

Lecture 22 Cancer Genomics

Michael Schatz

April 12, 2021

JHU 600.749: Applied Comparative Genomics



Preliminary Project Report

Assignment Date: March 24, 2021

Due Date: Monday, April 7, 2021 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to [GradeScope](#) by 11:59pm on Wednesday April 7.

The preliminary report should have at least:

- 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 preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result
- 5+ 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. [Overleaf](#) is recommended for LaTeX submissions. [Google Docs](#) is recommended for non-latex submissions, especially group projects. [Paperpile](#) is recommended for citation management.

Later, you will present your project in class starting the week of April 21. You will also submit your final written report (5-7 pages) of your project by May 13

Please use Piazza if you have any general questions!

/ JHU EN.600.749: Computational Genomics: Applied Comparative Genomics

Project Presentations

Presentations will be a total of 15 minutes: 12 minutes for the presentation, followed by 3 minutes for questions. We will strictly keep to the schedule to ensure that all groups can present in-class!

Schedule of Presentations

Day	Time	Team Name	Students	Title
Wed Apr 21	1:30-1:45	Team T-Cell	Hanzhi (Gary) Wang, Shao (Lynn) Yin	Sequence-Based Prediction of Cross-Reactivity in T Cell against SARS-CoV2 Epitopes
Wed Apr 21	1:45-2:00	dellertKare	Margaret Starostik, Katherine Jenike	Identifying RNA modifications in direct RNA sequencing data from Oxford Nanopore
Wed Apr 21	2:00-2:15	Daniel's Team	Daniel Burdick	Carnegie Project
Wed Apr 21	2:15-2:30	Team Blake	Blake Johnson	scCNS from scRNA
Wed Apr 21	2:30-2:45	Team Omar	Omar Ahmad	Developing a Realistic Nanopore Signal Simulator Using a Generative Adversarial Network
Mon Apr 26	1:30-1:45	The Contextualizer(s)	Theron Palmer	Single-cell alternative splicing analysis using non-negative matrix factorization and differentially expressed splice junctions.
Mon Apr 26	1:45-2:00	The Genome Pals	James Furuta, Solchiro Asami	Evaluation of Single-Cell Monocle Algorithm with Low Read Coverage
Mon Apr 26	2:00-2:15	SW	Sanku Panda, Ido Shinder, Natalie Wilson	Uncovering the Origin of Polyploidy in Coast Redwood
Mon Apr 26	2:15-2:30	Team Amy	Amy Gill	Variant analysis of a whole exome trio to investigate a heritable neurological phenotype
Mon Apr 26	2:30-2:45	Team Yuxin	Yuxin Wang	Pacific single-molecule fluorescence resonance energy transfer (FRET) analysis pipeline
Wed Apr 28	1:30-1:45	Team Metamutes	Bert Endygu, Yuchen Ge	Calling genetic variants from long-read RNA sequencing of SK-BR-3 breast cancer cell line
Wed Apr 28	1:45-2:00	Team Bohao	Bohao Tang	Robustness of Single-Cell Pseudo-time Reconstruction Methods
Wed Apr 28	2:00-2:15	Team Yash	Yash Senthil	Applying ML classification models to accurately predict cell-type among datasets using expression data from single-cell RNA seq experiments
Wed Apr 28	2:15-2:30	Team Maddy	Maddy Scott	Entropy of DNA

Recommended outline for your talk (~1 minute per slide):

1. Title Slide: Who are you, title, date
2. Intro 1: Whats the big idea???
3. Intro 2: More specifically, what are you trying to learn?
4. Methods 1: What did you try?
5. Methods 2: What is the key idea?
6. Data 1: What data are you looking at?
7. Data 2: Anything notable about the data?
8. Results 1: What did you see!
9. Results 2: Does it work?
10. Results 3: How does it compare to other methods/data/ideas?
11. Discussion 1: What did you learn from this study?
12. Discussion 2: What does this mean for the future?
13. Acknowledgements: Who helped you along the way?
14. Thank you!

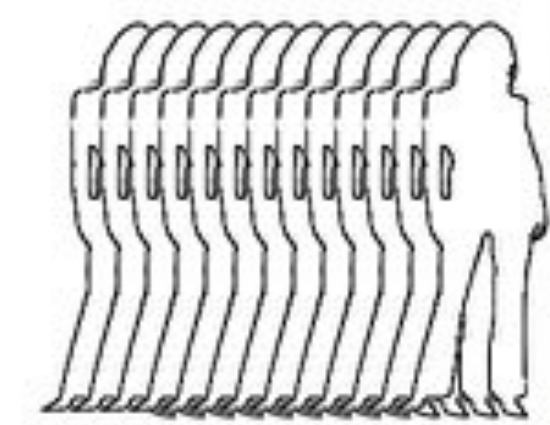
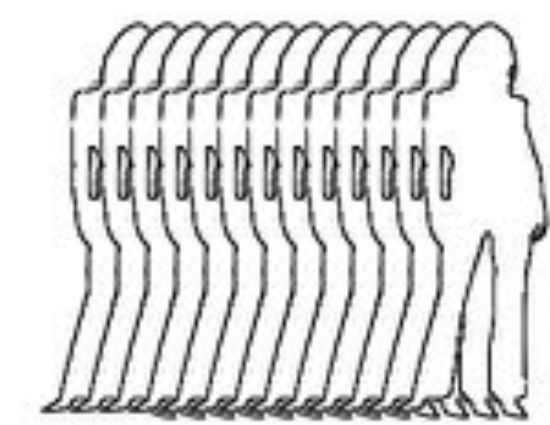
I strongly discourage you from trying to give a live demo as they are too unpredictable for a short talk. If you have running software you want to show, use a "cooking show" approach, where you have screen shots of the important steps.



Part I:

Genetic Medicine

Genome Wide Association (GWAS)

	SNP1	SNP2	SNP...
	Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	Cases Count of G: 1648 of 4000 Frequency of G: 41.2%	<i>Repeat for all SNPs</i>
	Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	Controls Count of G: 2532 of 6000 Frequency of G: 42.2%	
			

Are these significant
differences in frequencies?

Pearson's Chi-squared test

The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^n \frac{(O_i/N - p_i)^2}{p_i}$$

where

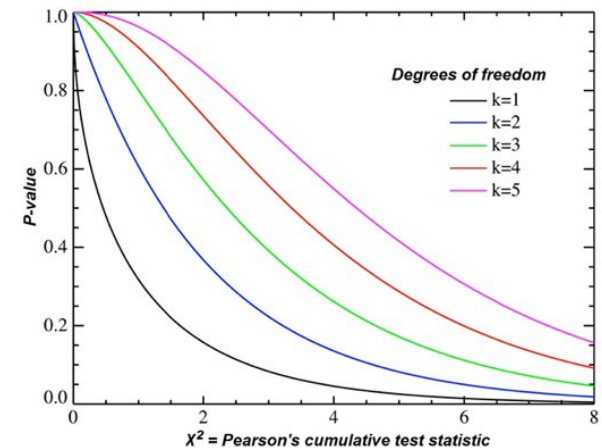
χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.

O_i = the number of observations of type i .

