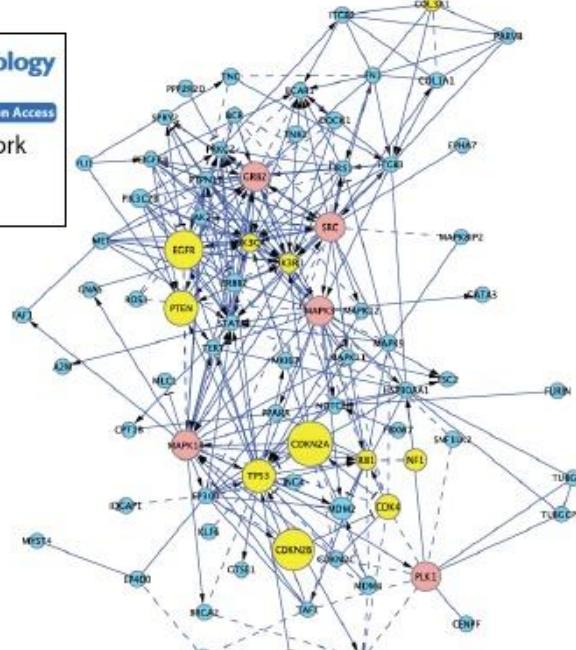
Wu et al. Genome Biology 2010, 11:853
<http://genomebiology.com/2010/11/5/853>

RESEARCH Open Access

A human functional protein interaction network and its application to cancer data analysis

Guoming Wu¹, Xin Feng¹ and Lincoln Stein^{1,2}

Subnetwork derived from
The Cancer Genome Atlas (TCGA)
of somatic mutation data set:
77 cancer genes
and
5 linker genes



Dr J De Las Rivas - 2011 36

Future research

- Novel sequencing techniques
- More robust pathway signaling analysis
- Prioritization of variants (non-coding regions included)
- Learning algorithms (Machine Learning, Random forest, Multiple Kernel Estimators, Similarity Network Fusion),...
- Study effect of drugs on variants (clinical trials)

ANY
QUESTIONS?
?