

# Lecture 20. Cancer Genetics & Genomics

Michael Schatz

April 20, 2017

JHU 600.649:Applied Comparative Genomics



# Preliminary Project Report

## Due Sunday (4/23) at 11:59pm

The screenshot shows a GitHub repository page for `schatzlab/appliedgenomics`. The repository has 5 stars, 15 forks, and 1 issue. The `Code` tab is selected, showing the file `preliminaryreport.md`. The file was created by `mschatz` 17 minutes ago. It contains 24 lines of code (17 sloc) and is 1.16 KB in size. The file content is as follows:

```
Preliminary Project Report

Assignment Date: April 20, 2017
Due Date: Sunday, April 23, 2017 @ 11:59pm

Each team should email a PDF of your preliminary project proposal (1 to 2 pages) to "jhuappliedgenomics@gmail.com" by 11:59pm on Sunday April 23.

The preliminary report should have:

• Title of your project
• List of team members and email addresses
• 1 paragraph abstract summarizing the project
• 1 paragraph of Introduction
• 1 paragraph of Methods that you are using
• 1 paragraph of Results, describing the data evaluated and any any preliminary results
• 1 paragraph of Discussion (what you have seen or expect to see)
• 1 figure showing a preliminary result
• References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission\_online

Later, you will present your project in class during the week of May 5. You will also submit your final written report (5-7 pages) of your project by May 17

Please use Piazza if you have any general questions!
```

At the bottom of the page, there are links to GitHub's Terms, Privacy, Security, Status, Help, Contact GitHub, API, Training, Shop, Blog, and About.

# Assignment 2

The screenshot shows a GitHub repository page for 'appliedgenomics/assignments/assignment2/README.md'. The page includes a header with tabs for Code, Issues, Pull requests, Projects, Pulse, Graphs, and Settings. A sidebar on the left shows a file tree with 'README.md' as the root. The main content area displays the README file's content, which includes sections for Assignment Overview, Question 1, Question 2, and Question 3, along with download links and hints.

Inbox - michael.schatz@gmail.com | Google Calendar - Month of March | appliedgenomics/README.md | PLOS Collections: Article collection

GitHub, Inc. [US] | https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment2/README.md

JHUMail Daily schatzlab SL cshl jhu Media edit Rm Cookies GoPerf

schatzlab / appliedgenomics

Michael

Branch: master | appliedgenomics / assignments / assignment2 / README.md

mschatz Update README.md | caeb8dd on Mar 17

1 contributor

233 lines (150 sloc) | 11.5 KB

[Raw](#) [Blame](#) [History](#) [Edit](#) [Delete](#)

## Assignment 2: Variant Analysis

Assignment Date: Tuesday, March 7, 2017  
Due Date: Thursday, March 16, 2017 @ 11:59pm

### Assignment Overview

In this assignment, you will identify variants in a human genome and then analyze the properties for them. Make sure to show your work in your writeup! As before, any questions about the assignment should be posted to [Piazza](#). Some of the tools you will need to use only run in a linux environment. If you do not have access to a linux machine, download and install a virtual machine following the directions here: <https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md>

#### Question 1. Gene Annotation Preliminaries [10 pts]

Download the annotation of build 38 of the human genome from here:  
[ftp://ftp.ensembl.org/pub/release-87/gtf/homo\\_sapiens/Homo\\_sapiens.GRCh38.87.gtf.gz](ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh38.87.gtf.gz)

- Question 1a. How many GTF data lines are in this file? [Hint: The first few lines in the file beginning with "#" are so-called "header" lines describing things like the creation date, the genome version (more on that later in the course), etc. Header lines should not be counted as data lines.]
- Question 1b. How many annotated protein coding genes are on each chromosome of the human genome? [Hint: Protein coding genes will contain the following text: gene\_biotype "protein\_coding"]
- Question 1c. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes? [Hint: use the same genes as those identified in 1b]
- Question 1d. What is the maximum, minimum, mean, and standard deviation in the number of exons for protein coding genes? [Hint: you should separately consider each isoform for each protein coding gene]

#### Question 2. Genome Sequence Analysis [10 pts]

Download chromosome 22 from build 38 of the human genome from here:  
<http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz>

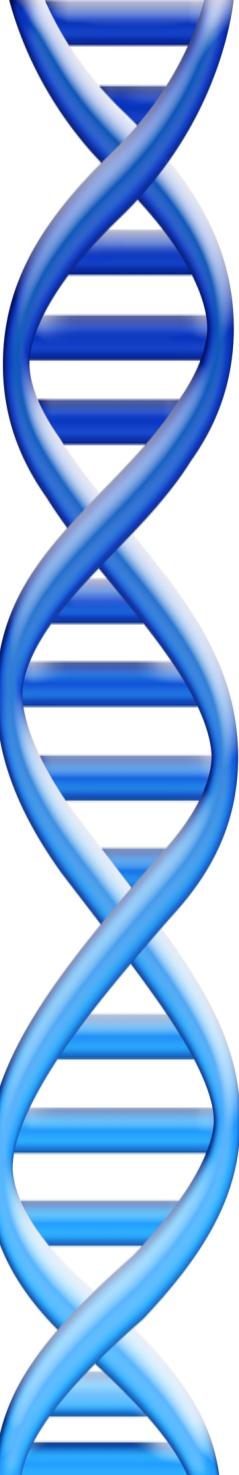
- Question 2a. What is the length of chromosome 22? [Hint: You should include Ns in the length]
- Question 2b. How many Ns are in chromosome 22? What is the GC content? [Hint: You should exclude Ns when computing GC content]
- Question 2c. Restriction enzymes cleave DNA molecules at or near a specific sequence of bases. For example, the HindIII enzyme cuts at the "/" in either this motif: 5'-A/AGCTT-3' or its reverse complement, 3'-TTCGA/A-5'. How many perfectly matching HindIII restriction enzyme cut sites are there on chr22?
- Question 2d. How many HindIII cut sites are there on chr22, assuming that a mutant form of HindIII will tolerate a mismatch in the second position? Think about ways in which you could best test for all the possible DNA combinations. [Hint: There are many valid approaches]

#### Question 3. Small Variant Analysis [10 pts]

Download the read set from here:  
<http://schatzlab.cshl.edu/data/teaching/sample.tgz>

For this question, you may find this tutorial helpful:  
<http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html>

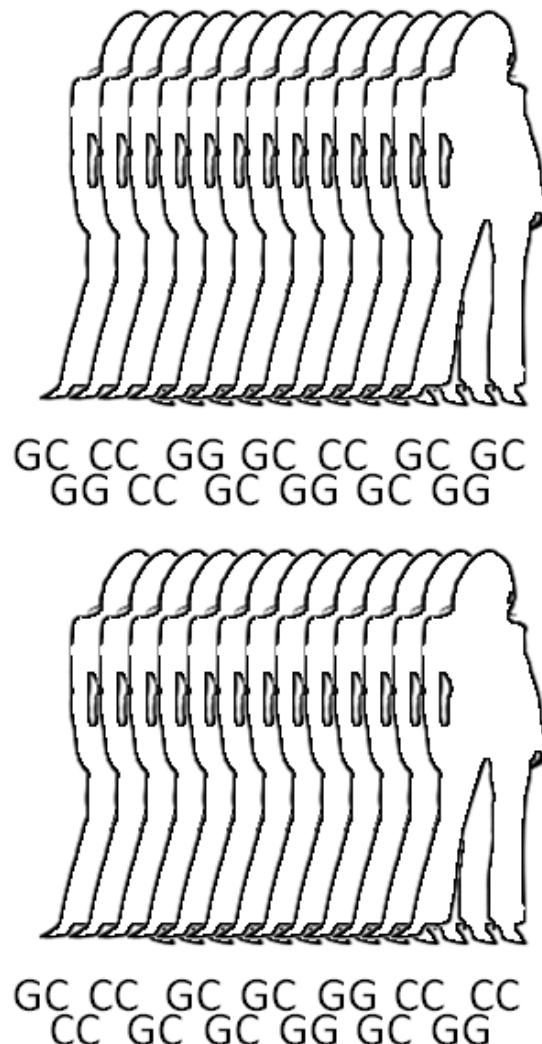
- Question 3a. How many single nucleotide and indel variants does the sample have? [Hint: Align reads using `bwa mem`, identify variants using `freebayes`, filter using `vcffilter -f "QUAL > 20"`, and summarize using `vcfstats`; Make sure to set the read group when running `bwa mem`: `bwa mem -R "@RG\tID:id\tSM:sample\tLB:lib" chr22.fa pair.1.fq pair.2.fq`]



# Part 2:

## Post-genome Inherited Diseases

# Genome Wide Association (GWAS)



*SNP1*

**Cases**

Count of G:  
2104 of 4000

Frequency of G:  
52.6%

**Controls**

Count of G:  
2676 of 6000

Frequency of G:  
44.6%

**P-value:**

$5.0 \cdot 10^{-15}$

*SNP2*

**Cases**

Count of G:  
1648 of 4000

Frequency of G:  
41.2%

**Controls**

Count of G:  
2532 of 6000

Frequency of G:  
42.2%

**P-value:**

0.33

*SNP ...*

*Repeat for all  
SNPs*

With a (much) larger population, this might be a significant difference:  
 $25320/60000 \Rightarrow p = 5e-7$

Chi-squared or similar test

# Adventures in Overfitting

Inbox - michael.schatz@gmail.com Google Calendar - Month of M schatzlab/appliedgenomics: JI PLOS Collections: Article colle Adventures in Overfitting! Michael

https://raw.githubusercontent.com/schatzlab/appliedgenomics/master/lectures/20.overfitting/overfitting.html

20 JHUMail Daily Y f G schatzlab P SL cshl jhu Media edit Rm Cookies GoPerf Other Bookmarks

1 Generate some noisy data  
2 Fitting Lines to (multiple) points  
3 Linear and Polynomial Regression

## Adventures in Overfitting!

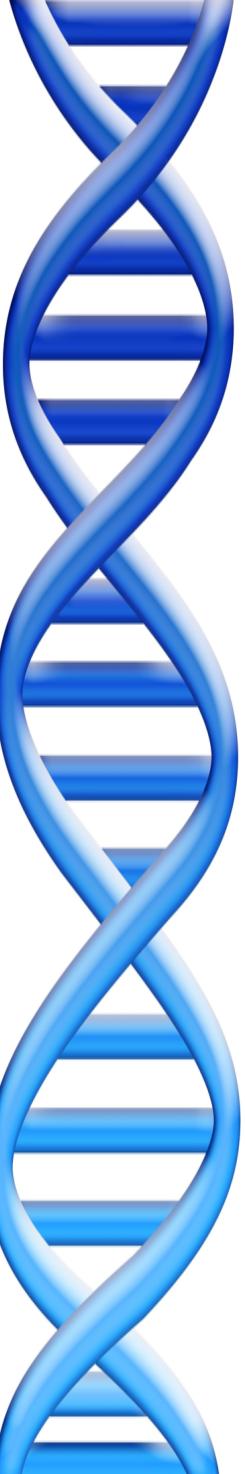
Michael Schatz ([mschatz@jhu.edu](mailto:mschatz@jhu.edu))  
April 20, 2017

### 1 Generate some noisy data

#### 1.1 Generate the true relationship

```
## Initialize Random Number Generator so plots are consistent
set.seed(20)

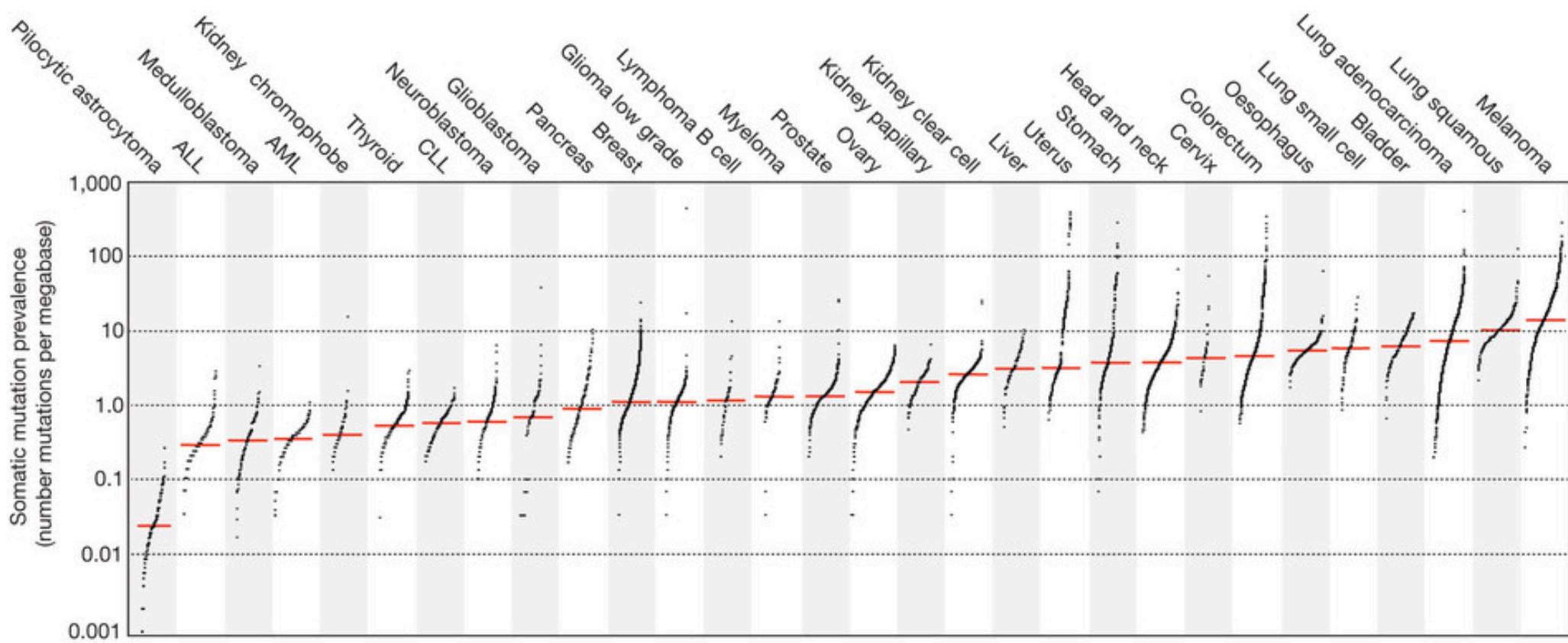
## Now compute points along a cubic function
q <- seq(from=-15, to=15, by=.5)
y <- q*q*q-3*q*q-144*q+432
plot(q,y,col='firebrick1',lwd=3, typ="l", ylim=c(min(y)-200, max(y)+200))
```