N = total number of observations

$E_i = Np_i$ = the expected (theoretical) frequency of type i , asserted by the null hypothesis that the fraction of type i in the population is p_i

n = the number of cells in the table.



$$P(\chi_P^2(\{p_i\}) > T) \sim C \int_{\sum_{i=1}^{m-1} y_i^2 > T} \left\{ \prod_{i=1}^{m-1} dy_i \right\} \prod_{i=1}^{m-1} \exp \left[-\frac{1}{2} \left(\sum_{i=1}^{m-1} y_i^2 \right) \right]$$

	has G	Not G	Marginal Row Totals
Cases	2104 (1912) [19.28]	1896 (2088) [17.66]	4000
Controls	2676 (2868) [12.85]	3324 (3132) [11.77]	6000
Marginal Column Totals	4780	5220	10000 (Grand Total)

Cases/hasG expected: $4000 * (4780/10000) = 1912$ expected

Cases/hasG squared deviation: $(2104 - 1912)^2 / 1912 = 19.28$ deviation

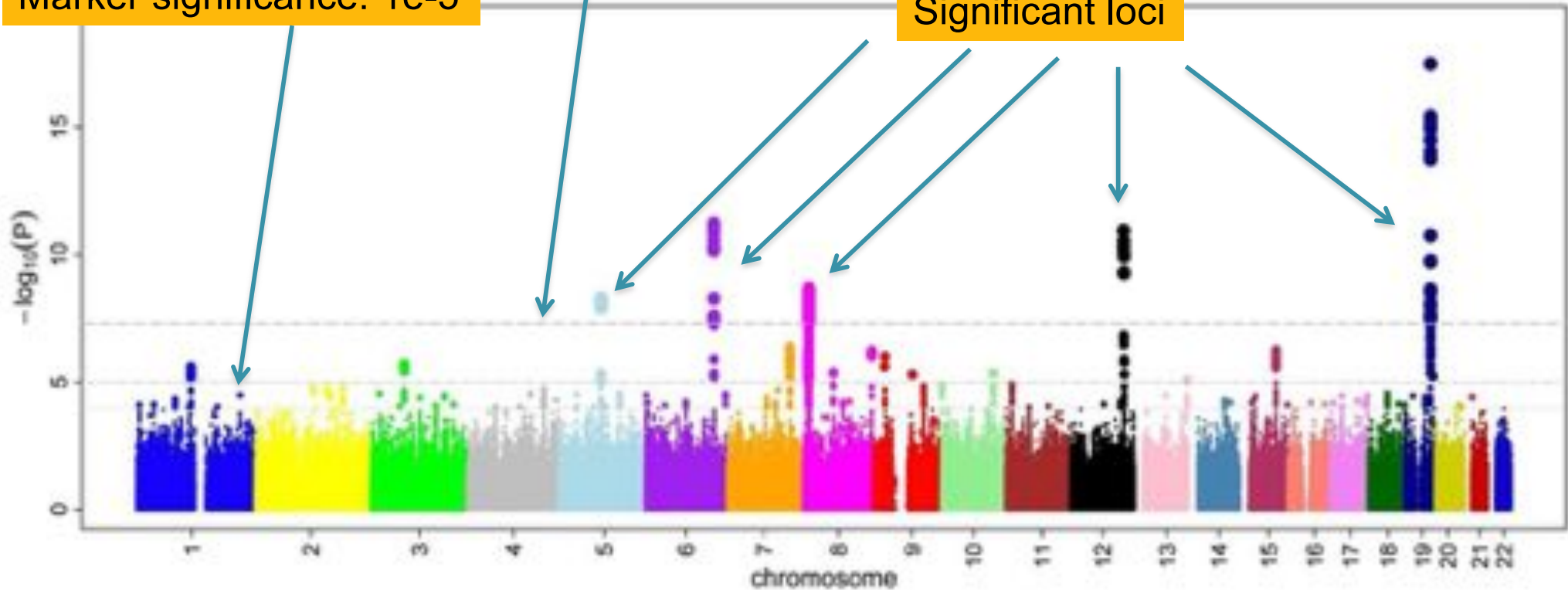
The chi-square statistic is $19.28 + 17.66 + 12.85 + 11.77 = 61.56$. The p-value is $5e-15$

Manhattan Plot

Genome-wide significance: $5e-8$

Marker significance: $1e-5$

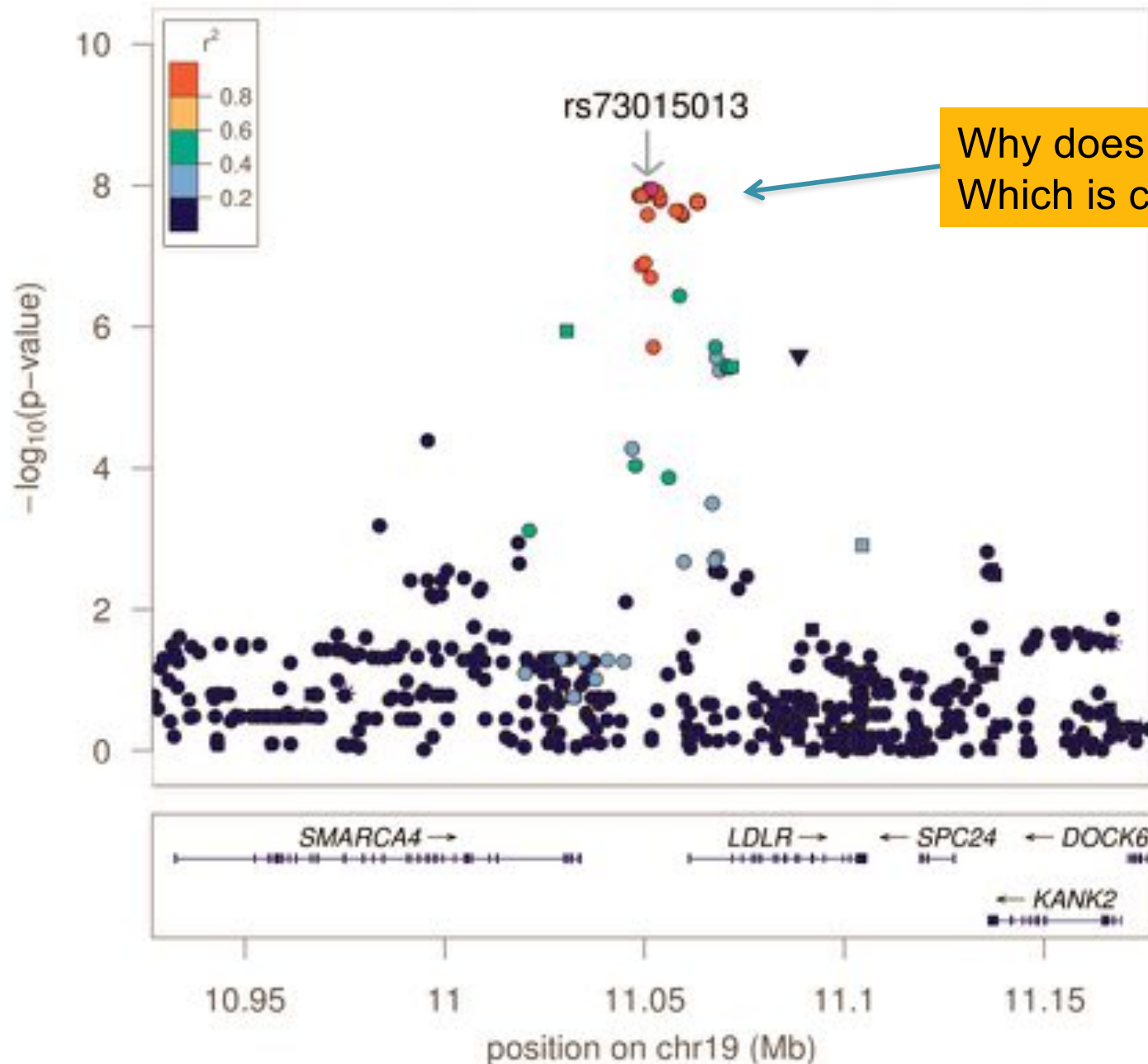
Significant loci



Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo

Ikram et al (2010) PLOS Genetics. doi: 10.1371/journal.pgen.1001184

Regional Association Plot



GWAS Catalog

As of 2021-03-25, the GWAS Catalog contains
4961 publications and 251,401 associations.



<http://www.ebi.ac.uk/gwas/diagram>

Biological insights from 108

schizophrenia

Schizophrenia W

Schizophrenia alleles of small phrenia genociations span previously reported findings. and several genes relevance to schizophrenia in brain, associated support for the

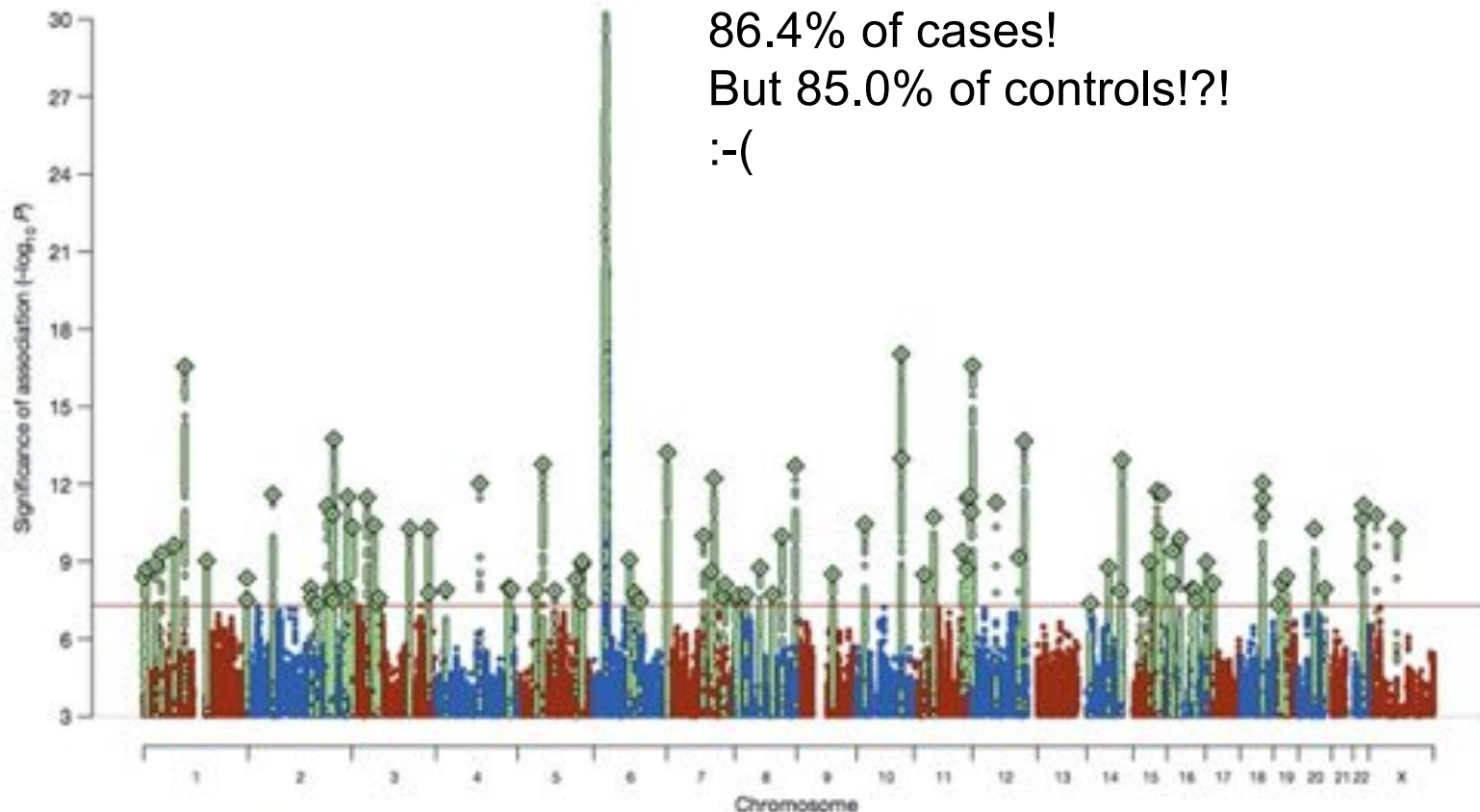


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The x axis is chromosomal

position and the y axis is the significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

rs115329265: A -> G substitution
86.4% of cases!
But 85.0% of controls!?!
:-(

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher^{1,5}, Daniela M Witten^{2,5}, Preti Jain^{3,4}, Brian J O’Roak^{1,4}, Gregory M Cooper³ & Jay Shendure¹

Current methods for annotating and interpreting human genetic variation tend to exploit a single information type (for example, conservation) and/or are restricted in scope (for example, to missense changes). Here we describe Combined Annotation-Dependent Depletion (CADD), a method for objectively integrating many diverse annotations into a single measure (C score) for each variant. We implement CADD as a support vector machine trained to differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants. We precompute C scores for all 8.6 billion possible human single-nucleotide variants and enable scoring of short insertions-deletions. C scores correlate with allelic diversity, annotations of functionality, pathogenicity, disease severity, experimentally measured regulatory effects and complex trait associations, and they highly rank known pathogenic variants within individual genomes. The ability of CADD to prioritize functional, deleterious and pathogenic variants across many functional categories, effect sizes and genetic architectures is unmatched by any current single-annotation method.

comparable, making it difficult to evaluate the relative importance of distinct variant categories or annotations. Third, annotation methods trained on known pathogenic mutations are subject to major ascertainment biases and may not be generalizable. Fourth, it is a major practical challenge to obtain, let alone to objectively evaluate or combine, the existing panoply of partially correlated and partially overlapping annotations; this challenge will only increase in size as large-scale projects such as the Encyclopedia of DNA Elements (ENCODE)¹¹ continually increase the amount of relevant data available. The net result of these limitations is that many potentially relevant annotations are ignored, while the annotations that are used are applied and combined in *ad hoc* and subjective ways that undermine their usefulness.

Here we describe a general framework, Combined Annotation-Dependent Depletion (CADD), for integrating diverse genome annotations and scoring any possible human single-nucleotide variant (SNV) or small insertion-deletion (indel) event. The basis of CADD is to contrast the annotations of fixed or nearly fixed derived alleles in humans with those of simulated variants. Deleterious variants—that is, variants that reduce organismal fitness—are depleted by natural selection in fixed but not simulated variation. CADD therefore

CADD Key Idea: Evaluate amino acid substitutions AND allele frequencies in 1000 genomes project AND ENCODE regions AND ... (63 annotations total :)

Genetic Basis of Autism Spectrum Disorders



Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

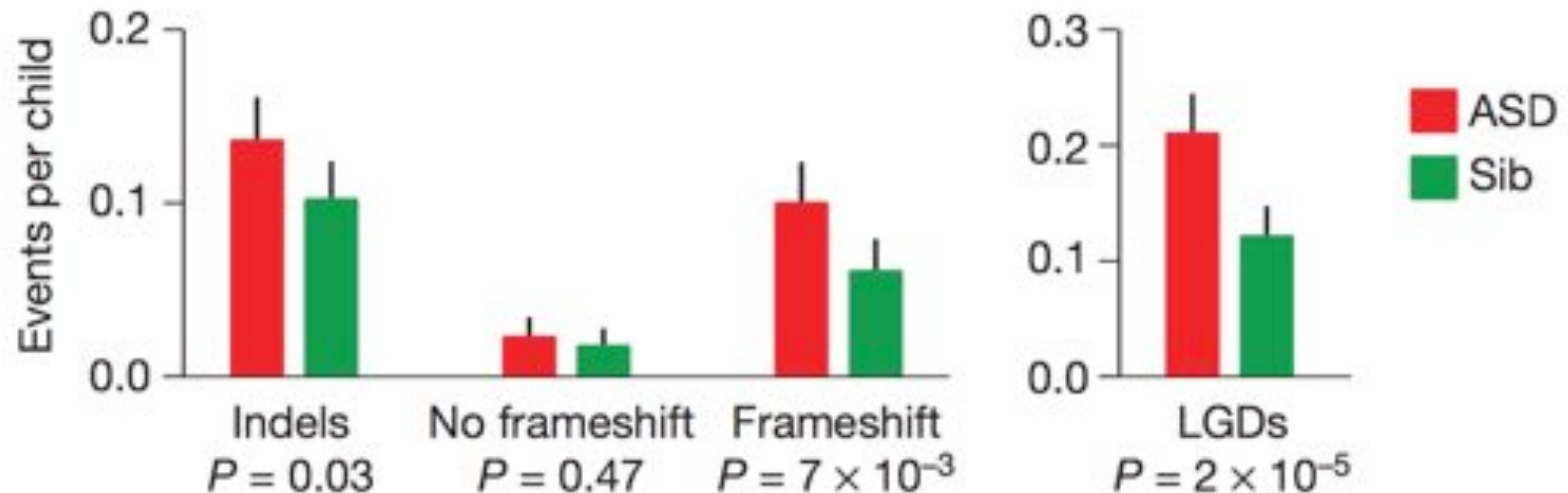
U.S. CDC identify around 1 in 68 American children as on the autism spectrum

- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

What is Autism?