# Part 3:

# Cancer Genetics

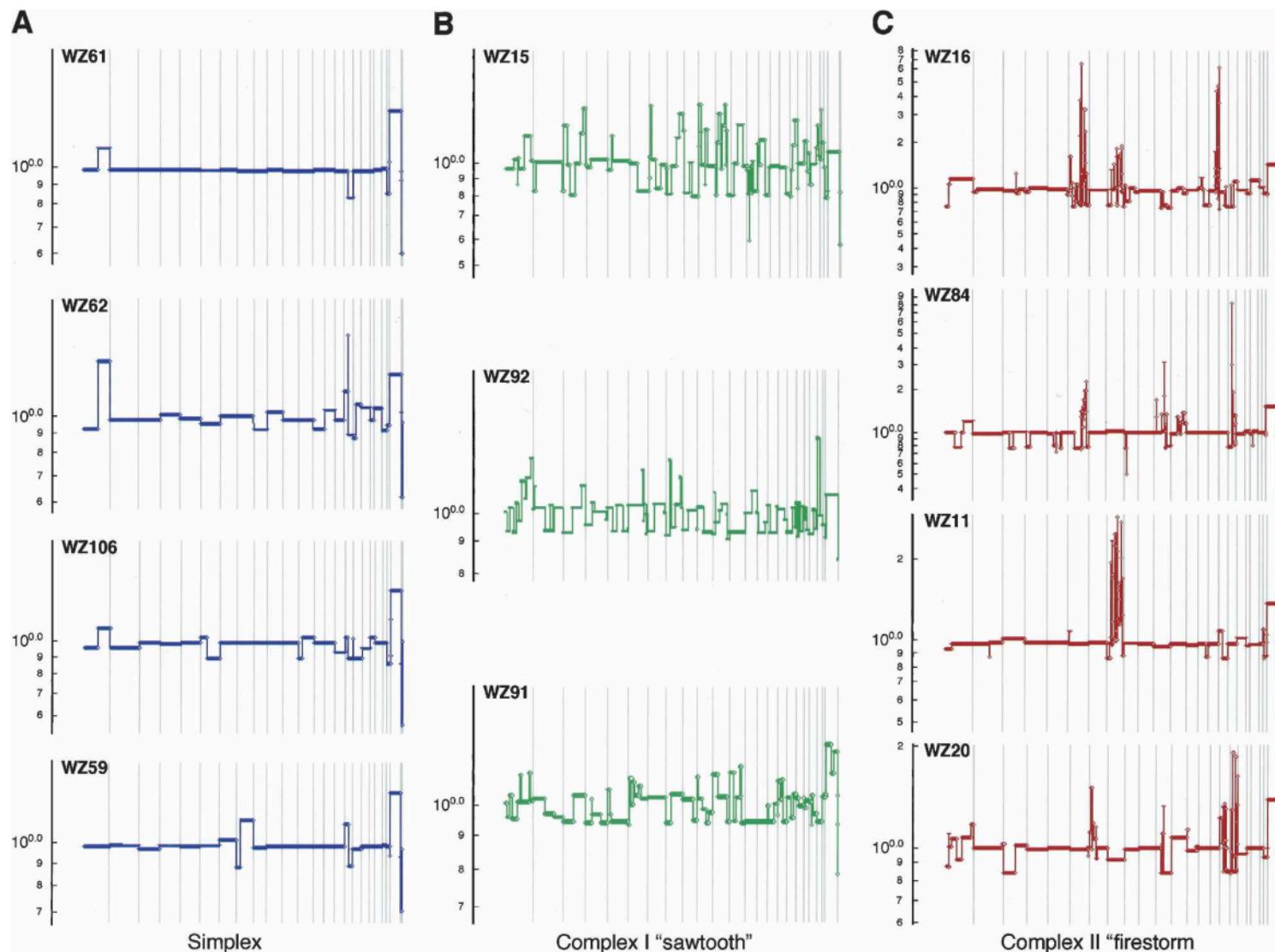
# Somatic Mutations In Cancer



**Signatures of mutational processes in human cancer**

Alexandrov et al (2013) *Nature*. doi:10.1038/nature12477

# A firestorm in cancer



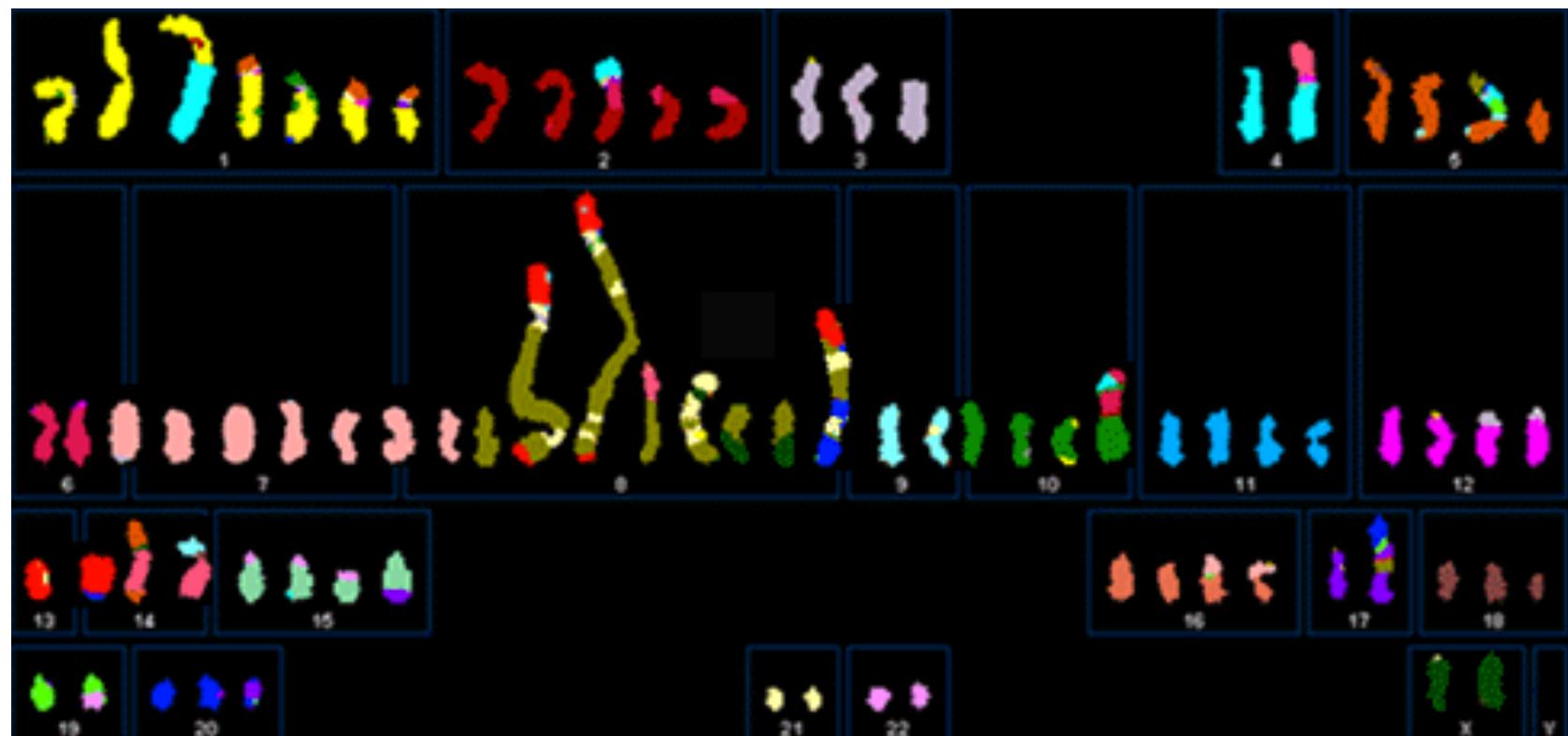
**Figure 2.** Major types of tumor genomic profiles. Segmentation profiles for individual tumors representing each category: (A) simplex; (B) complex type I or sawtooth; (C) complex type II or firestorm. Scored events consist of a minimum of six consecutive probes in the same state. The y-axis displays the geometric mean value of two experiments on a log scale. Note that the scale of the amplifications in C is compressed relative to A and B owing to the high levels of amplification in firestorms. Chromosomes 1–22 plus X and Y are displayed in order from left to right according to probe position.

**Novel patterns of genome rearrangement and their association with survival in breast cancer**

Hicks et al (2006) Genome Research. Doi: 10.1101/gr.5460106

# SK-BR-3

Most commonly used Her2-amplified breast cancer cell line

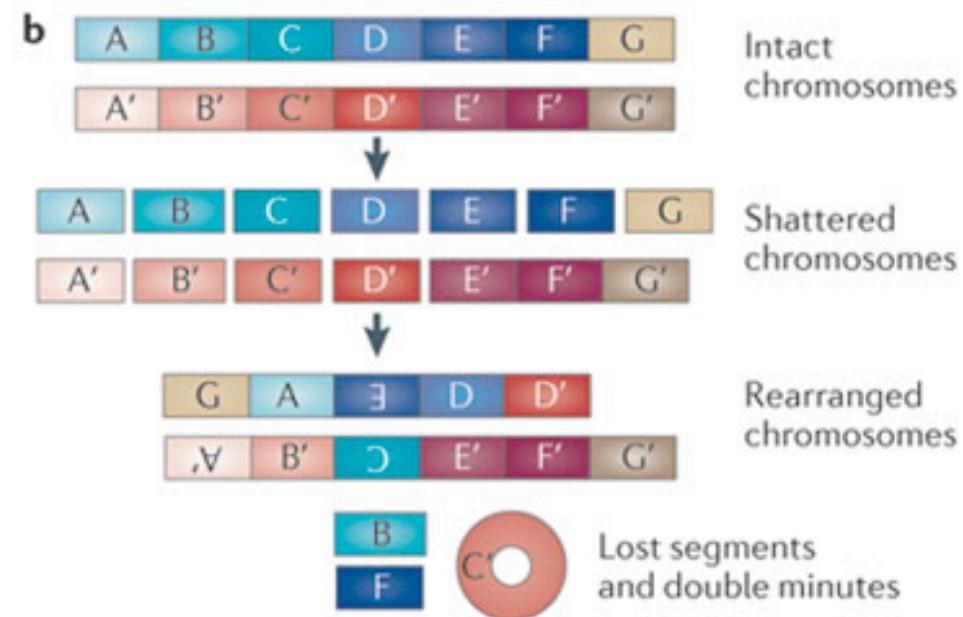
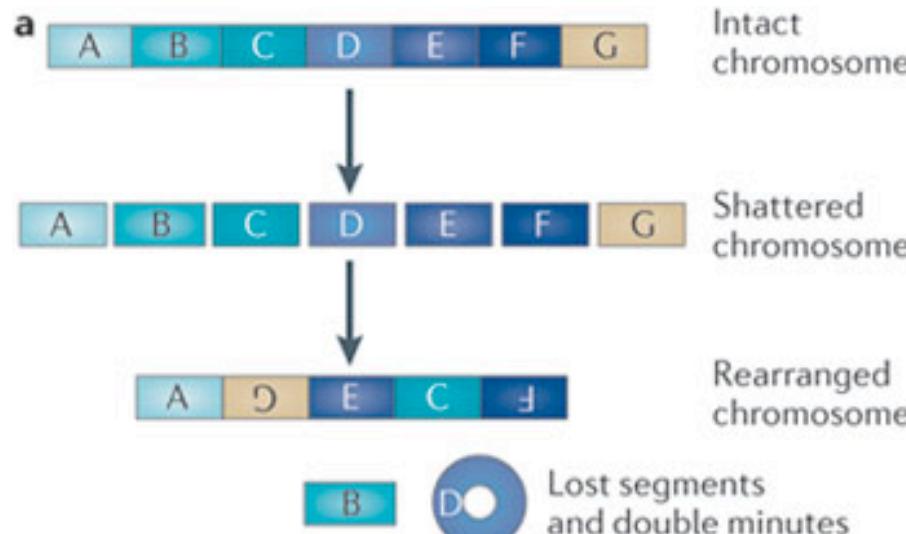


(Davidson et al, 2000)

80+ chromosomes,

Many are a patchwork of fragments of other chromosomes

# Aberrations in cancer genomes



**Chromothripsis**, which literally means 'chromosome shattering', is a phenomenon that has recently been reported to occur in cells harbouring complex genomic rearrangements (CGRs). Has 3 defining characteristics:

- (1) Occurrence of remarkable numbers of rearrangements in localized chromosomal regions;
- (2) Low number of copy number states (generally between one or two) across the rearranged region;
- (3) Alternation in the chromothriptic areas of regions where heterozygosity is preserved with regions presenting loss of heterozygosity (LOH).

## Chromothripsis and cancer: causes and consequences of chromosome shattering

Forment et al (2012) Nature Reviews Cancer. doi:10.1038/nrc3352

# Hypomethylation distinguishes genes of some human cancers from their normal counterparts

Andrew P. Feinberg & Bert Vogelstein

Cell Structure and Function Laboratory, The Oncology Center,  
Johns Hopkins University School of Medicine, Baltimore,  
Maryland 21205, USA

It has been suggested that cancer represents an alteration in DNA, heritable by progeny cells, that leads to abnormally regulated expression of normal cellular genes; DNA alterations such as mutations<sup>1,2</sup>, rearrangements<sup>3-5</sup> and changes in methylation<sup>6-8</sup> have been proposed to have such a role. Because of increasing evidence that DNA methylation is important in gene expression (for review see refs 7, 9-11), several investigators have studied DNA methylation in animal tumours, transformed cells and leukaemia cells in culture<sup>8,12-30</sup>. The results of these studies have varied; depending on the techniques and systems used, an increase<sup>12-19</sup>, decrease<sup>20-24</sup>, or no change<sup>25-29</sup> in the degree of methylation has been reported. To our knowledge, however, primary human tumour tissues have not been used in such studies. We have now examined DNA methylation in human cancer with three considerations in mind: (1) the methylation pattern of specific genes, rather than total levels of methylation, was determined; (2) human cancers and adjacent analogous normal tissues, unconditioned by culture media, were analysed; and (3) the cancers were taken from patients who had received neither radiation nor chemotherapy. In four of five patients studied, representing two histological types of cancer, substantial hypomethylation was found in genes of cancer cells compared with their normal counterparts. This hypomethylation was progressive in a metastasis from one of the patients.

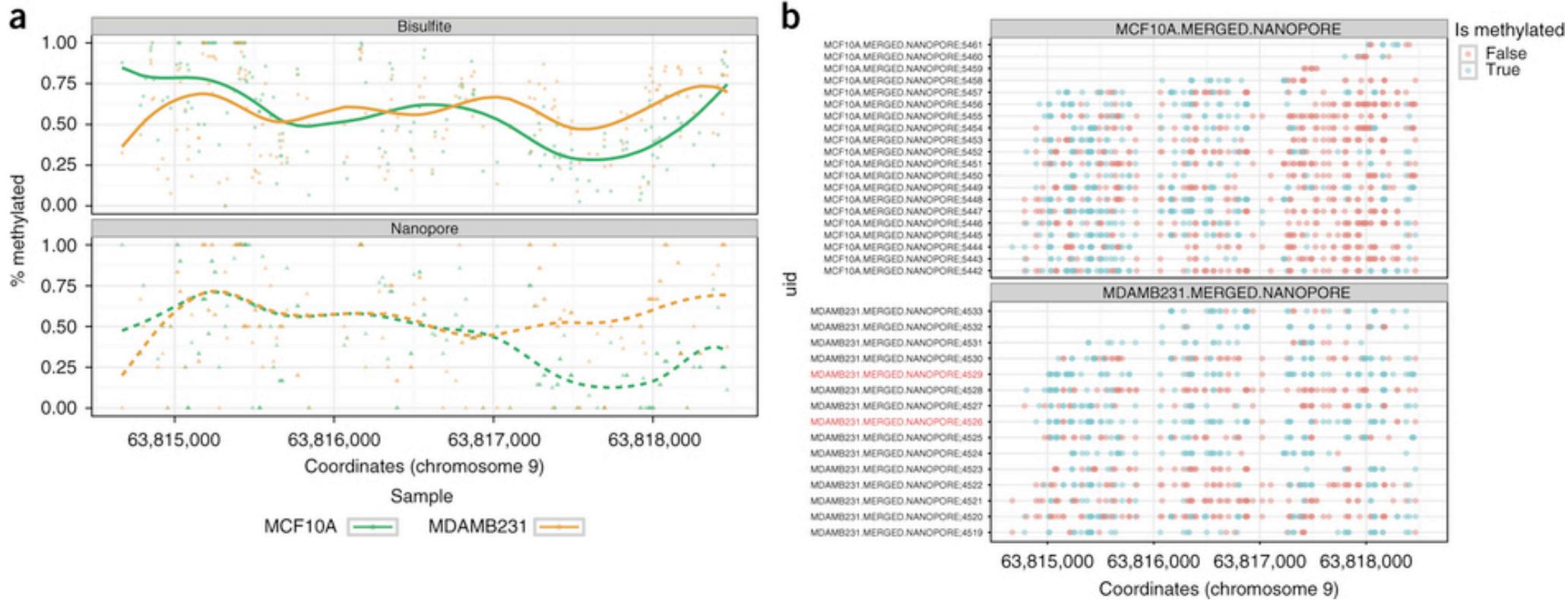
and (3) *Hpa*II and *Hha*I cleavage sites should be present in the regions of the genes.

The first cancer studied was a grade D (ref. 43), moderately well differentiated adenocarcinoma of the colon from a 67-yr-old male. Tissue was obtained from the cancer itself and also from colonic mucosa stripped from the colon at a site just outside the histologically proven tumour margin. Figure 1 shows the pattern of methylation of the studied genes. Before digestion with restriction enzymes, all DNA samples used in the study had a size >25,000 base pairs (bp). After *Hpa*II cleavage, hybridization with a probe made from a cDNA clone of human growth hormone (HGH) showed that significantly more of the DNA was digested to low-molecular weight fragments in DNA from the cancer (labelled C in Fig. 1) than in DNA from the normal colonic mucosa (labelled N). In the hybridization conditions used, the HGH probe detected the human growth hormone genes as well as the related chorionic somatotropin

Table 1 Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues

Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	{ <i>Hpa</i> II	<10	35	—
			{ <i>Hha</i> I	<10	39	—
			{ <i>Hpa</i> II	<10	52	—
		{ <i>Hha</i> I	<10	39	—	—
			{ <i>Hpa</i> II	<10	<10	—
			{ <i>Hha</i> I	<10	<10	—
2	Colon	HGH	{ <i>Hpa</i> II	<10	76	—
			{ <i>Hha</i> I	<10	85	—
			{ <i>Hpa</i> II	<10	58	—
		{ <i>Hha</i> I	<10	23	—	—
			{ <i>Hpa</i> II	<10	<10	—
			{ <i>Hha</i> I	<10	<10	—
3	Colon	HGH	{ <i>Hpa</i> II	<10	41	—
			{ <i>Hha</i> I	<10	38	—
		{ <i>Hpa</i> II	<10	50	—	—
		{ <i>Hha</i> I	<10	22		

# Methylation changes in cancer detected by Nanopore Sequencing

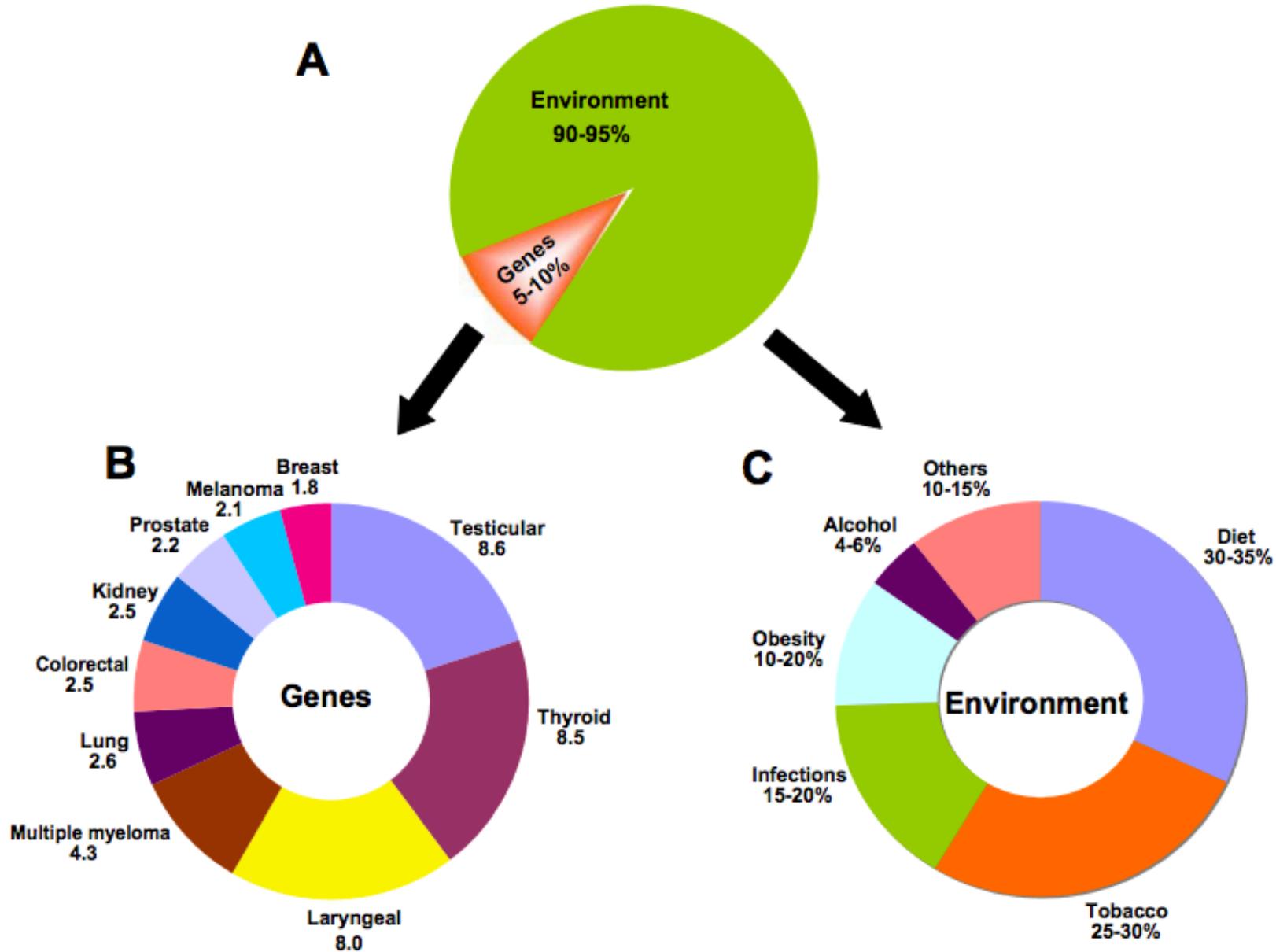


Comparison of bisulfite sequencing and nanopore-based R7.3 data in reduced representation data sets from cancer and normal cells. (a) Raw data (points) and smoothed data (lines) for methylation, as determined by bisulfite sequencing (top) and nanopore-based sequencing using an R7.3 pore (bottom), in a genomic region from the human mammary epithelial cell line MCF10A (green) and metastatic mammary epithelial cell line MDA-MB-231 (orange). (b) Same region as in a but with individual nanopore reads plotted separately. Each CpG that can be called is a point. Blue indicates methylated; red indicates unmethylated.