<https://autisticadvocacy.org/about-asan/about-autism/>

De novo Genetics of Autism



- In 2,500 family quads we see significant enrichment in de novo **likely gene disruptions (LGDs)** in the autistic children
 - Overall rate of de novo mutations basically 1:1
 - 2:1 enrichment in frameshift indels, nonsense mutations
 - Contributed dozens of new autism candidate genes, highly enriched for neuron development or chromatin modifiers

The contribution of de novo coding mutations to autism spectrum disorder
lossifov et al (2014) *Nature*. doi:10.1038/nature13908



Part 2: Cancer Genetics & Genomics

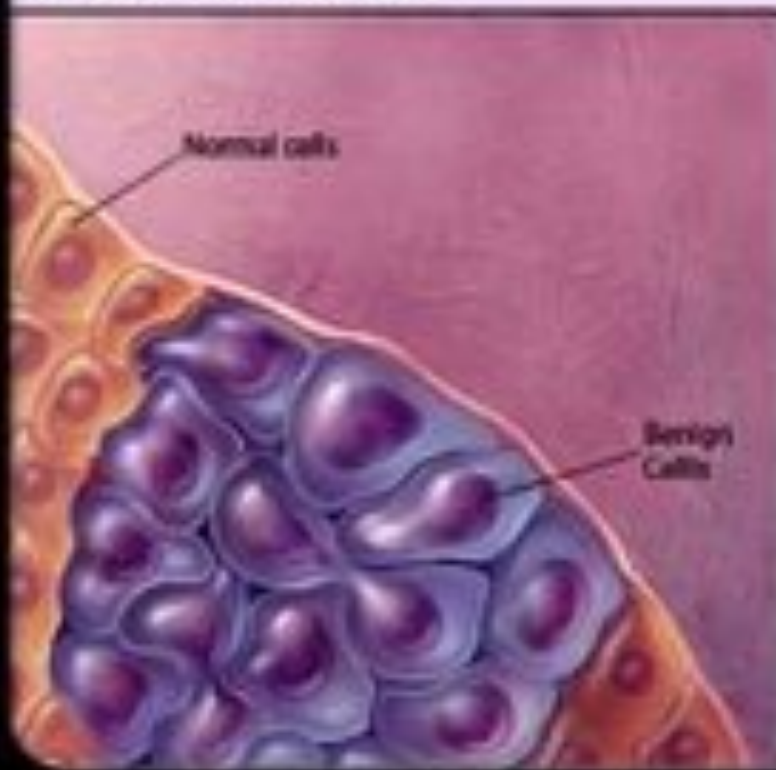


A tumor removed by surgery in 1689.

Benign vs. Malignant

Benign vs. Malignant Tumors

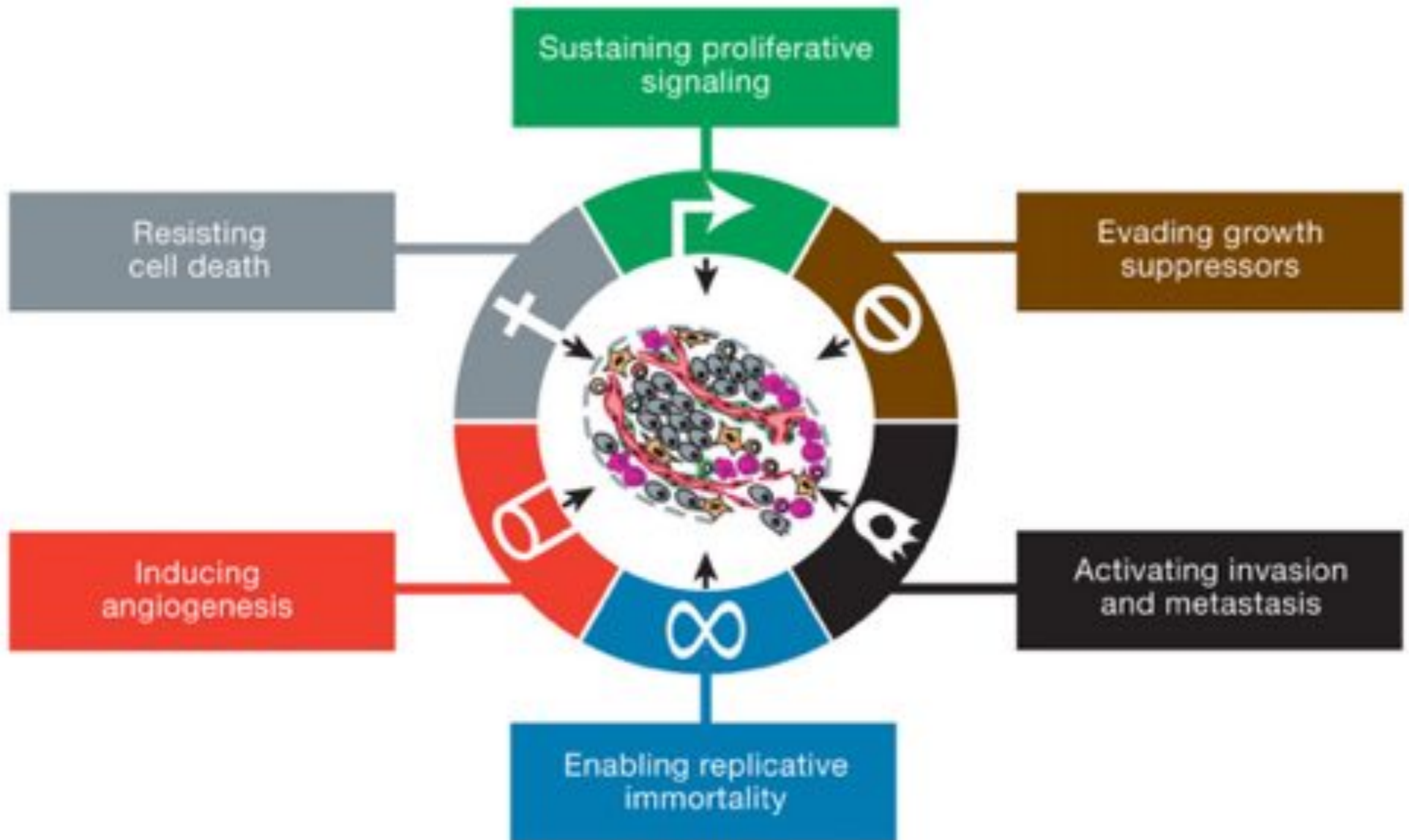
Benign (not cancer) tumor cells grow only locally and cannot spread by invasion or metastasis



Malignant (cancer) cells invade neighboring tissues, enter blood vessels, and metastasize to different sites