## Detecting DNA cytosine methylation using nanopore sequencing

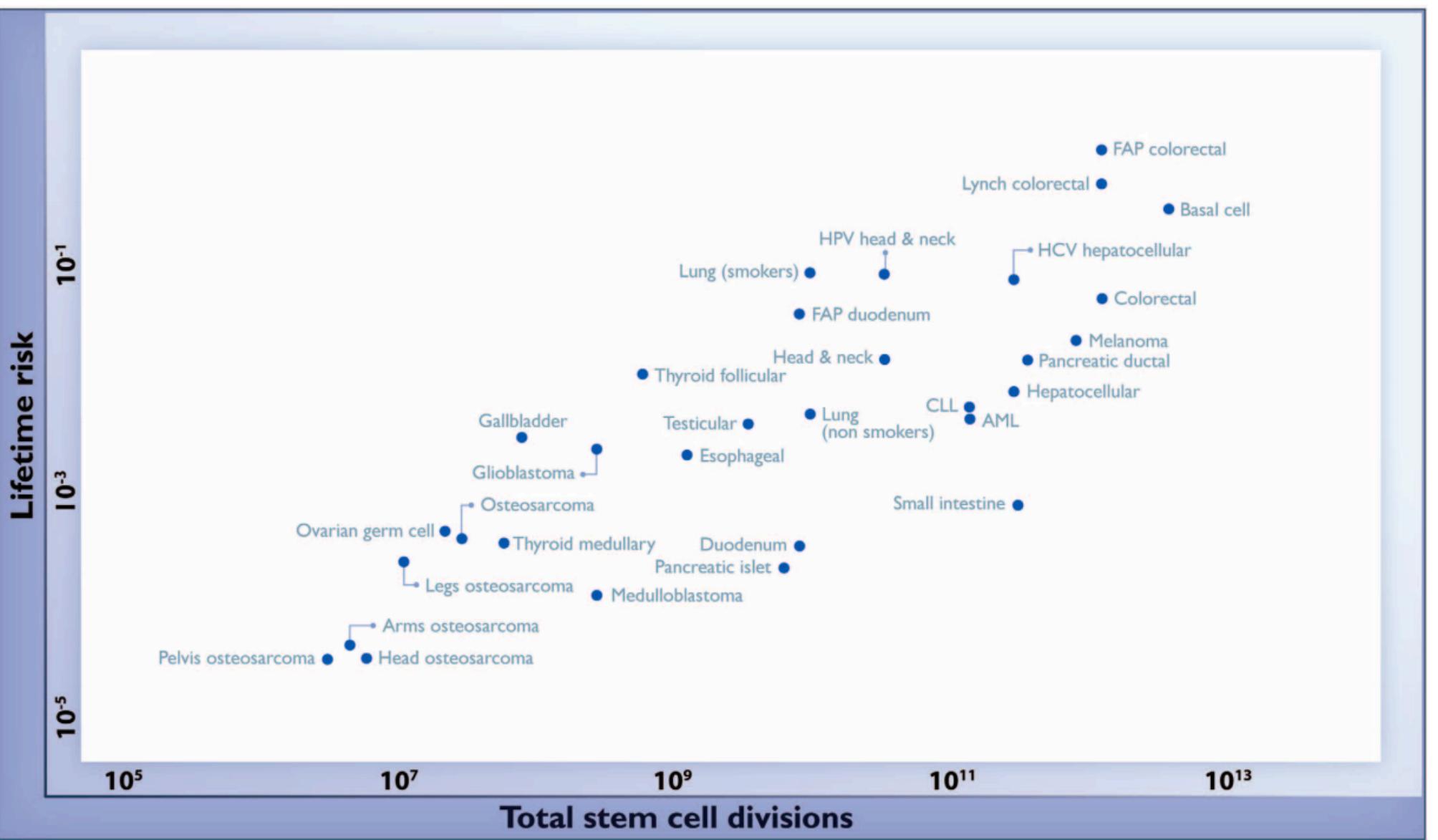
Simpson, Workman, Zuzarte, David, Dursi, Timp (2017) Nature Methods. doi:10.1038/nmeth.4184

# Causes of Cancer



**Cancer is a Preventable Disease that Requires Major Lifestyle Changes**

Anand et al (2008) Pharmaceutical Research. doi: 10.1007/s11095-008-9661-9

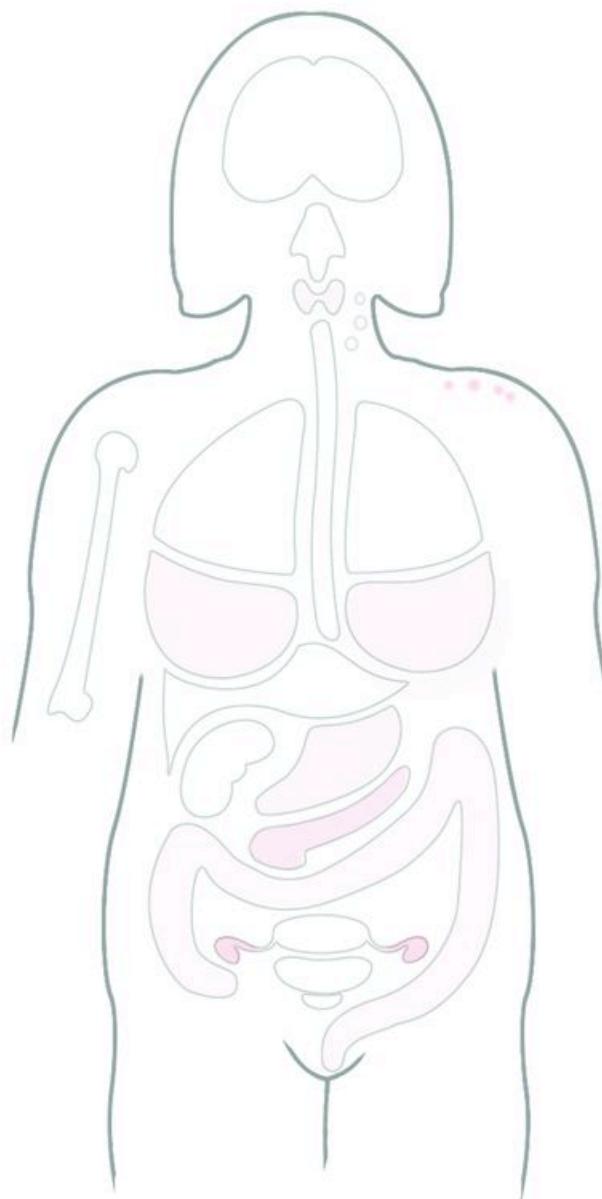


FAP = Familial Adenomatous Polyposis ◆ HCV = Hepatitis C virus ◆ HPV = Human papillomavirus ◆ CLL = Chronic lymphocytic leukemia ◆ AML = Acute myeloid leukemia

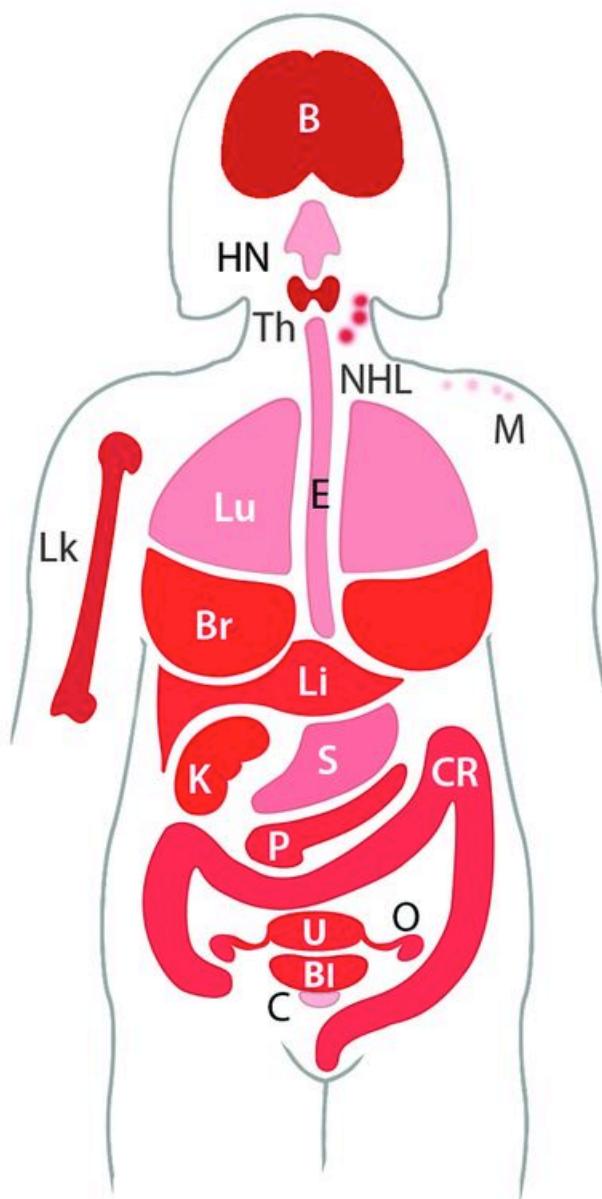
**Fig. 1. The relationship between the number of stem cell divisions in the lifetime of a given tissue and the lifetime risk of cancer in that tissue.**  
Values are from table S1, the derivation of which is discussed in the supplementary materials.

**Variation in cancer risk among tissues can be explained by the number of stem cell divisions**  
Tomasetti and Vogelstein (2015) Science. DOI: 10.1126/science.1260825

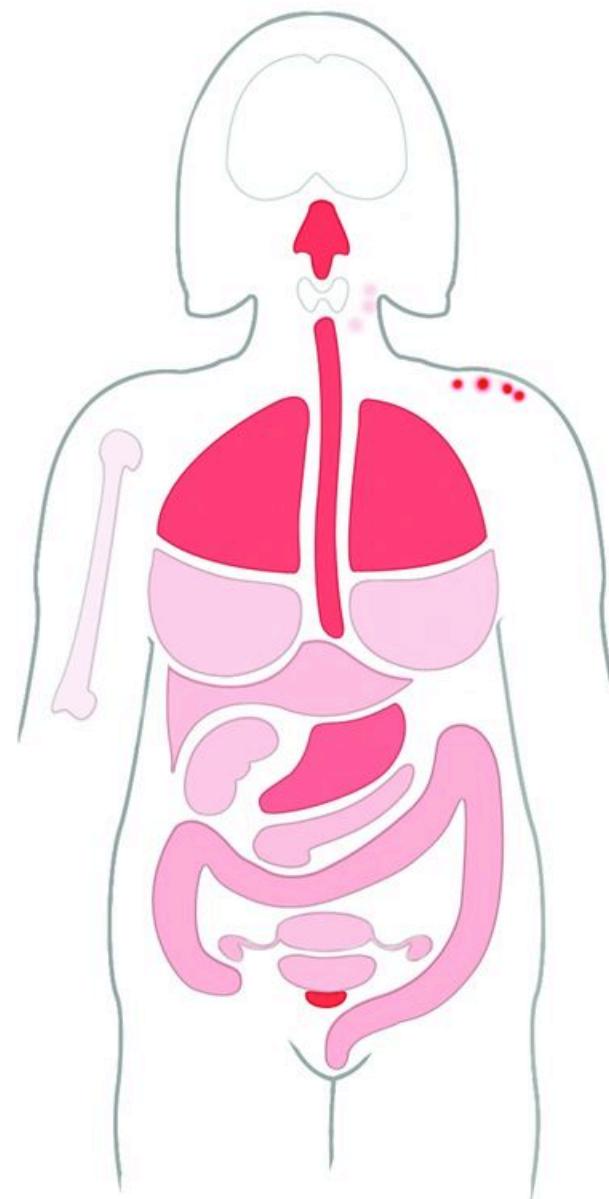
### Hereditary



### Replicative



### Environmental



0%

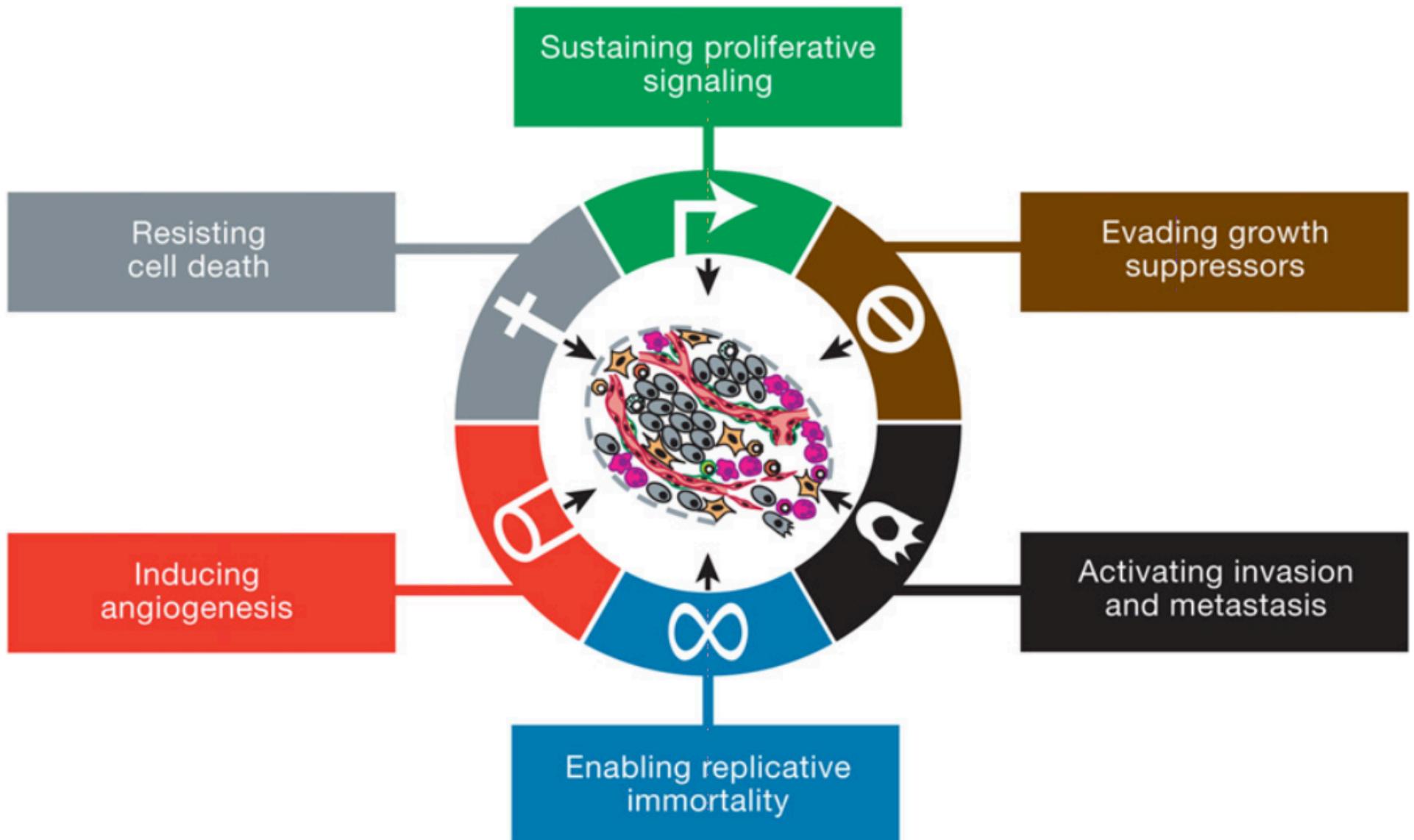
100%

Percentage of driver mutations attributable to each factor

**Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention**

Tomasetti, Li, and Vogelstein (2017) Science. DOI: 10.1126/science.aaf9011

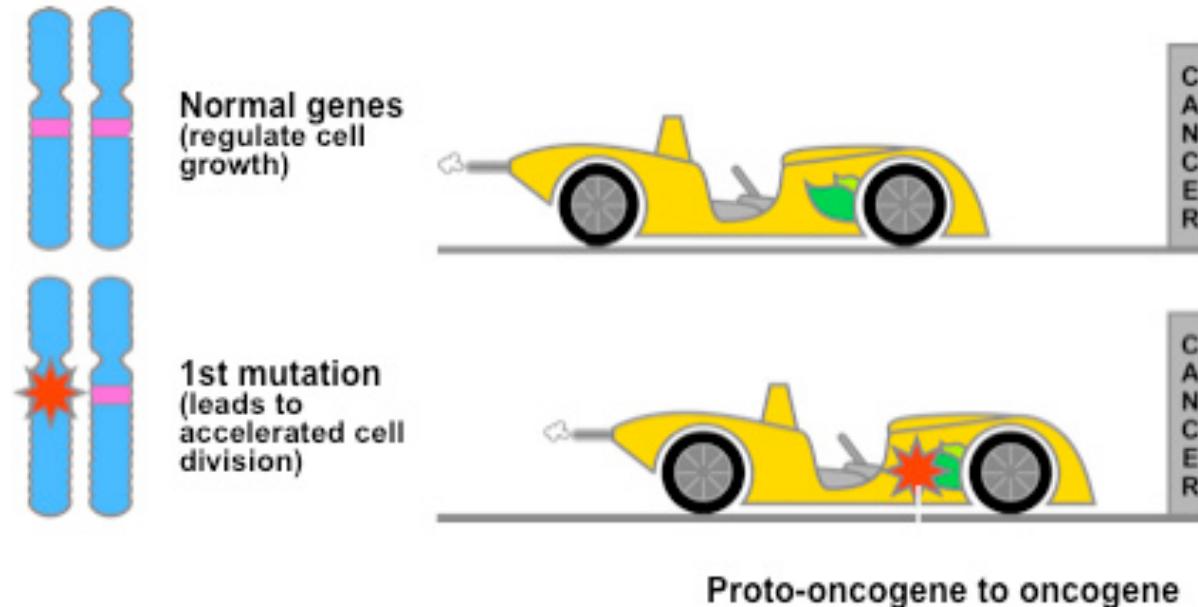
# The Six Hallmarks of Cancer



## Hallmarks of Cancer

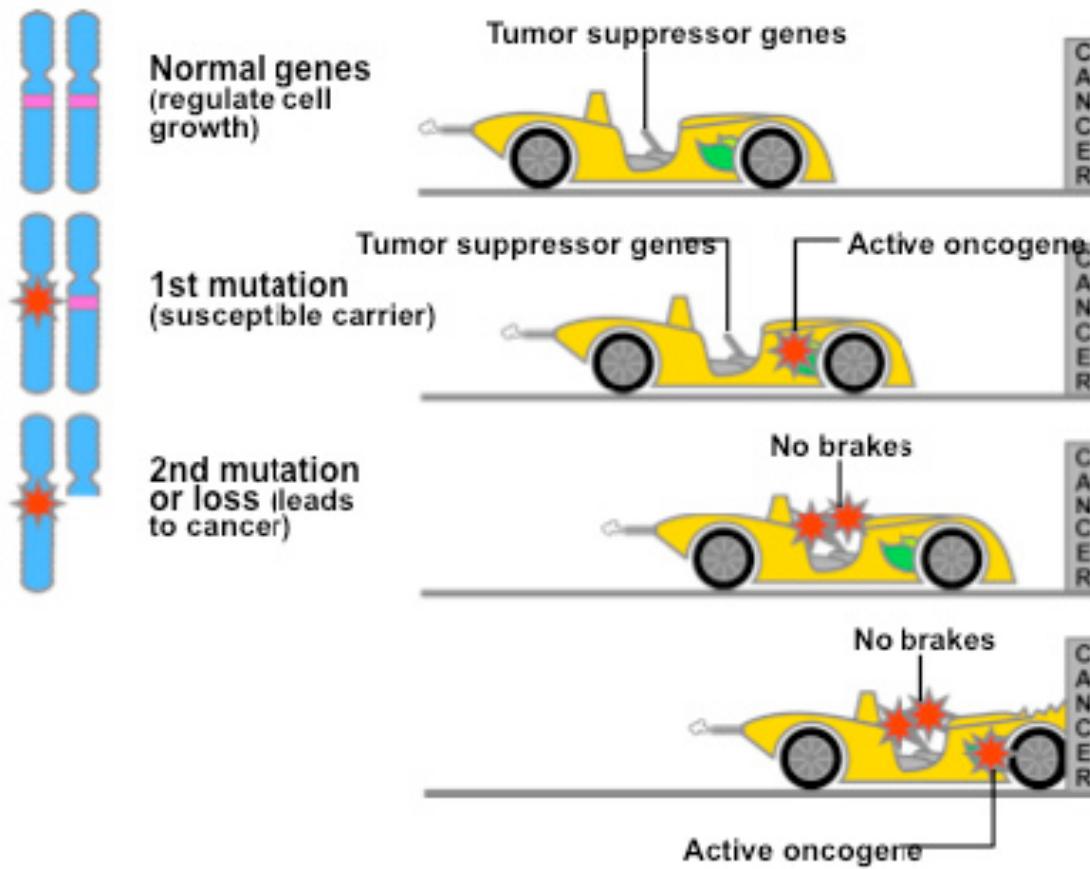
Hanahan and Weinberg (2000) Cell. [http://doi.org/10.1016/S0092-8674\(00\)81683-9](http://doi.org/10.1016/S0092-8674(00)81683-9)

# Oncogenes



- ***HER-2/neu***: encodes for a cell surface receptor that can stimulate cell division. The HER-2/neu gene is amplified in up to 30% of human breast cancers.
- ***RAS***: The Ras gene products are involved in kinase signaling pathways that ultimately control transcription of genes, regulating cell growth and differentiation.
- ***MYC***: The Myc protein is a transcription factor and controls expression of several genes.
- ***SRC***: First oncogene ever discovered. The Src protein is a tyrosine kinase, which regulates cell activity.
- ***hTER***: Codes for an enzyme (telomerase) that maintains chromosome ends.