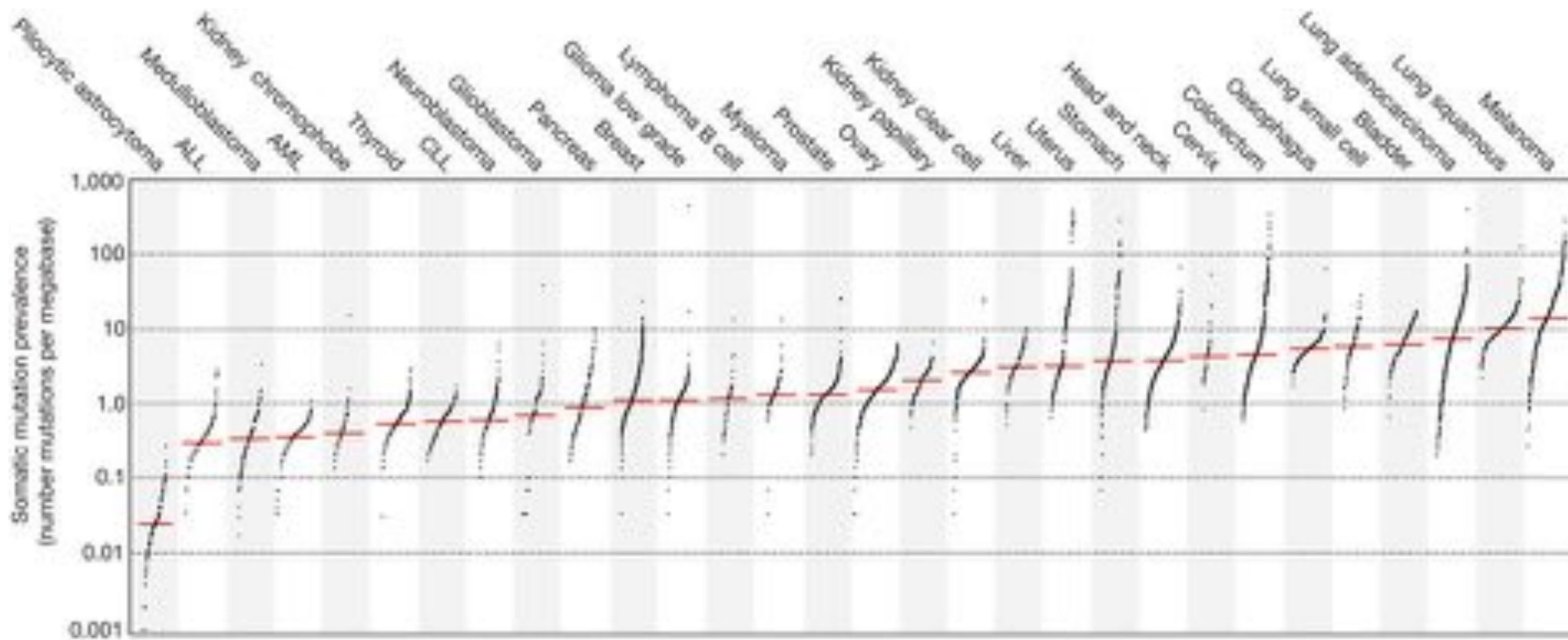
The Six Hallmarks of Cancer



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)

Somatic Mutations In Cancer

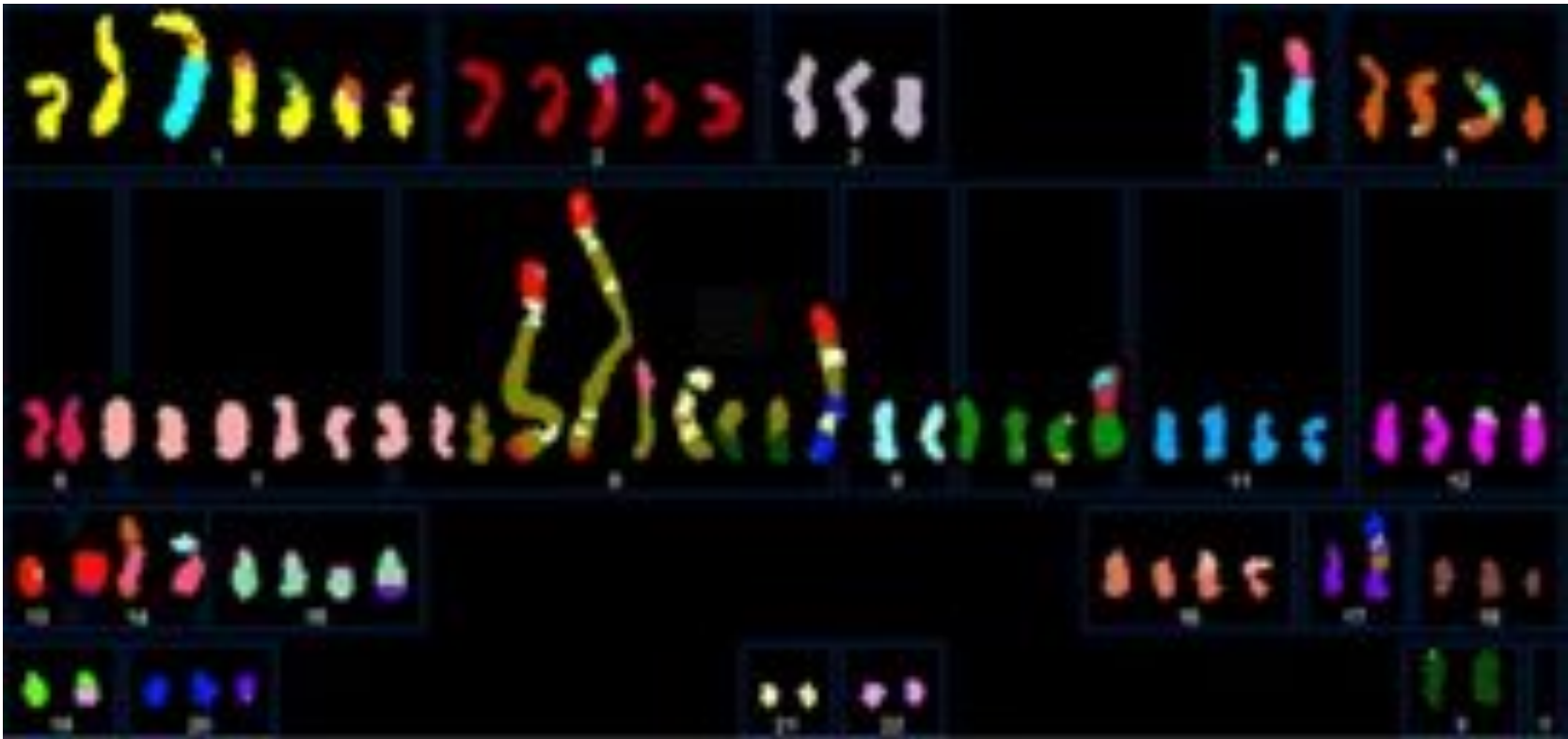


Signatures of mutational processes in human cancer

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

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

A firestorm in cancer

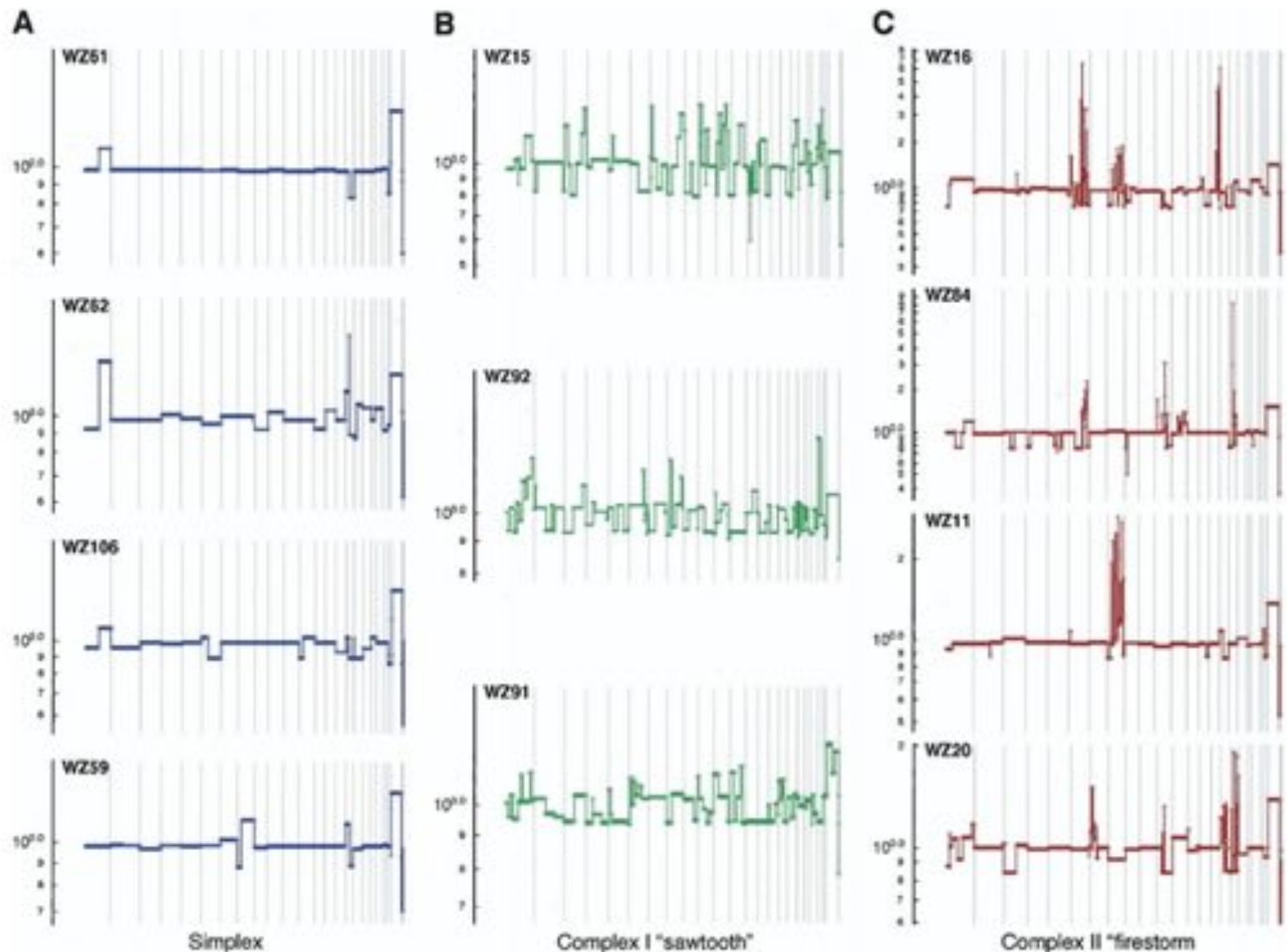
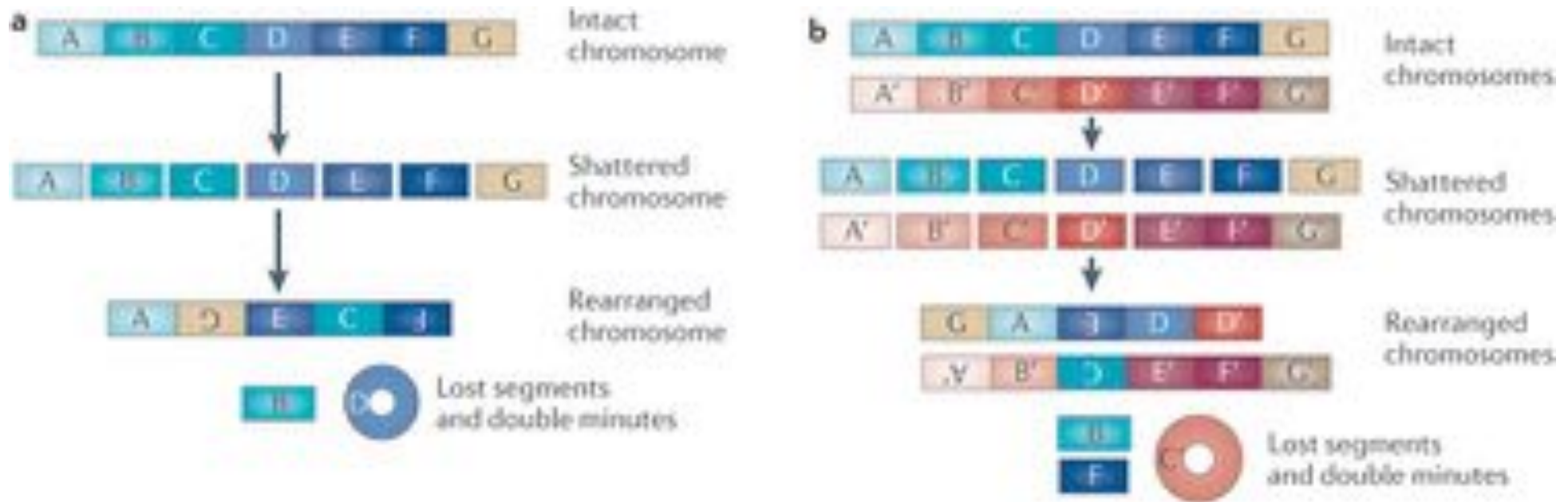


Figure 2. Major types of tumor genomic profiles. Segmentation profiles for individual tumors representing each category: (A) simplex; (B) complex type I or sawtooth; (C) complex type II or firestorm. Scored events consist of a minimum of six consecutive probes in the same state. The y-axis displays the 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

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^{1,2}, rearrangements^{3,4} and changes in methylation⁵⁻⁸ 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^{8,12-20}. The results of these studies have varied; depending on the techniques and systems used, an increase¹²⁻¹⁹, decrease²⁰⁻²⁴, or no change²⁵⁻²⁹ 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.

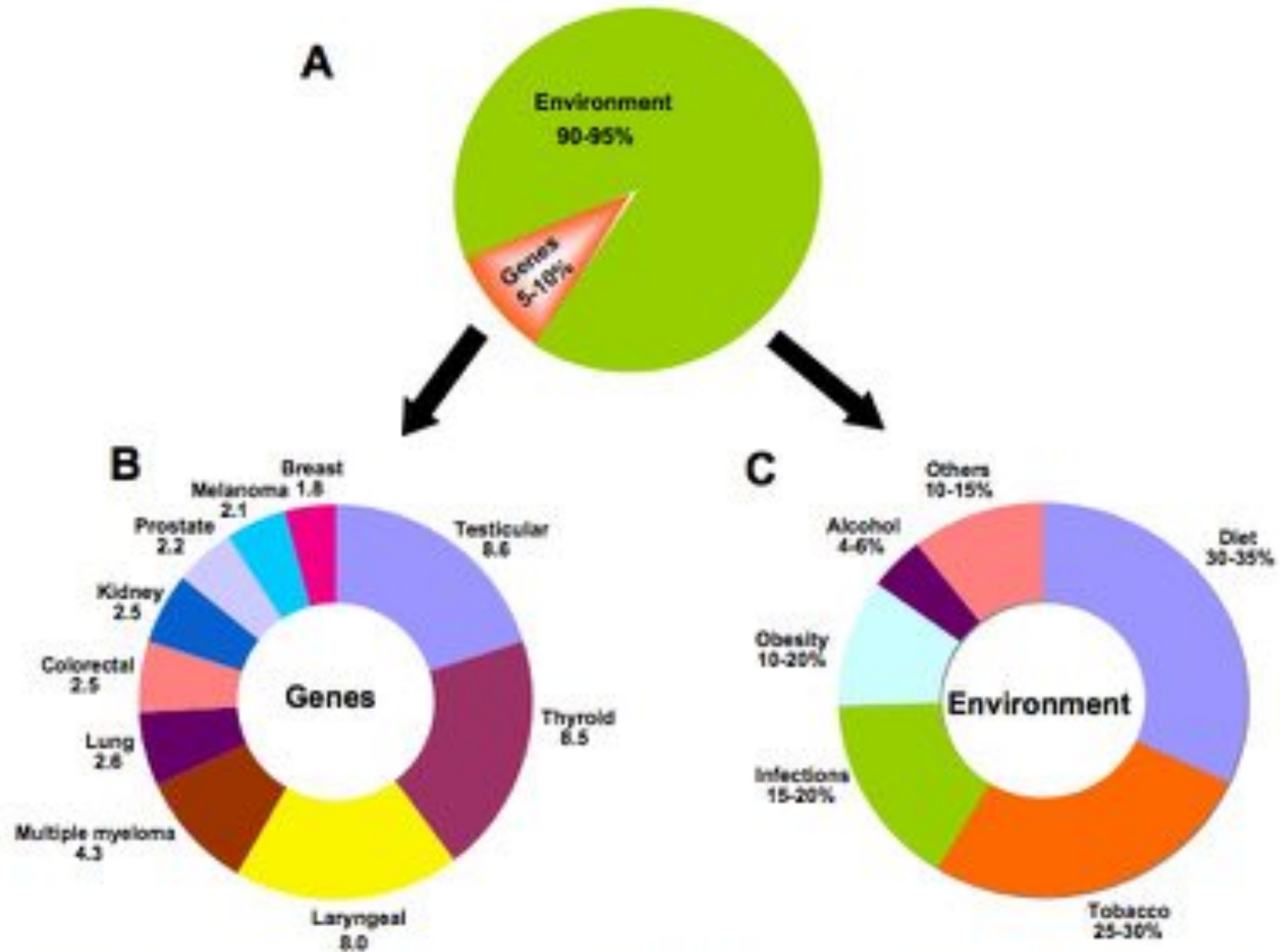
and (3) *Hpa*II and *Hha*I cleavage sites should be present in the regions of the genes.