# Tumor Suppressors



- **p53:** a transcription factor that regulates cell division and cell death.
- **Rb:** alters the activity of transcription factors and therefore controls cell division.
- **APC:** controls the availability of a transcription factor.
- **PTEN:** acts by opposing the action of PI3K, which is essential for anti-apoptotic, pro-tumorigenic Akt activation.

# TP53: The first and most important tumor suppressor

Mechanism of inactivating p53	Typical tumours	Effect of inactivation
Amino-acid-changing mutation in the DNA-binding domain	Colon, breast, lung, bladder, brain, pancreas, stomach, oesophagus and many others	Prevents p53 from binding to specific DNA sequences and activating the adjacent genes
Deletion of the carboxy-terminal domain	Occasional tumours at many different sites	Prevents the formation of tetramers of p53
Multiplication of the MDM2 gene in the genome	Sarcomas, brain	Extra MDM2 stimulates the degradation of p53
Viral infection	Cervix, liver, lymphomas	Products of viral oncogenes bind to and inactivate p53 in the cell, in some cases stimulating p53 degradation
Deletion of the p14 <sup>ARF</sup> gene	Breast, brain, lung and others, especially when p53 itself is not mutated	Failure to inhibit MDM2 and keep p53 degradation under control
Mislocalization of p53 to the cytoplasm, outside the nucleus	Breast, neuroblastomas	Lack of p53 function (p53 functions only in the nucleus)

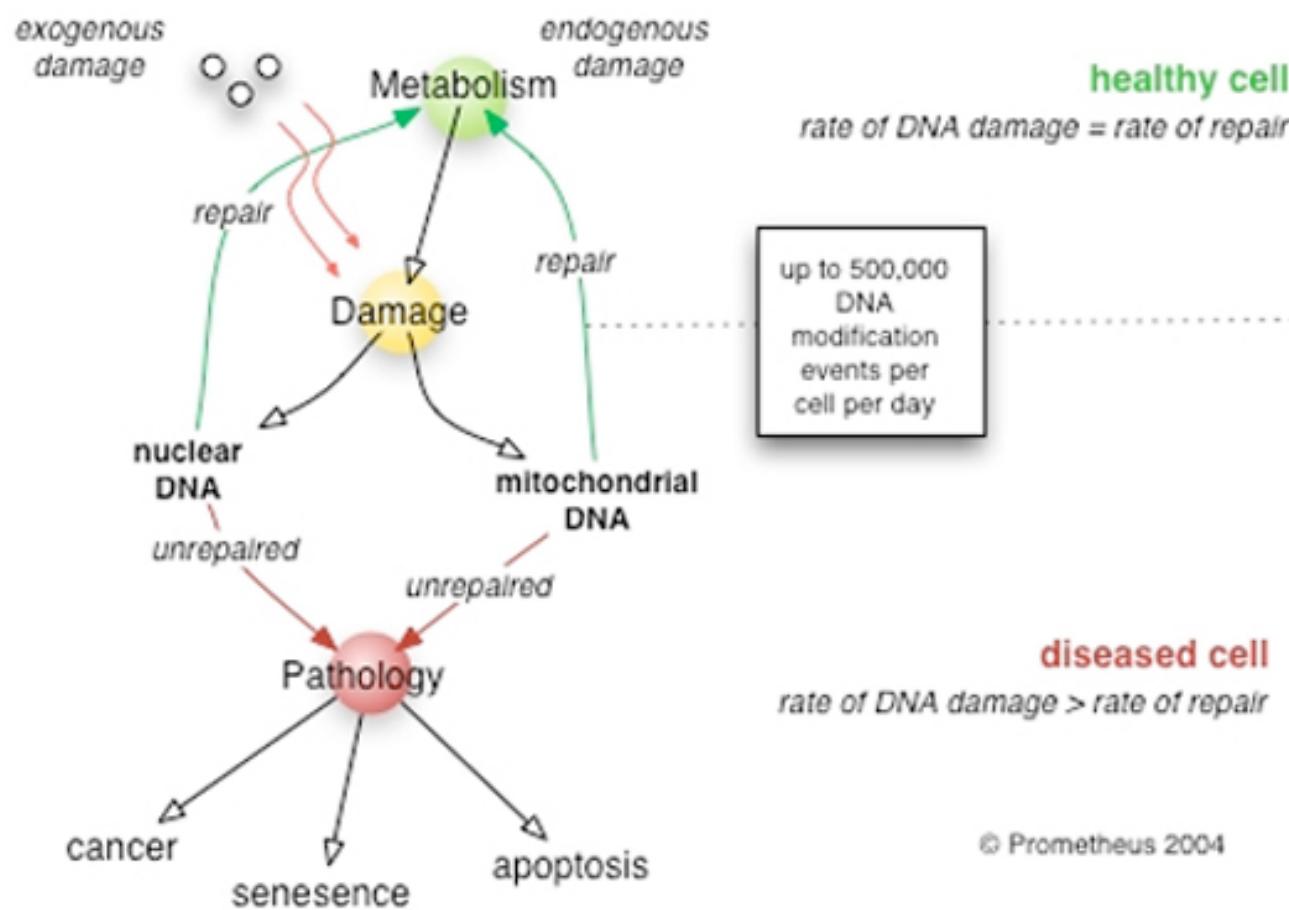
Figure 1 The many ways in which p53 may malfunction in human cancers.

>10,000 known mutations  
>17,000 publications

## Surfing the p53 network

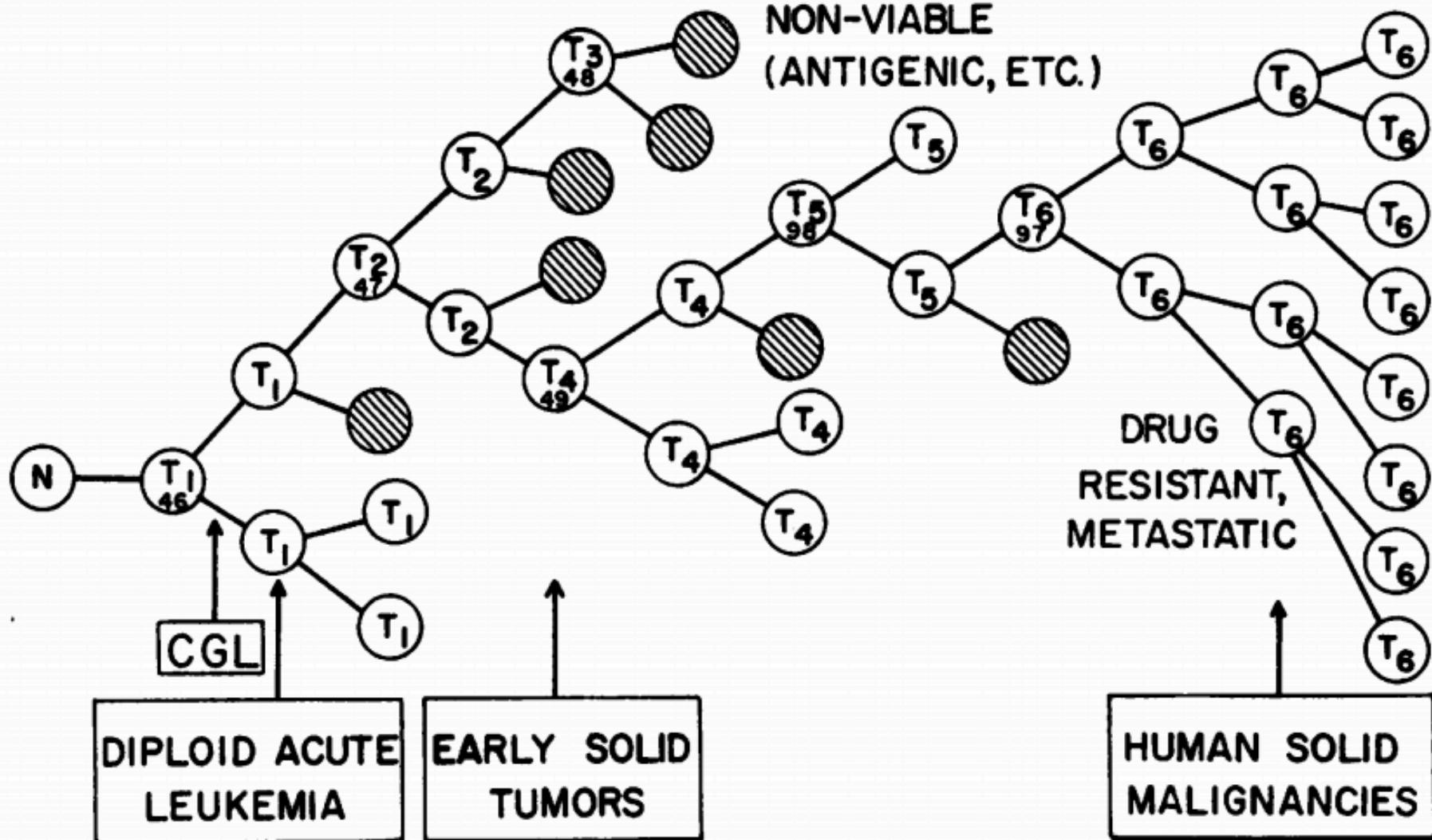
Volgelstein et al (2000) Nature. DOI: 10.1038/35042675

# DNA Repair Genes



- ***BRCA1 and BRCA2 (breast cancer type 1/2 susceptibility genes)*:** Normally expressed in the cells of breast and other tissue, where they help repair damaged DNA, or destroy cells if DNA cannot be repaired. They are involved in the repair of chromosomal damage with an important role in the error-free repair of DNA double-strand breaks

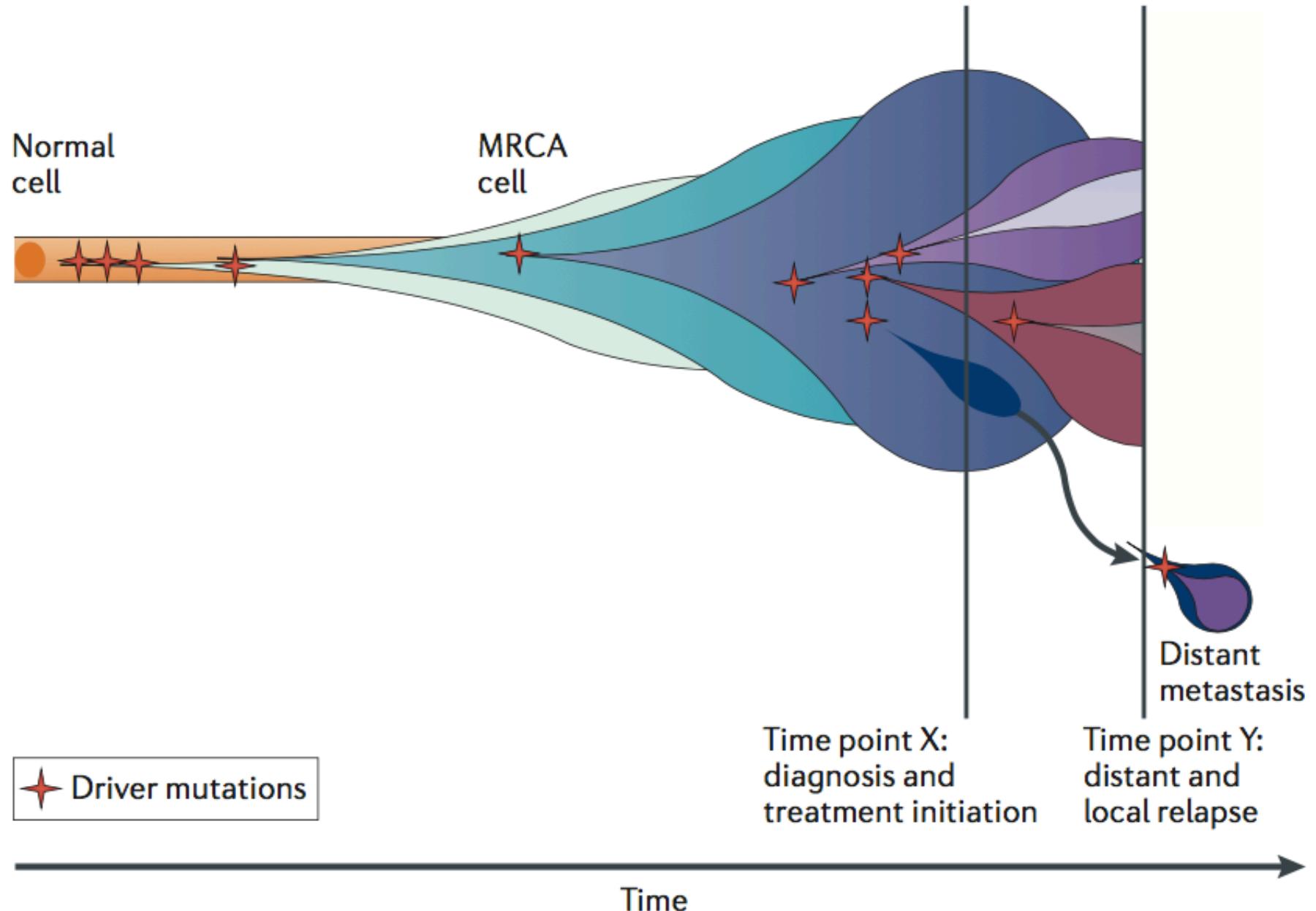
# Tumor Evolution



# The Clonal Evolution of Tumor Cell Populations

Peter C. Nowell (1976) *Science*. 194(4260):23-28 DOI: 10.1126/science.959840

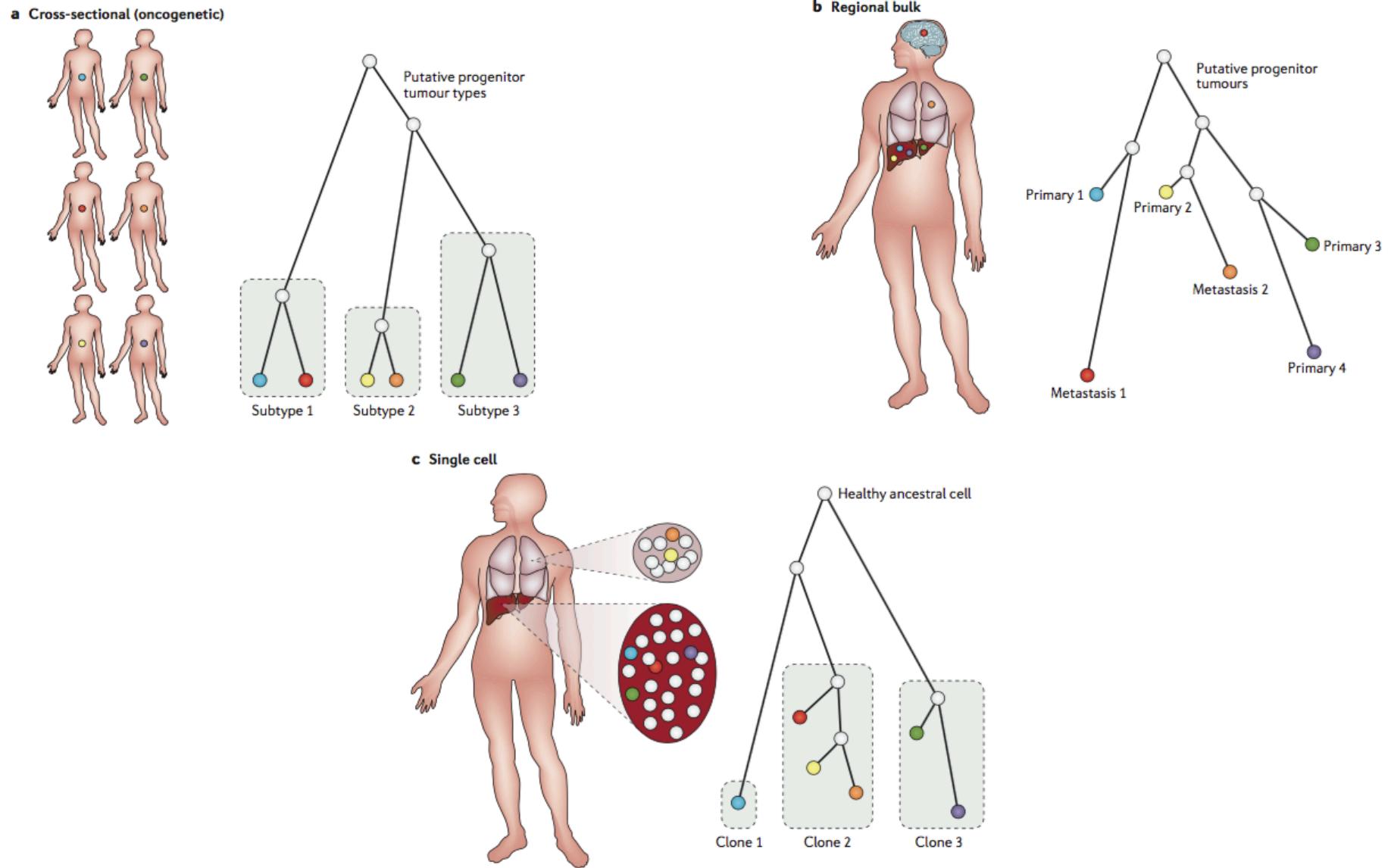
# Tumor Evolution



## Evolution of the cancer genome

Yates & Campbell (2012) Nature Review Genetics. doi:10.1038/nrg3317

# Tumor Heterogeneity

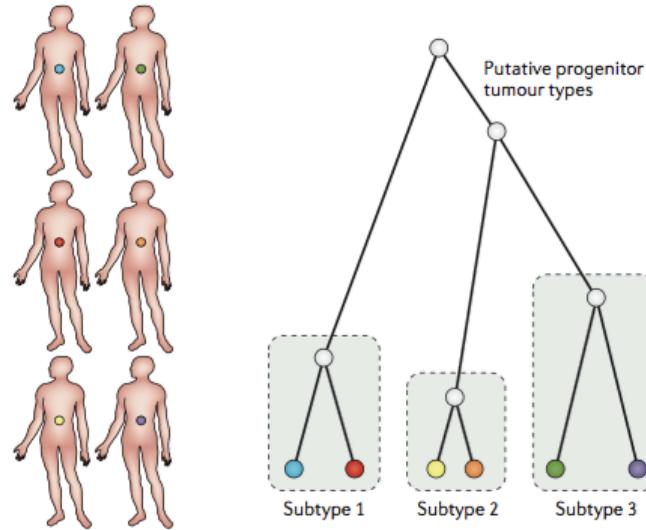


**The evolution of tumour phylogenetics: principles and practice**

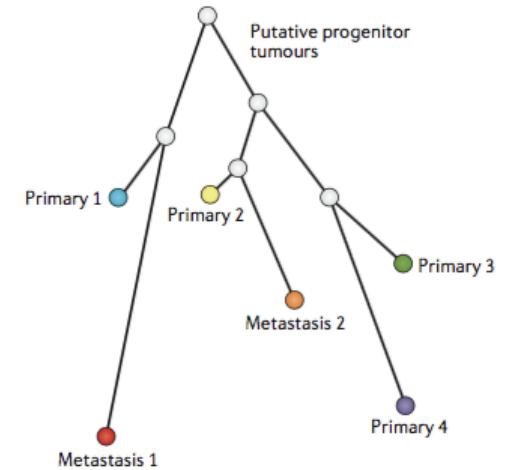
Schwarz and Schaffer (2017) *Nature Reviews Genetics*. doi:10.1038/nrg.2016.170

# Tumor Heterogeneity

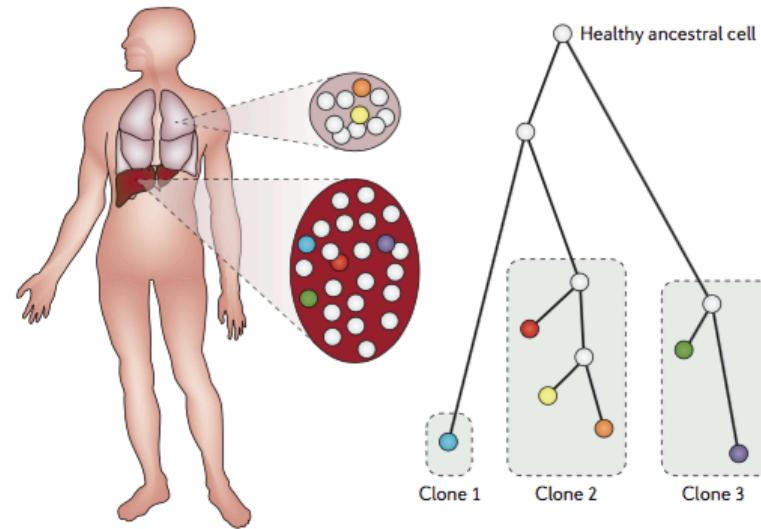
a Cross-sectional (oncogenetic)



b Regional bulk



c Single cell

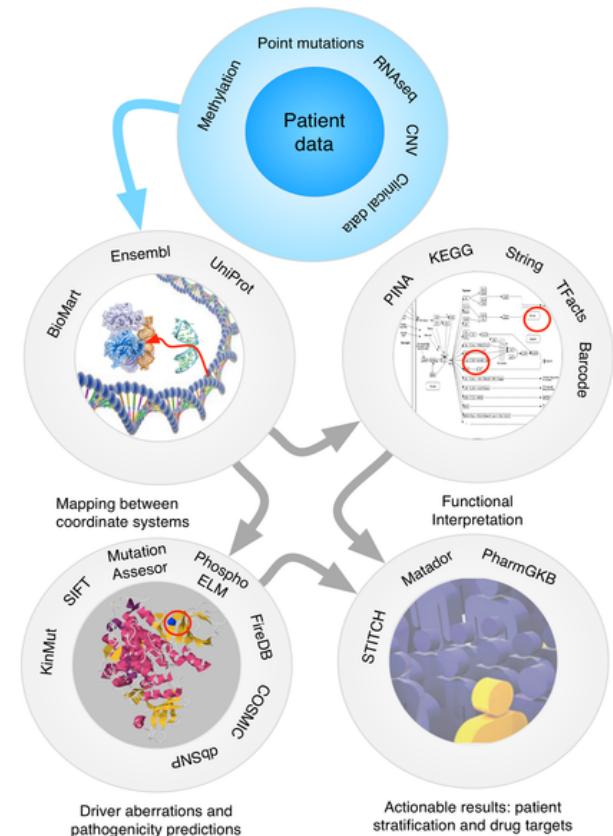


**The evolution of tumour phylogenetics: principles and practice**

Schwarz and Schaffer (2017) *Nature Reviews Genetics*. doi:10.1038/nrg.2016.170

# Cancer Mutation Analysis

Analysis Steps	Disciplines	Techniques	Challenges
Sequencing, alignment, and Variant calling.	NGS bioinformatics, alignments, and databasing.	Data management, alignment and variant calling, and quality control.	Sample preparation, clonal mosaicism, and alignment efficiency and biases.
Consequence Analysis. Mutation Recurrence. Classification of driver-passenger mutations.	Structural bioinformatics, regulatory genomics, and biostatistics.	Pathogenicity predictions and recurrence statistics.	Predicting effect of mutations in protein function, regulation, splicing, etc. Establishing mutational background.
Pathway and functional analysis.	Systems Biology, pathway modeling, and Interactomics.	Enrichment statistics and network analysis.	Multiple pathway definitions and violated statistical assumptions.
Integration, Visualization, and disease centered Interpretation.	Application development, Man-machine interfaces, Pharmacogenomics, Text-mining, and Information management.	Data integration and visualization.	Multiple, heterogeneous, experimental data sources, database formats, and software resources. Biomedical expertise required.



Vazquez M, de la Torre V, Valencia A (2012) Chapter 14: Cancer Genome Analysis. PLOS Computational Biology 8(12): e1002824. <https://doi.org/10.1371/journal.pcbi.1002824>

<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002824>

# First Cancer Genome

nature

Vol 456 | 6 November 2008 | doi:10.1038/nature07485

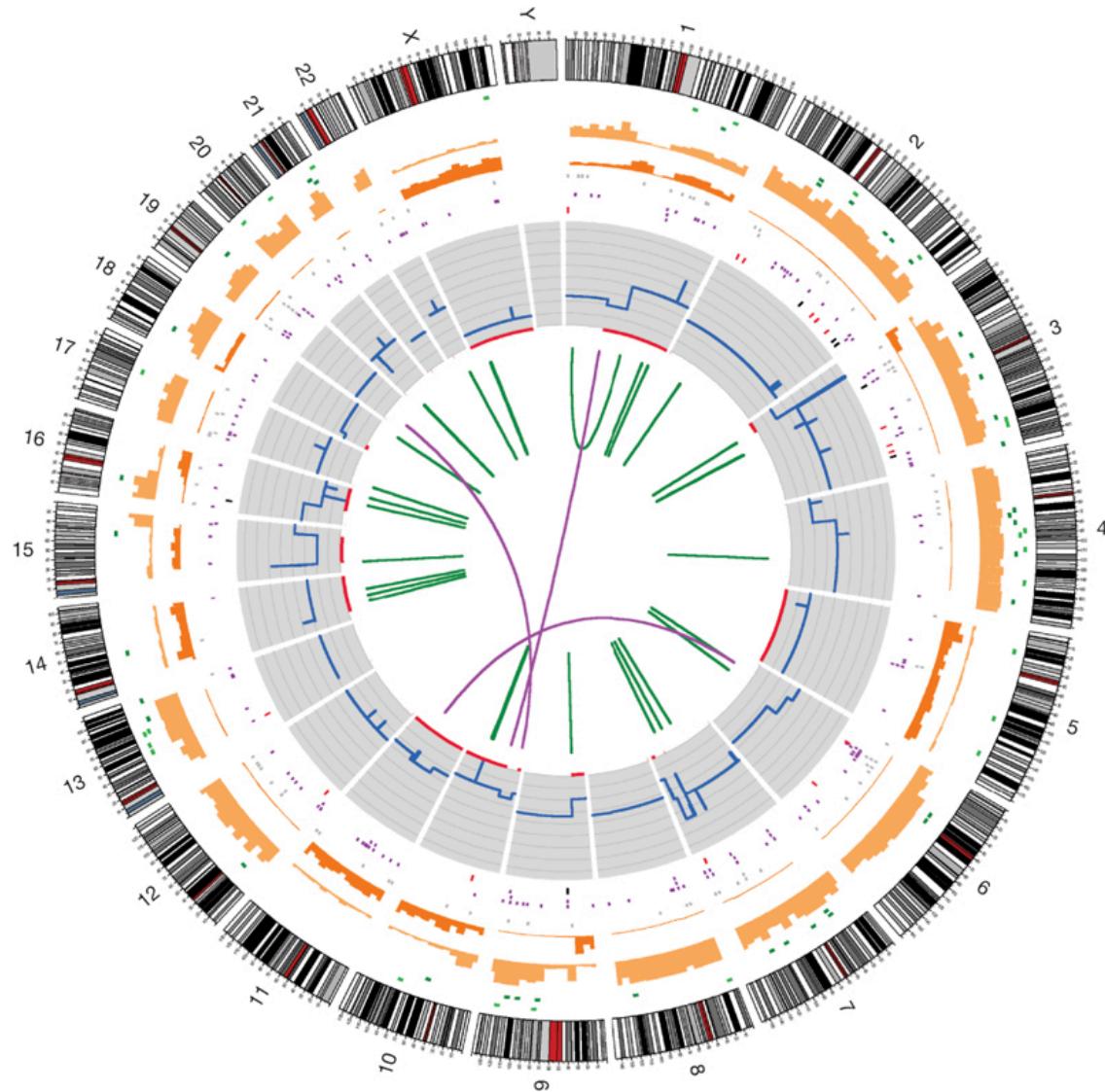
## ARTICLES

### DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley<sup>1,2,3,4\*</sup>, Elaine R. Mardis<sup>2,3\*</sup>, Li Ding<sup>2,3</sup>, Bob Fulton<sup>3</sup>, Michael D. McLellan<sup>3</sup>, Ken Chen<sup>3</sup>, David Dooling<sup>3</sup>, Brian H. Dunford-Shore<sup>3</sup>, Sean McGrath<sup>3</sup>, Matthew Hickenbotham<sup>3</sup>, Lisa Cook<sup>3</sup>, Rachel Abbott<sup>3</sup>, David E. Larson<sup>3</sup>, Dan C. Koboldt<sup>3</sup>, Craig Pohl<sup>3</sup>, Scott Smith<sup>3</sup>, Amy Hawkins<sup>3</sup>, Scott Abbott<sup>3</sup>, Devin Locke<sup>3</sup>, LaDeana W. Hillier<sup>3,8</sup>, Tracie Miner<sup>3</sup>, Lucinda Fulton<sup>3</sup>, Vincent Magrini<sup>2,3</sup>, Todd Wylie<sup>3</sup>, Jarret Glasscock<sup>3</sup>, Joshua Conyers<sup>3</sup>, Nathan Sander<sup>3</sup>, Xiaoqi Shi<sup>3</sup>, John R. Osborne<sup>3</sup>, Patrick Minx<sup>3</sup>, David Gordon<sup>8</sup>, Asif Chinwalla<sup>3</sup>, Yu Zhao<sup>1</sup>, Rhonda E. Ries<sup>1</sup>, Jacqueline E. Payton<sup>5</sup>, Peter Westervelt<sup>1,4</sup>, Michael H. Tomasson<sup>1,4</sup>, Mark Watson<sup>3,4,5</sup>, Jack Baty<sup>6</sup>, Jennifer Ivanovich<sup>4,7</sup>, Sharon Heath<sup>1,4</sup>, William D. Shannon<sup>1,4</sup>, Rakesh Nagarajan<sup>4,5</sup>, Matthew J. Walter<sup>1,4</sup>, Daniel C. Link<sup>1,4</sup>, Timothy A. Graubert<sup>1,4</sup>, John F. DiPersio<sup>1,4</sup> & Richard K. Wilson<sup>2,3,4</sup>

Acute myeloid leukaemia is a highly malignant haematopoietic tumour that affects about 13,000 adults in the United States each year. The treatment of this disease has changed little in the past two decades, because most of the genetic events that initiate the disease remain undiscovered. Whole-genome sequencing is now possible at a reasonable cost and timeframe to use this approach for the unbiased discovery of tumour-specific somatic mutations that alter the protein-coding genes. Here we present the results obtained from sequencing a typical acute myeloid leukaemia genome, and its matched normal counterpart obtained from the same patient's skin. We discovered ten genes with acquired mutations; two were previously described mutations that are thought to contribute to tumour progression, and eight were new mutations present in virtually all tumour cells at presentation and relapse, the function of which is not yet known. Our study establishes whole-genome sequencing as an unbiased method for discovering cancer-initiating mutations in previously unidentified genes that may respond to targeted therapies.

# First Melanoma Genome

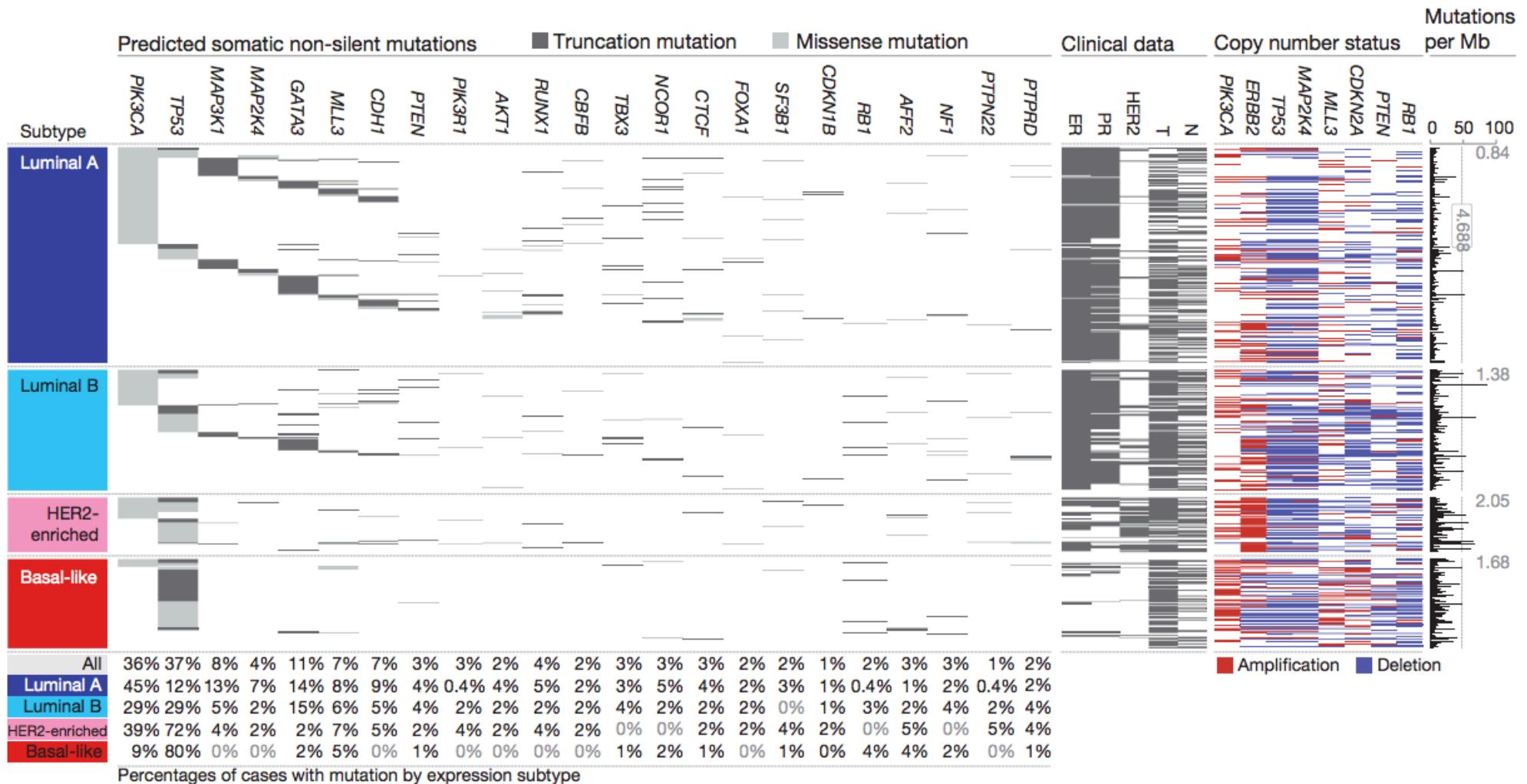


- Insertions (light-green rectangles);
- Deletions (dark-green rectangles);
- Heterozygous (light-orange bars) and Homozygous (dark-orange bars) Substitutions
- Coding substitutions (coloured squares: silent in grey, missense in purple, nonsense in red and splice site in black);
- Copy number (blue lines); regions of LOH (red lines);
- Intrachromosomal rearrangements (green lines);
- Interchromosomal rearrangements (purple lines).