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

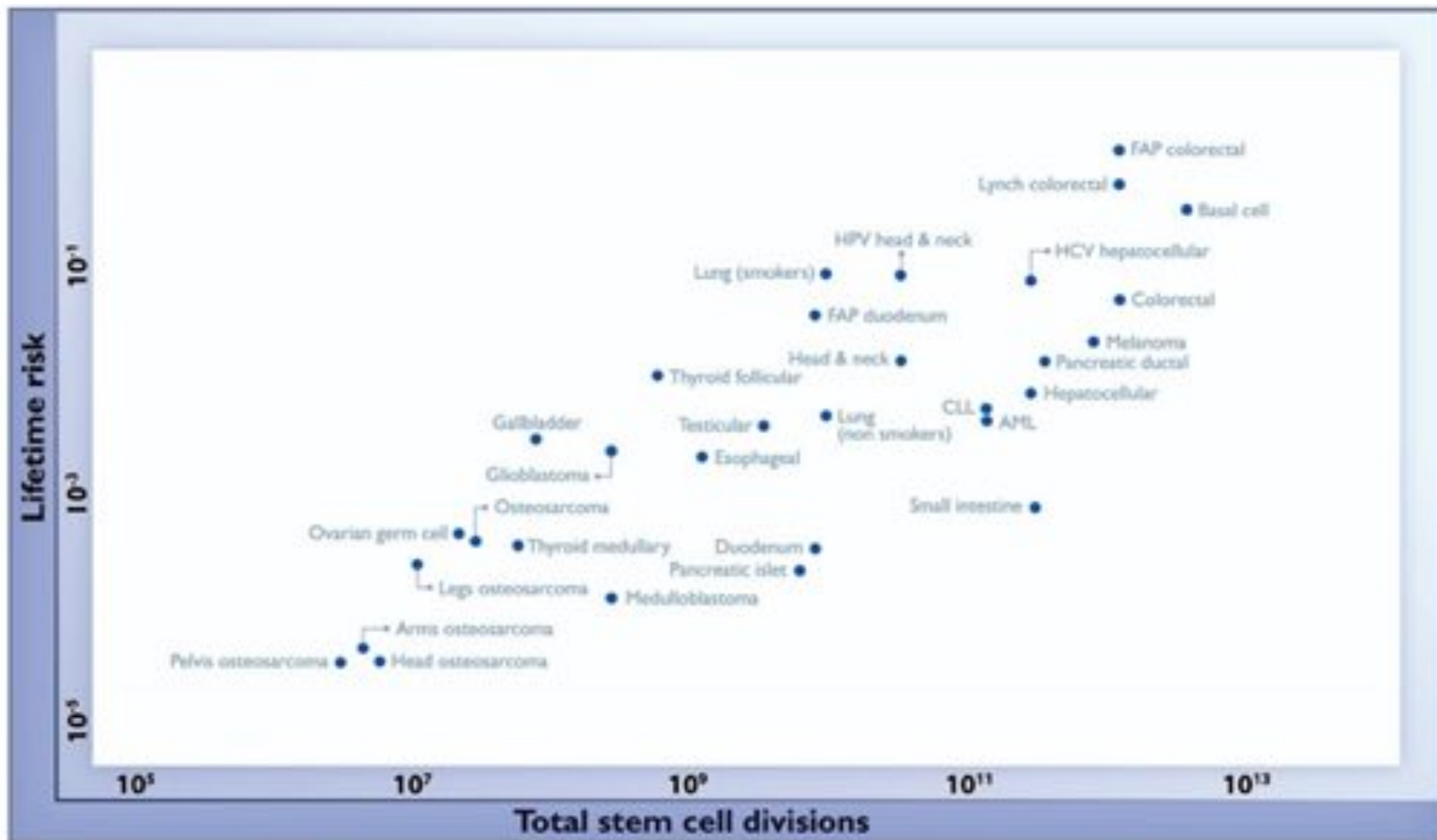
Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	{ <i>Hpa</i> II	<10	35	—
			{ <i>Hha</i> I	<10	39	—
		γ -Globin	{ <i>Hpa</i> II	<10	52	—
			{ <i>Hha</i> I	<10	39	—
		α -Globin	{ <i>Hpa</i> II	<10	<10	—
			{ <i>Hha</i> I	<10	<10	—
2	Colon	HGH	{ <i>Hpa</i> II	<10	76	—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