**A comprehensive catalogue of somatic mutations from a human cancer genome**

Pleasance et al (2010) Nature. doi:10.1038/nature08658

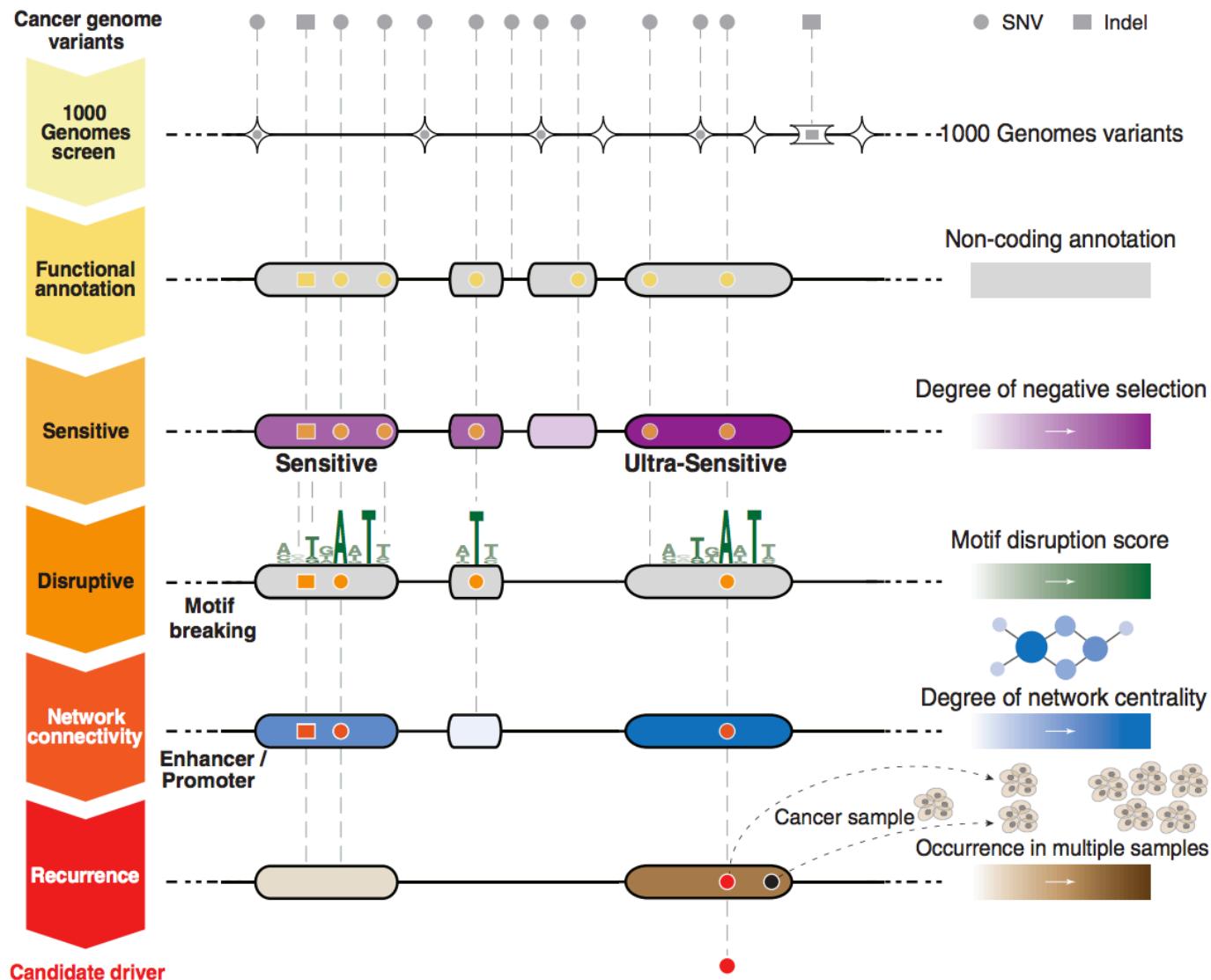
# Mutations in Breast Cancer



# Comprehensive molecular portraits of human breast tumours

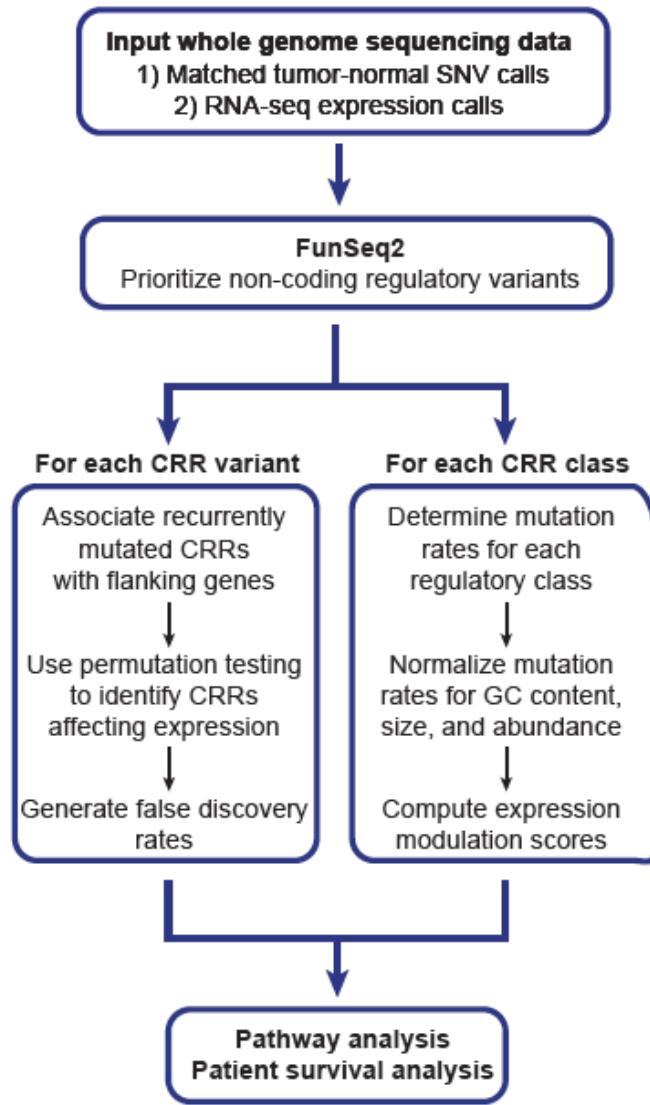
Cancer Genome Atlas Network (2012) Nature. doi:10.1038/nature11412

# Finding Driving Mutations



**Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics**  
Khurana et al (2013) Science. DOI: 10.1126/science.1235587

# Regulatory mutations in PDAC



**Coding alterations of PDAC are now fairly well established but non-coding mutations (NCMs) largely unexplored**

- Developed GECCO to analyze the thousands of somatic mutations observed from hundreds of tumors to find potential drivers of gene expression and pathogenesis

- NCMs are enriched in known and novel pathways
- NCMs correlate with changes in gene expression
- NCMs can demonstrably modulate gene expression
- NCMs correlate with novel clinical outcomes

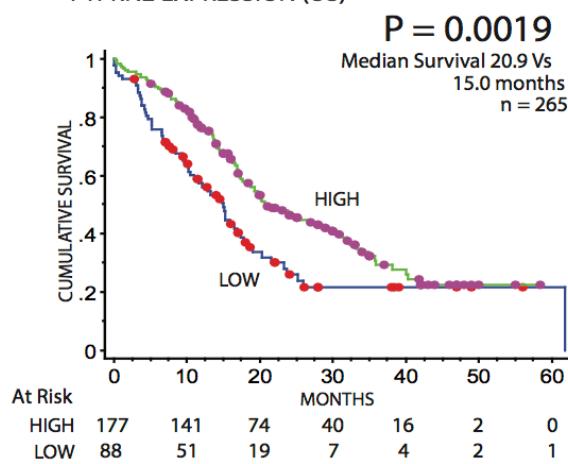
**NCMs are an important mechanism for tumor genome evolution**

# Driving Non-Coding Mutations

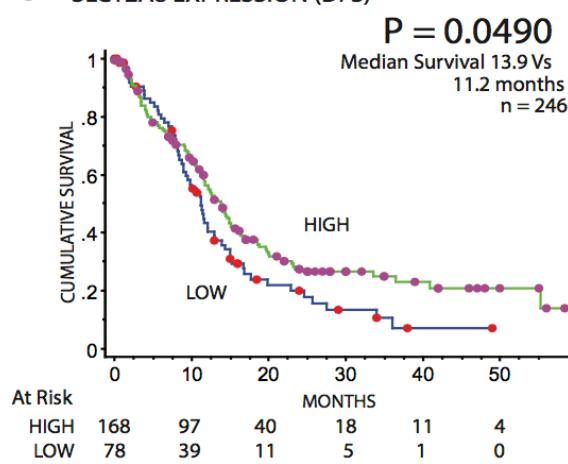
## a NCMs correlate with gene expression changes

CRR (MUT#)	Nearest gene	MUT allele	WT allele	Fold change	p-value	q-value
MAX (5)	<i>PTPRN2</i>	0.82	10.92	0.075	0.00593	0.09689
FOSL2 (7)	<i>KCNQ1</i>	0.85	6.39	0.133	0.02456	0.18212
TAF7 (9)	<i>SNRPN</i>	0.46	3.4	0.135	0.00818	0.11818
NFKB1 (7)	<i>GYPC</i>	1.08	7.29	0.148	0.01845	0.15157
TAF1 (6)	<i>PDPN</i>	2.09	13.08	0.160	0.03544	0.22016
BCLAF1 (5)	<i>PRSS12</i>	1.07	6.46	0.166	0.01107	0.14144
MAFK (3)	<i>SOX5</i>	0.29	1.63	0.178	0.02851	0.20379
POU2F2 (6)	<i>MIR4420</i>	8.16	40.24	0.203	0.01773	0.15157
WRNIP1 (3)	<i>IKZF1</i>	0.64	3.15	0.203	0.01811	0.15157
GATA3 (3)	<i>PCLO</i>	0.35	1.67	0.210	0.01113	0.14144
JUND (3)	<i>TUSC7</i>	0.98	4.53	0.216	0.02909	0.20560
REST (3)	<i>MTERF4</i>	1.46	5.78	0.253	0.02209	0.16542
GATA1 (3)	<i>FNIP2</i>	7.59	18.32	0.414	0.02588	0.18929
CEBPB (3)	<i>PNPLA8</i>	5.69	13.62	0.418	0.01726	0.15157
EGR1 (5)	<i>SLC12A8</i>	4.34	7.99	0.542	0.04185	0.23823
SIN3A (3)	<i>FAM192A</i>	20.31	30.48	0.666	0.01788	0.15157

## b PTPRN2 EXPRESSION (OS)



## c SLC12A8 EXPRESSION (DFS)

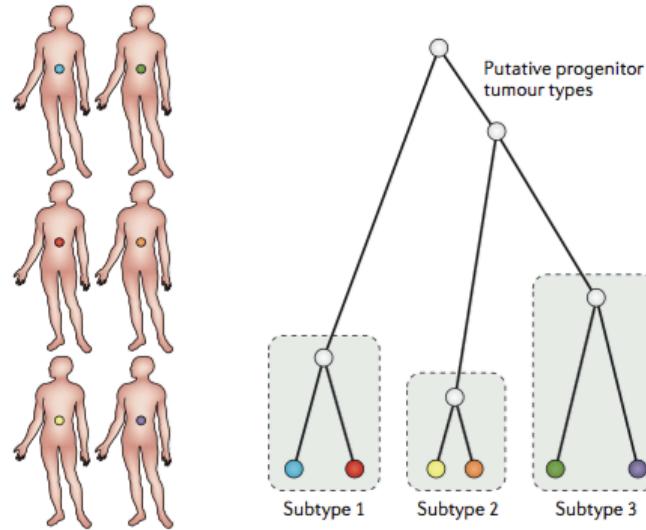


Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma

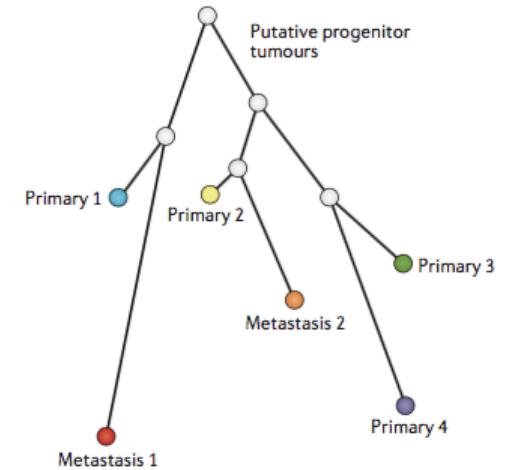
Feigin, M, Garvin, T et al. (2017) Nature Genetics. In press

# Tumor Heterogeneity

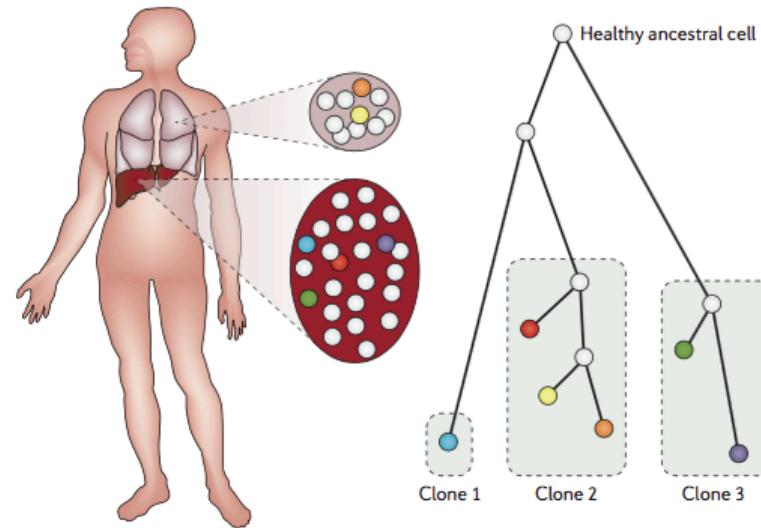
a Cross-sectional (oncogenetic)



b Regional bulk



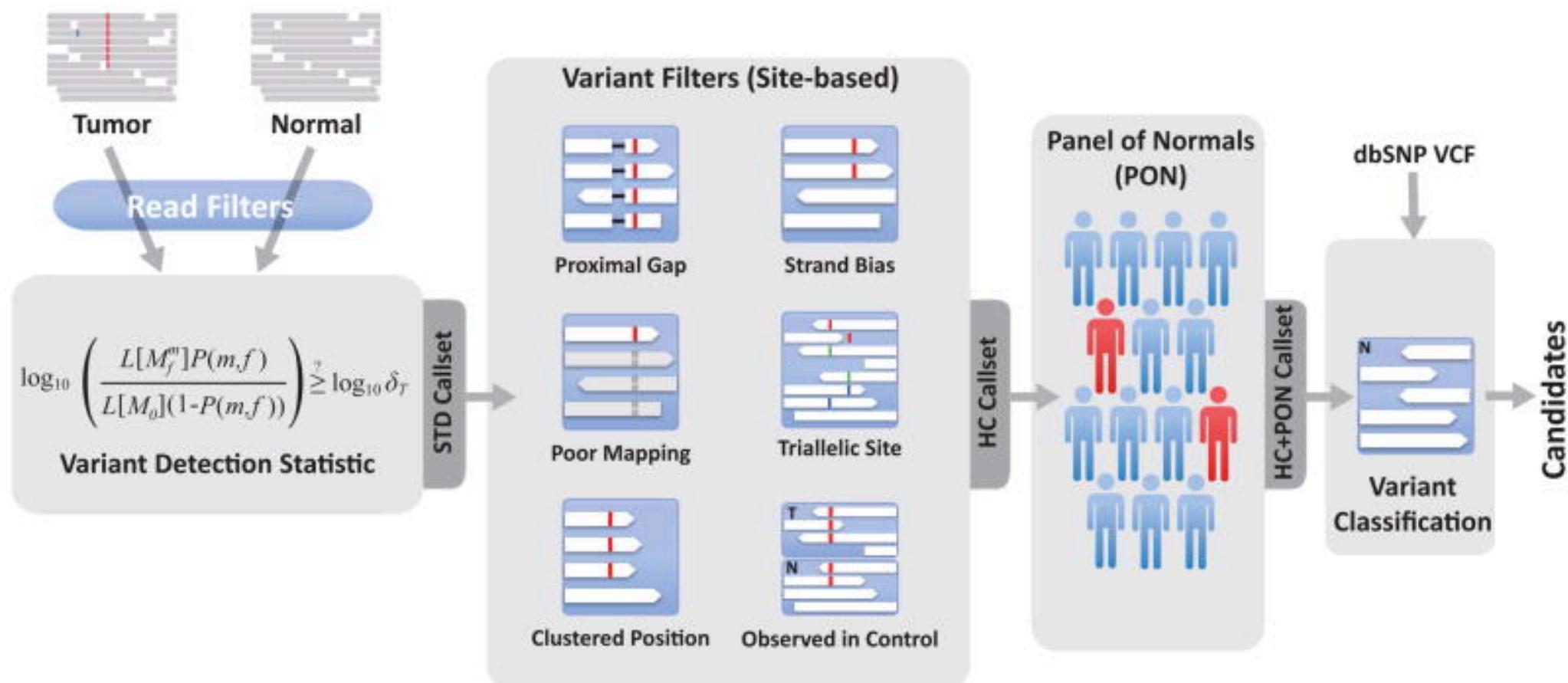
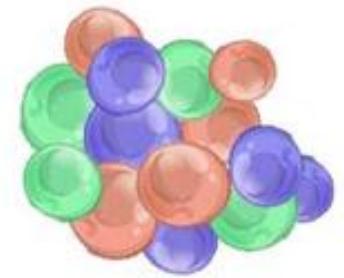
c Single cell



**The evolution of tumour phylogenetics: principles and practice**

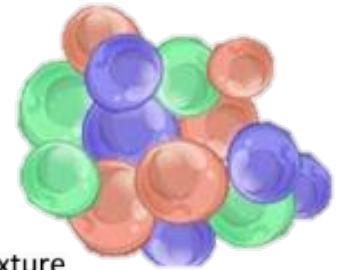
Schwarz and Schaffer (2017) *Nature Reviews Genetics*. doi:10.1038/nrg.2016.170

# Tumor Heterogeneity

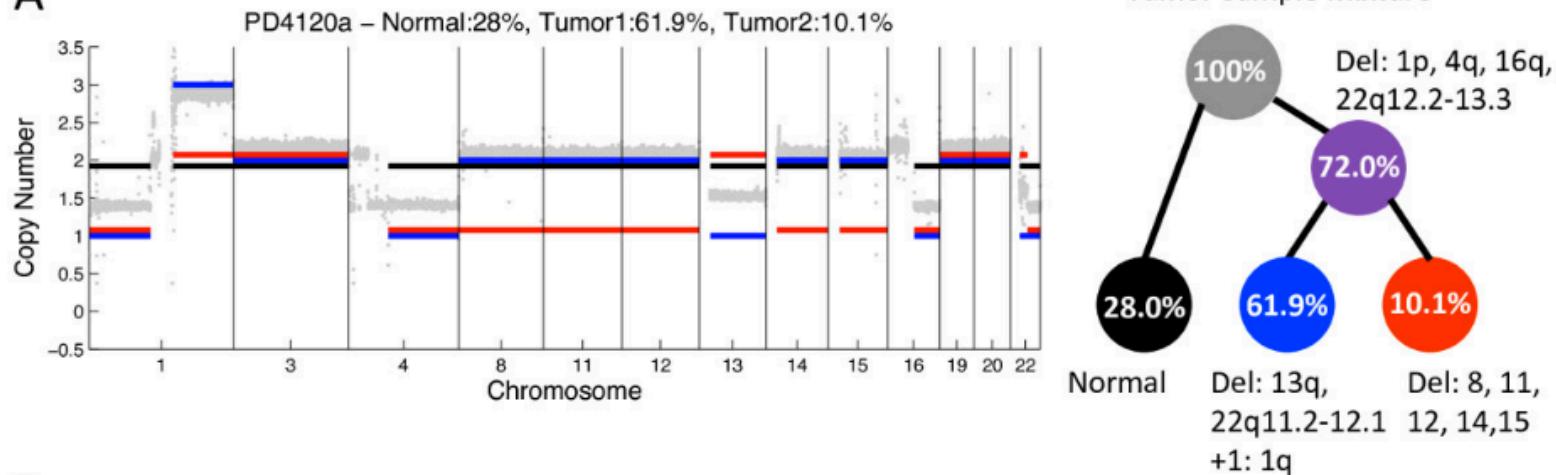


**Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples**  
Cibulskis et al (2013) Nature Biotech. doi:10.1038/nbt.2514

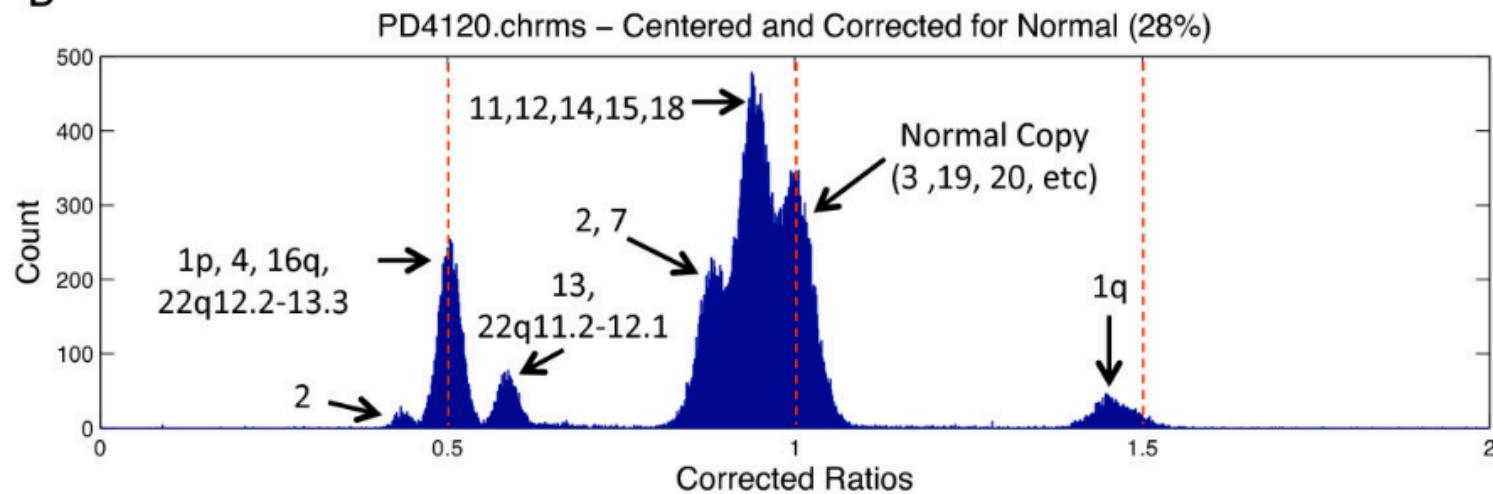
# Tumor Heterogeneity



A



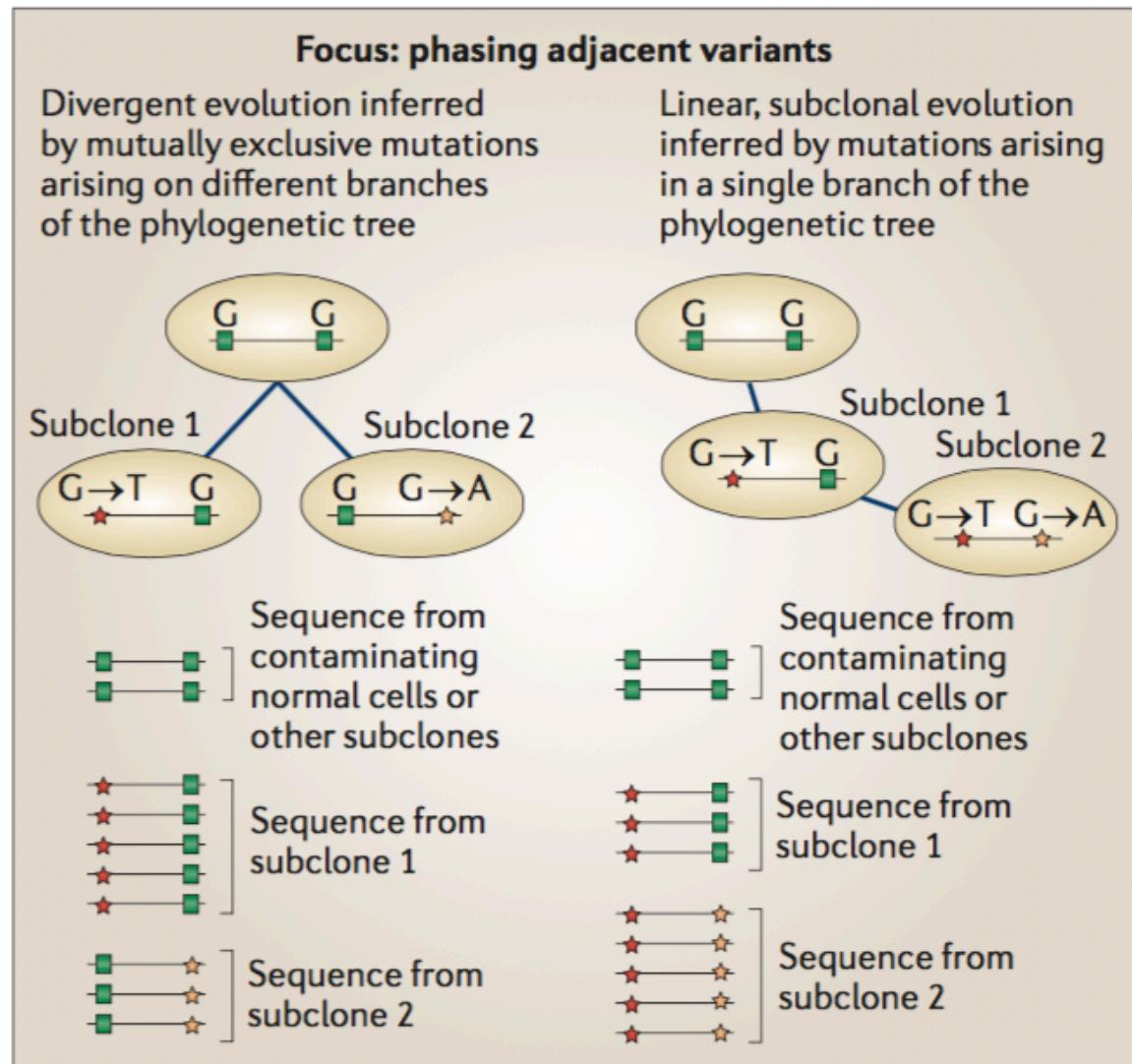
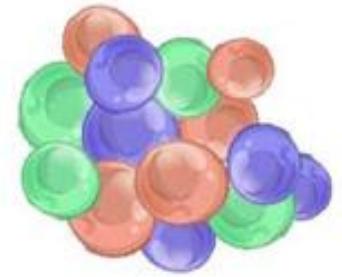
B



**THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data**

Oesper et al (2013) Genome Biology. DOI: 10.1186/gb-2013-14-7-r80

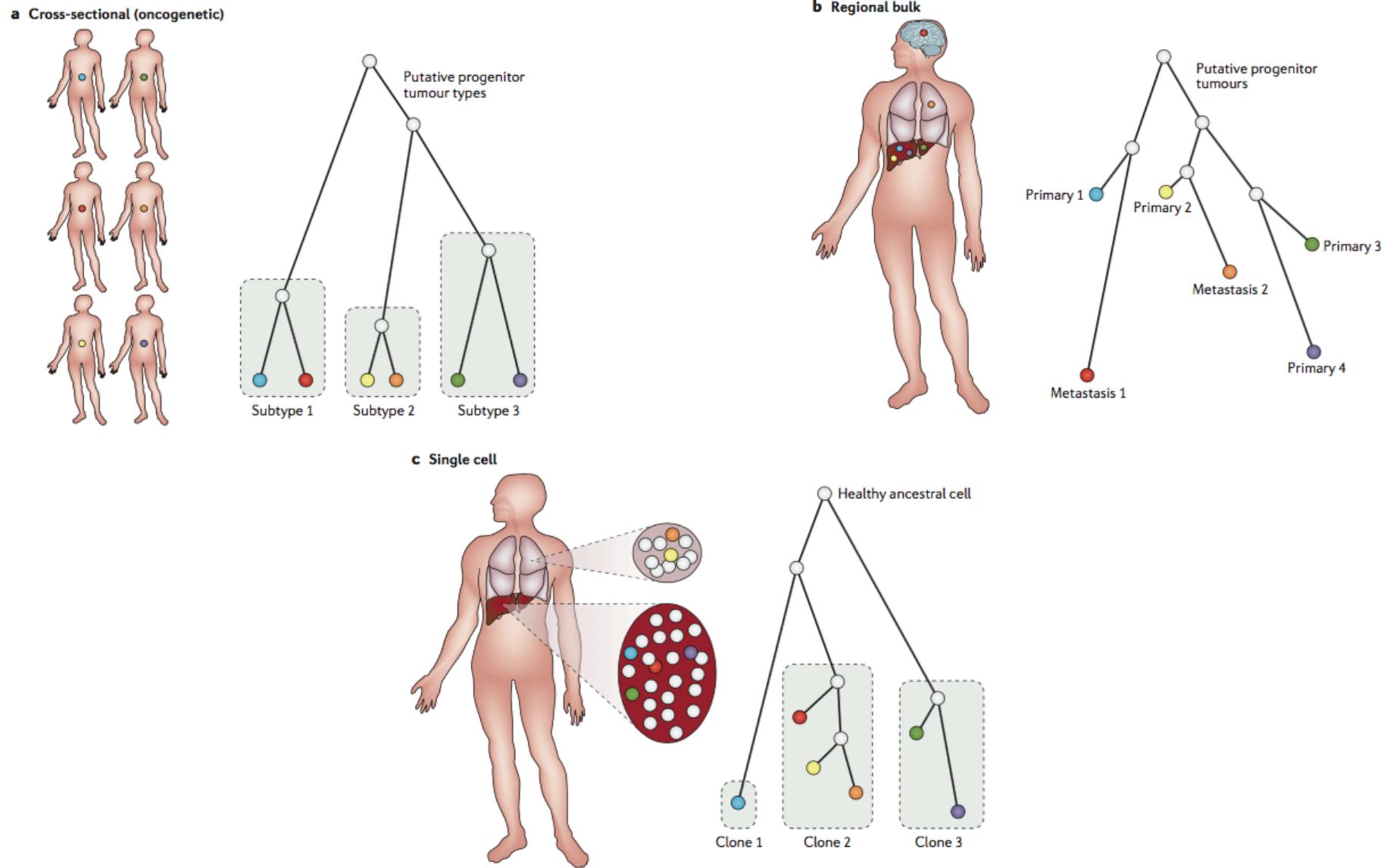
# Tumor Heterogeneity



## Evolution of the cancer genome

Yates & Campbell (2012) Nature Review Genetics. doi:10.1038/nrg3317

# Tumor Heterogeneity

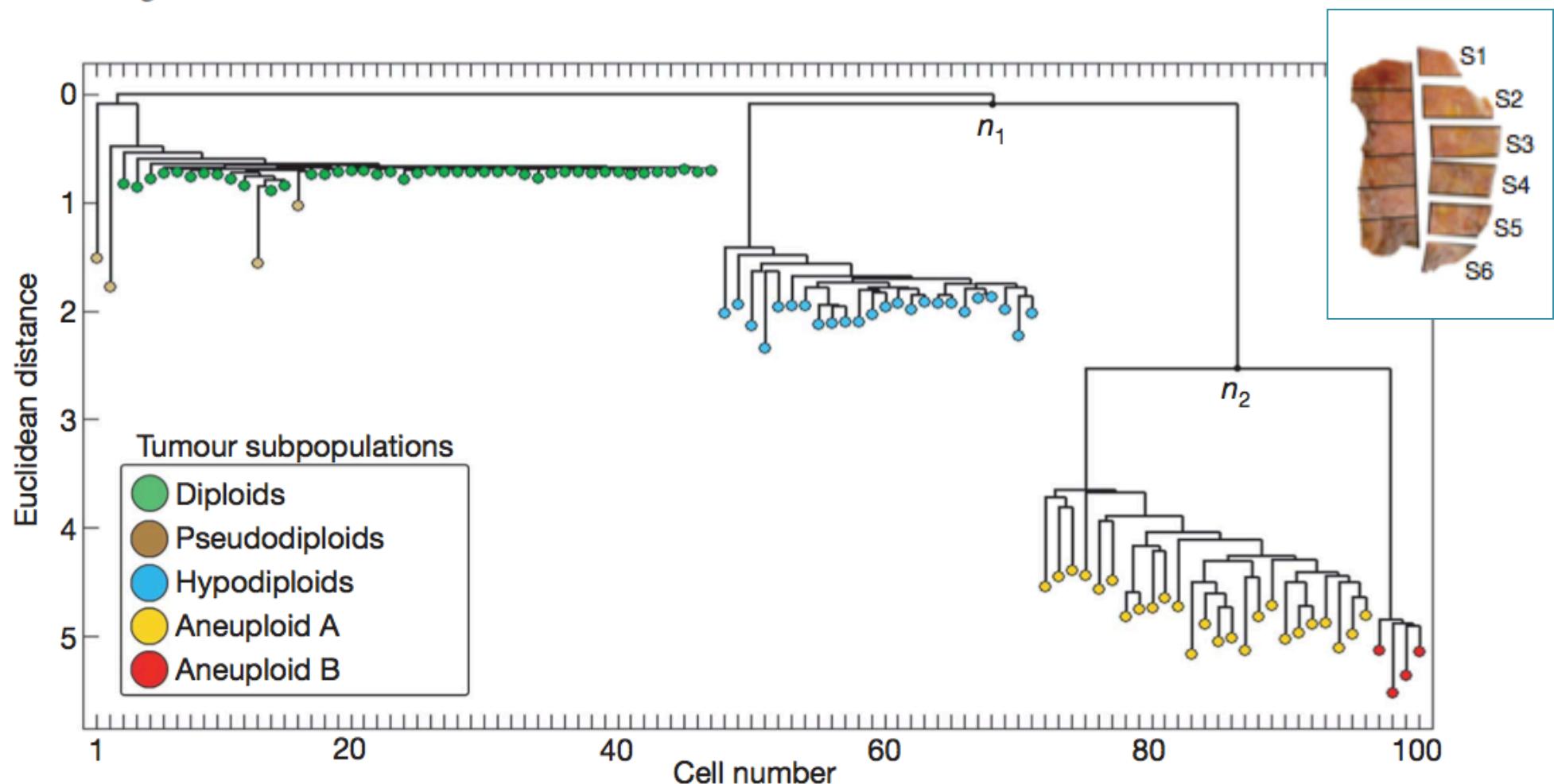


**The evolution of tumour phylogenetics: principles and practice**

Schwarz and Schaffer (2017) *Nature Reviews Genetics*. doi:10.1038/nrg.2016.170

# Tumour evolution inferred by single-cell sequencing

Nicholas Navin<sup>1,2</sup>, Jude Kendall<sup>1</sup>, Jennifer Troge<sup>1</sup>, Peter Andrews<sup>1</sup>, Linda Rodgers<sup>1</sup>, Jeanne McIndoo<sup>1</sup>, Kerry Cook<sup>1</sup>, Asya Stepansky<sup>1</sup>, Dan Levy<sup>1</sup>, Diane Esposito<sup>1</sup>, Lakshmi Muthuswamy<sup>3</sup>, Alex Krasnitz<sup>1</sup>, W. Richard McCombie<sup>1</sup>, James Hicks<sup>1</sup> & Michael Wigler<sup>1</sup>



# Gingko

<http://qb.cshl.edu/ginkgo>



## ***Interactive Single Cell CNV analysis & clustering***

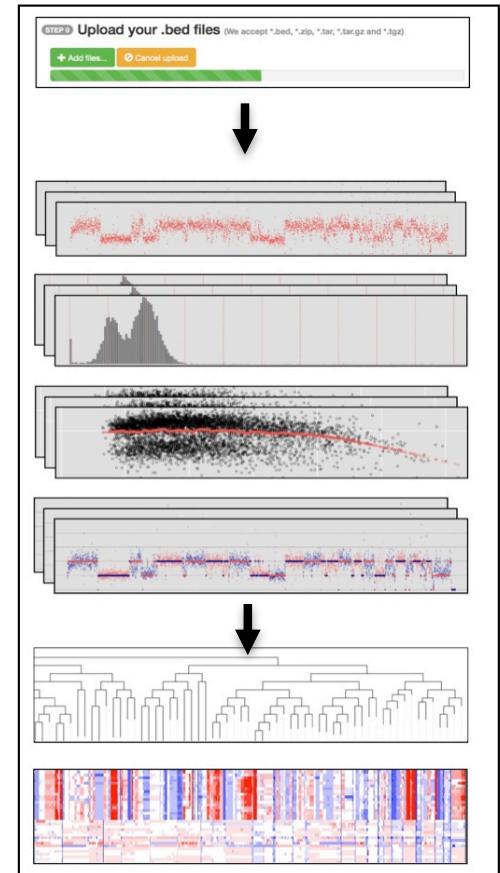
- Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
- Per cell through project-wide analysis in any species

## ***Compare MDA, DOP-PCR, and MALBAC***

- DOP-PCR shows superior resolution and consistency

## ***Available for collaboration***

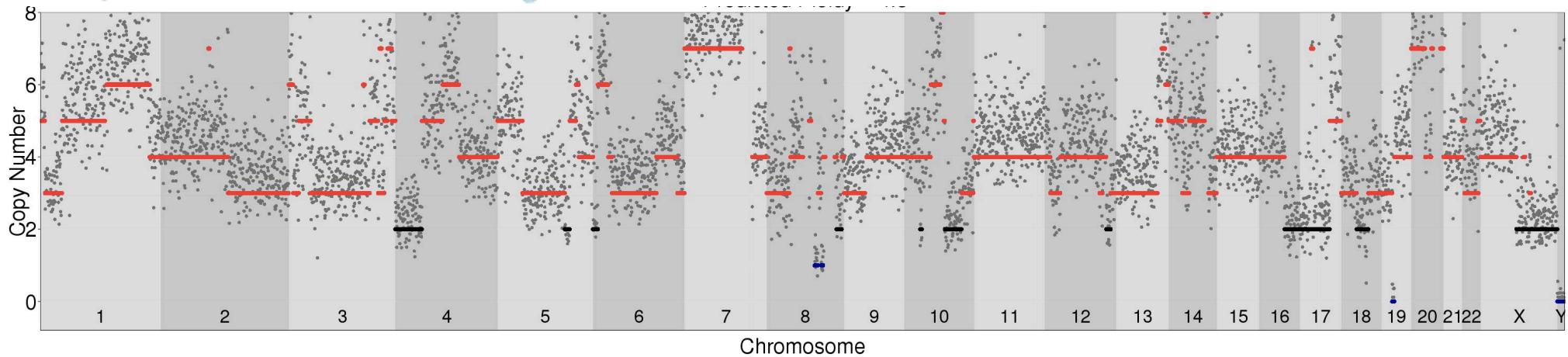
- Analyzing CNVs with respect to different clinical outcomes
- Extending clustering methods, prototyping scRNA



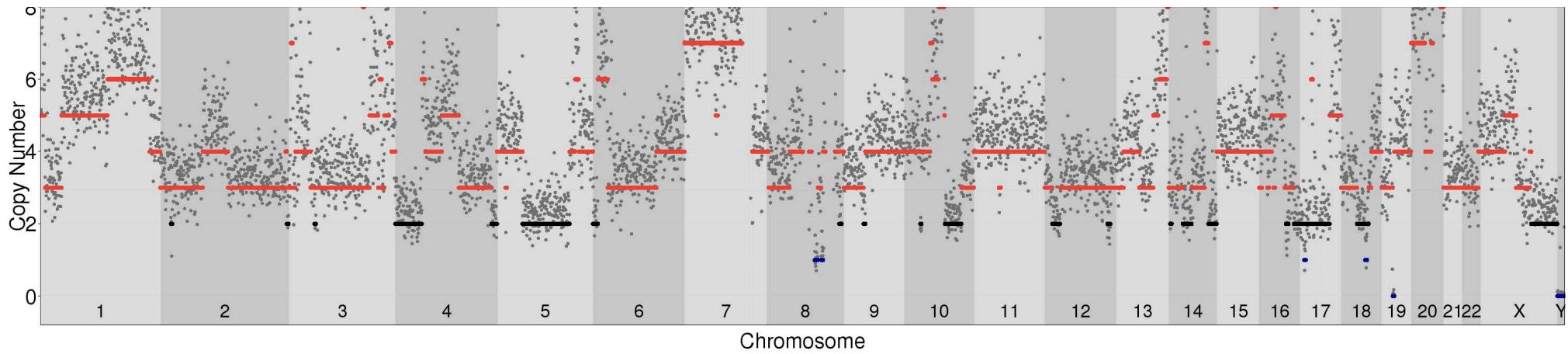
**Interactive analysis and assessment of single-cell copy-number variations.**

Garvin et al. (2015) Nature Methods doi:10.1038/nmeth.3578

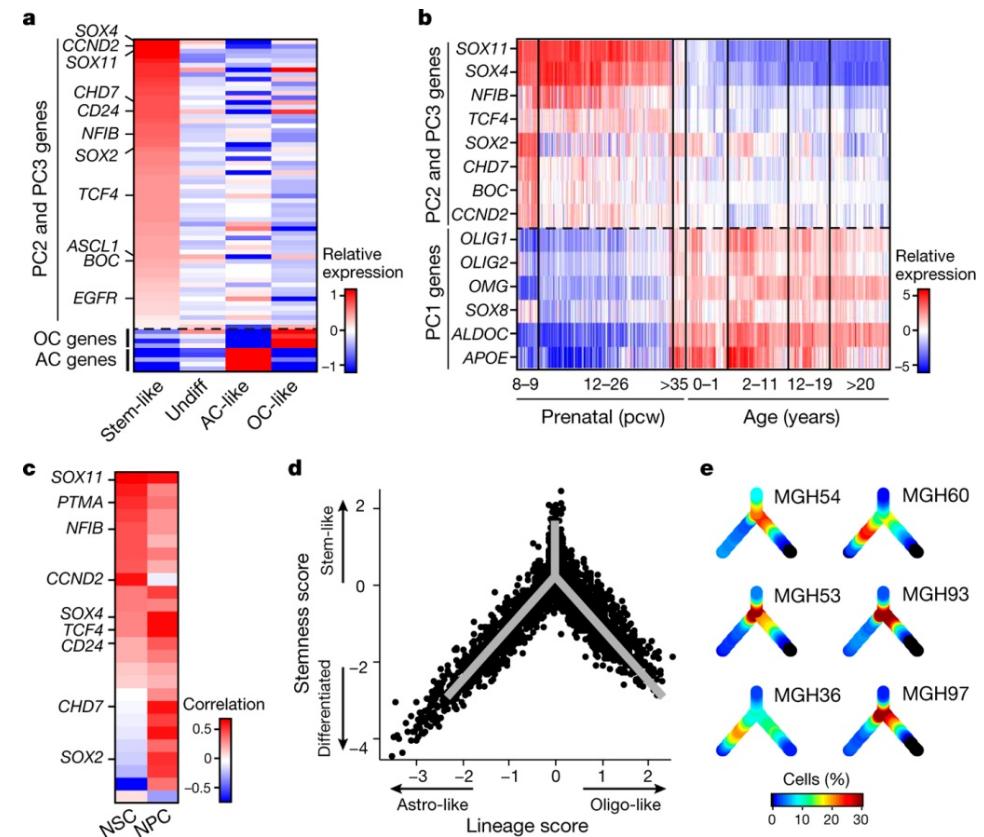
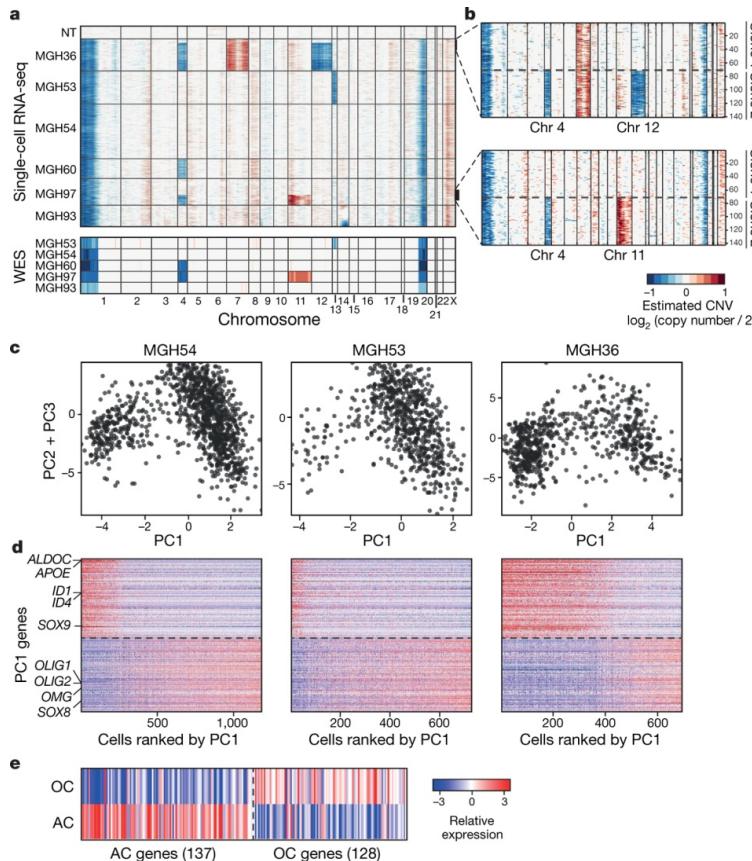
# Realtime CNV Analysis



illumina®

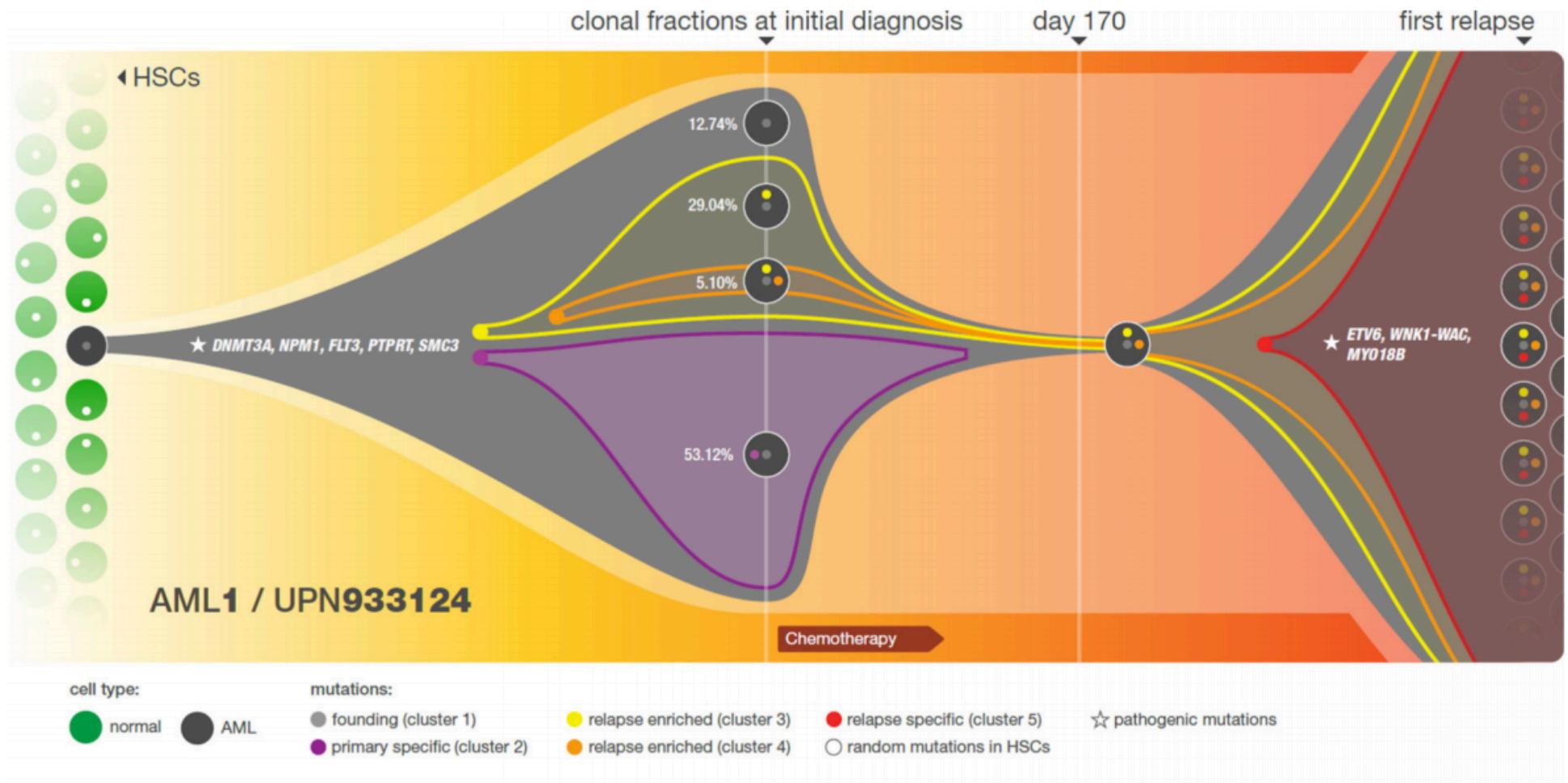


# Single Cell RNA-seq of Cancer



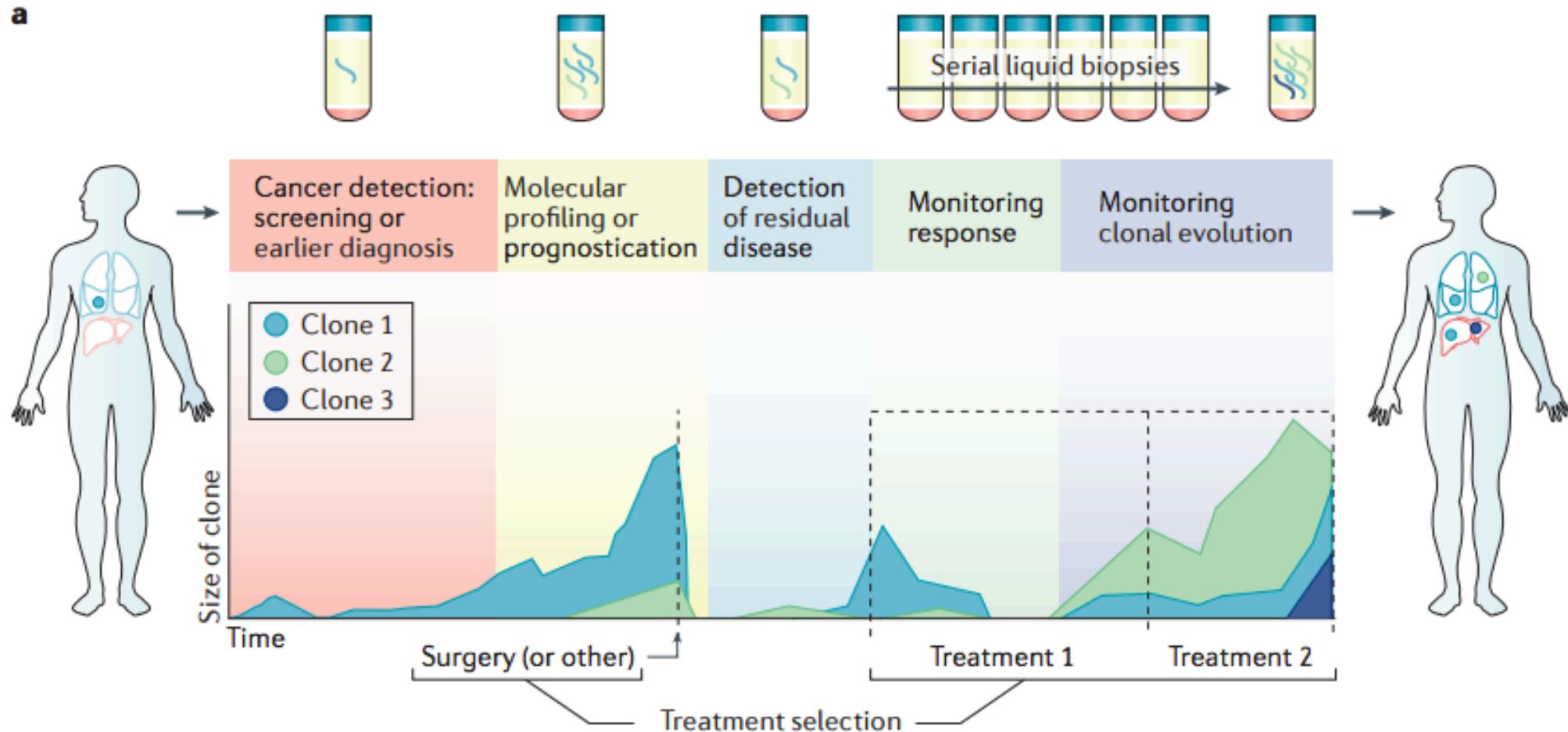
**Single-cell RNA-seq supports a developmental hierarchy in human oligodendrogloma**  
Tirosh et al (2016) Nature. doi:10.1038/nature20123

# Tumor Heterogeneity and Treatment



**Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing**  
Ding et al (2012) Nature. doi:10.1038/nature10738

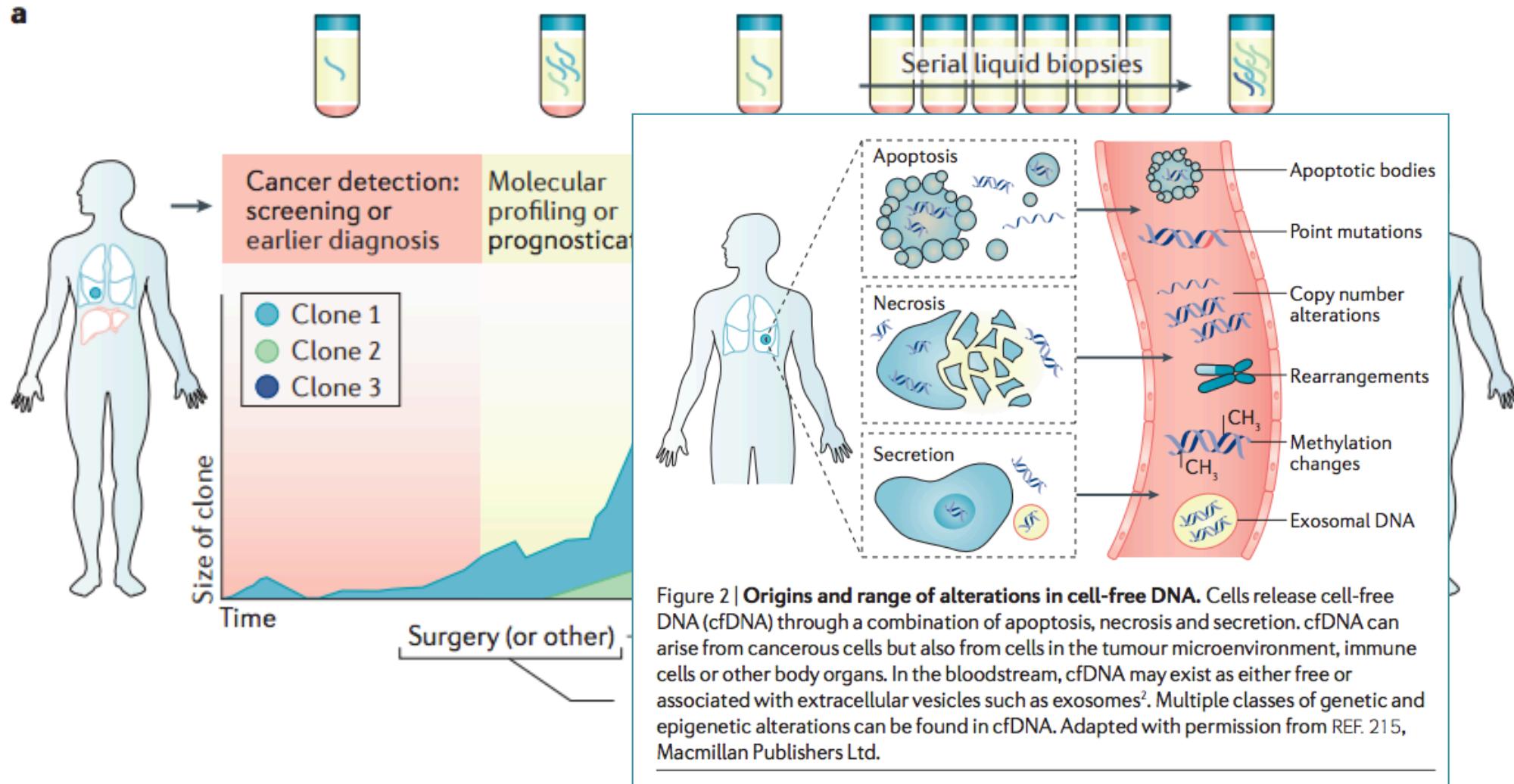
# Liquid Biopsies



Liquid biopsies come of age: towards implementation of circulating tumour DNA

Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7

# Liquid Biopsies

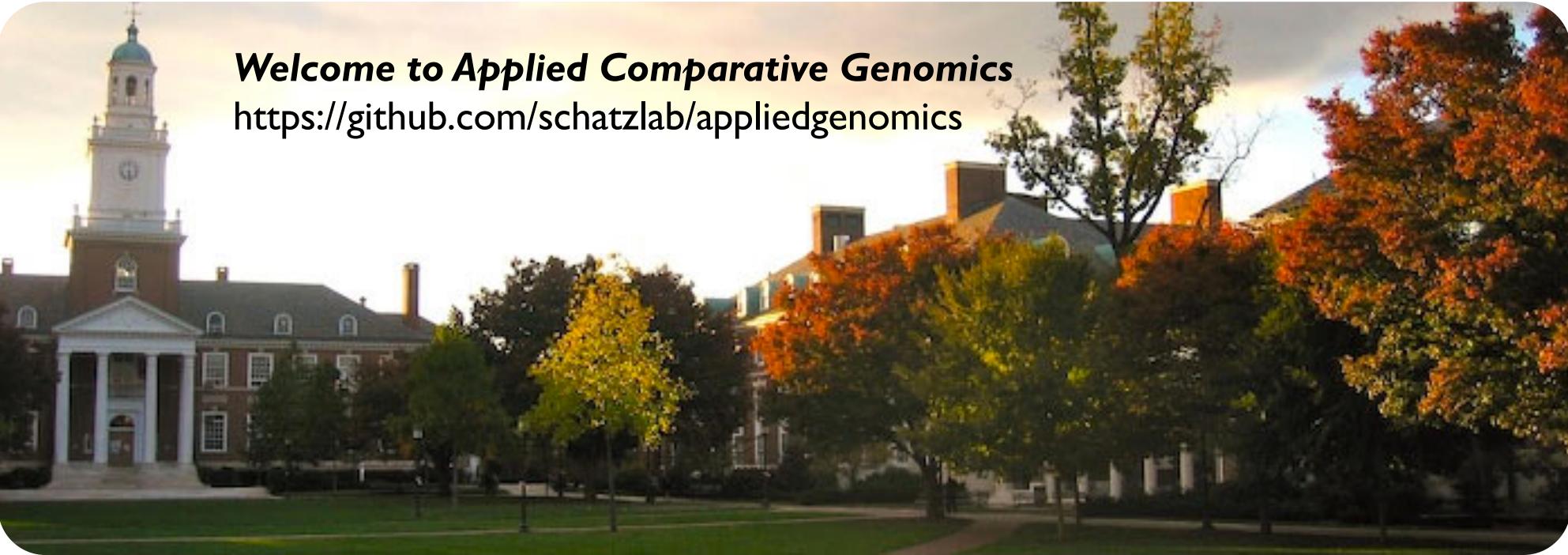


Liquid biopsies come of age: towards implementation of circulating tumour DNA

Wan et al (2017) Nature Review Cancer. doi:10.1038/nrc.2017.7

# Next Steps

1. Questions on project?
2. Check out the course webpage



**Welcome to Applied Comparative Genomics**

<https://github.com/schatzlab/appliedgenomics>

**Questions?**