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

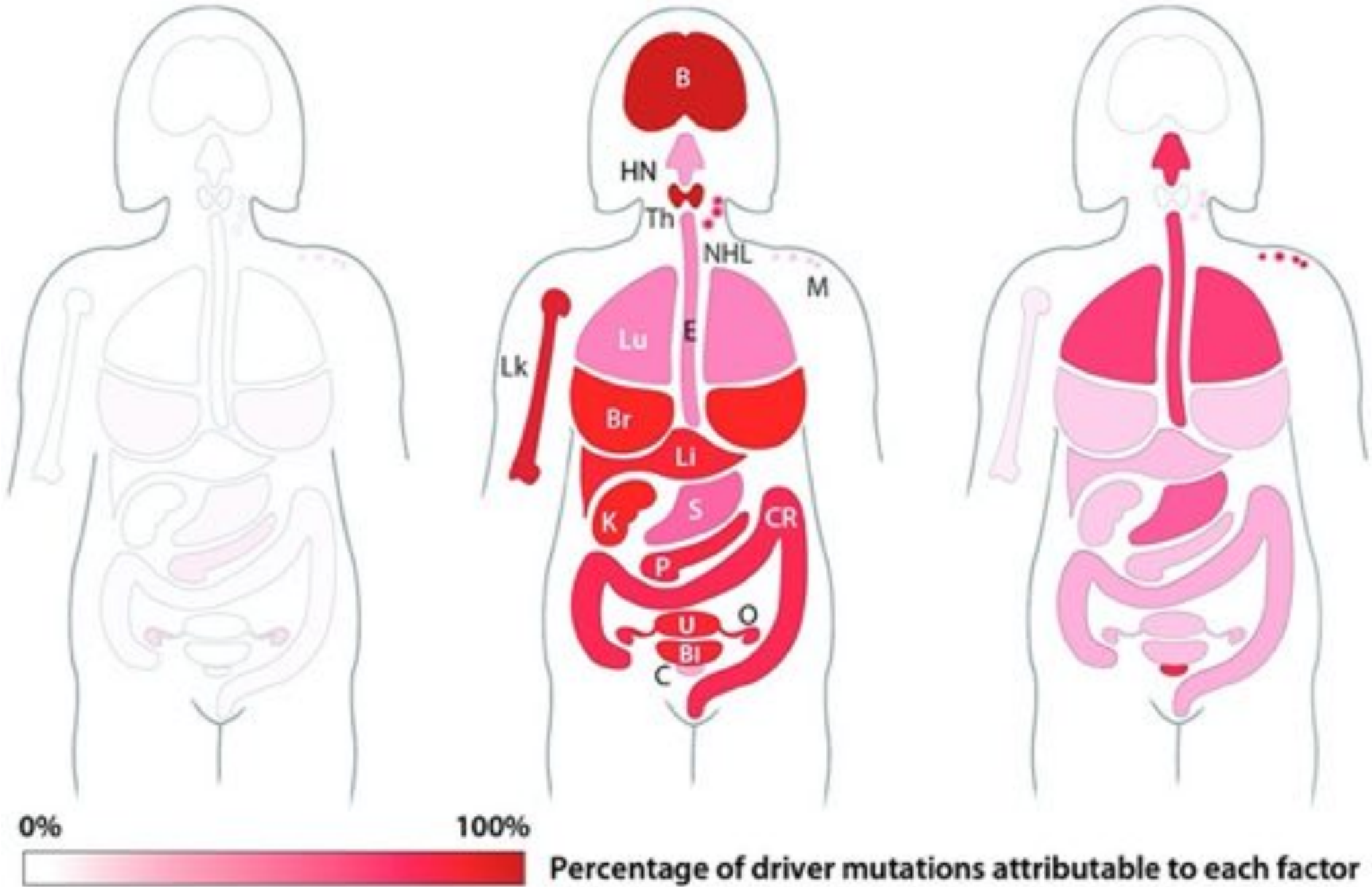
Fig. 1. The relationship between the number of stem cell divisions in the lifetime of a given tissue and the lifetime risk of cancer in that tissue. Values are from table S1, the derivation of which is discussed in the supplementary materials.

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

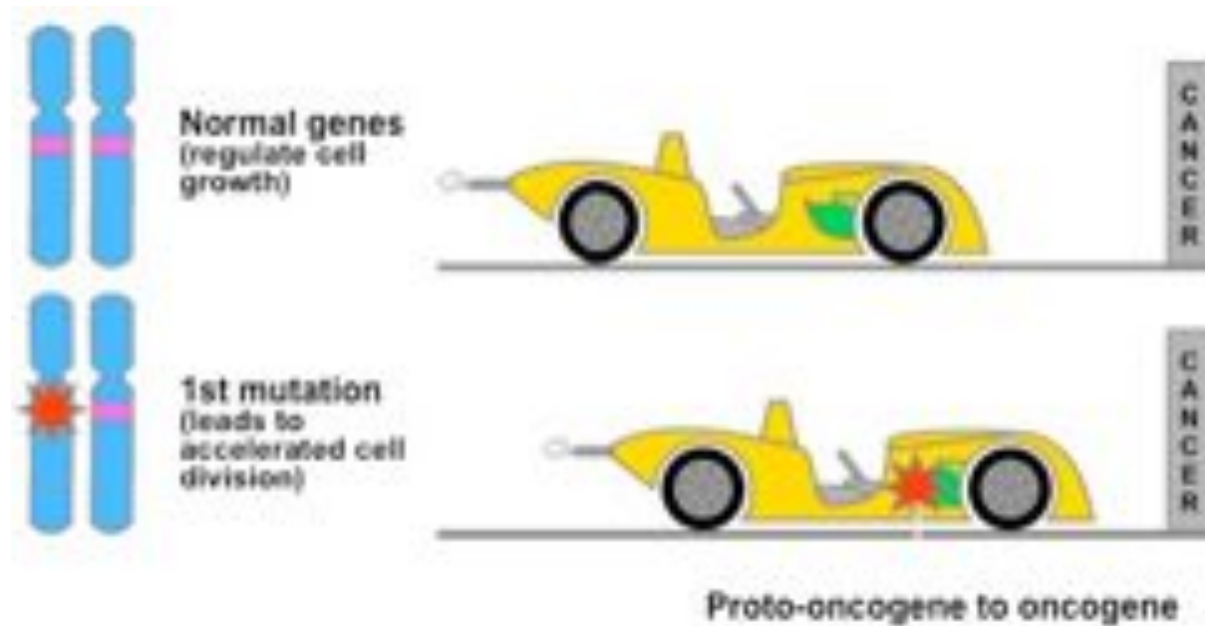
Environmental



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

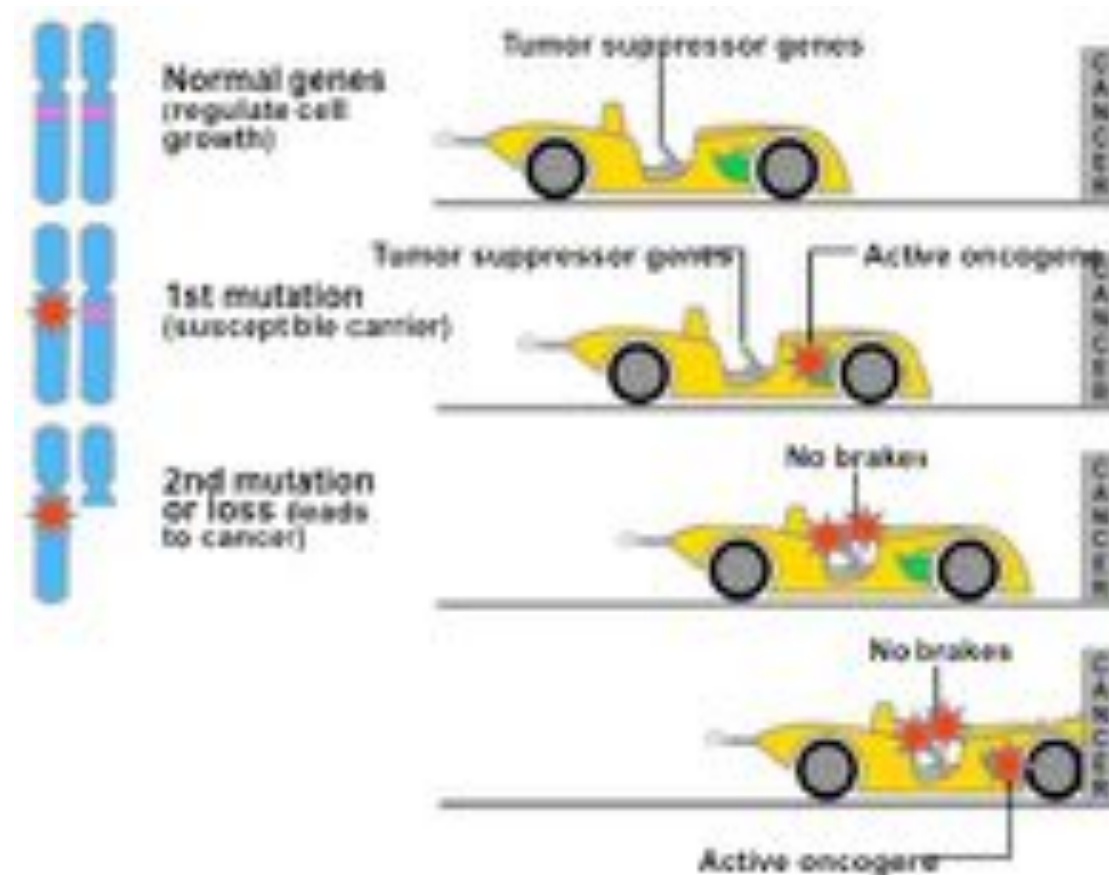
Tomasetti, Li, and Vogelstein (2017) Science. DOI: 10.1126/science.aaf9011

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



- **TP53:** 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 ^{ARF} 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.

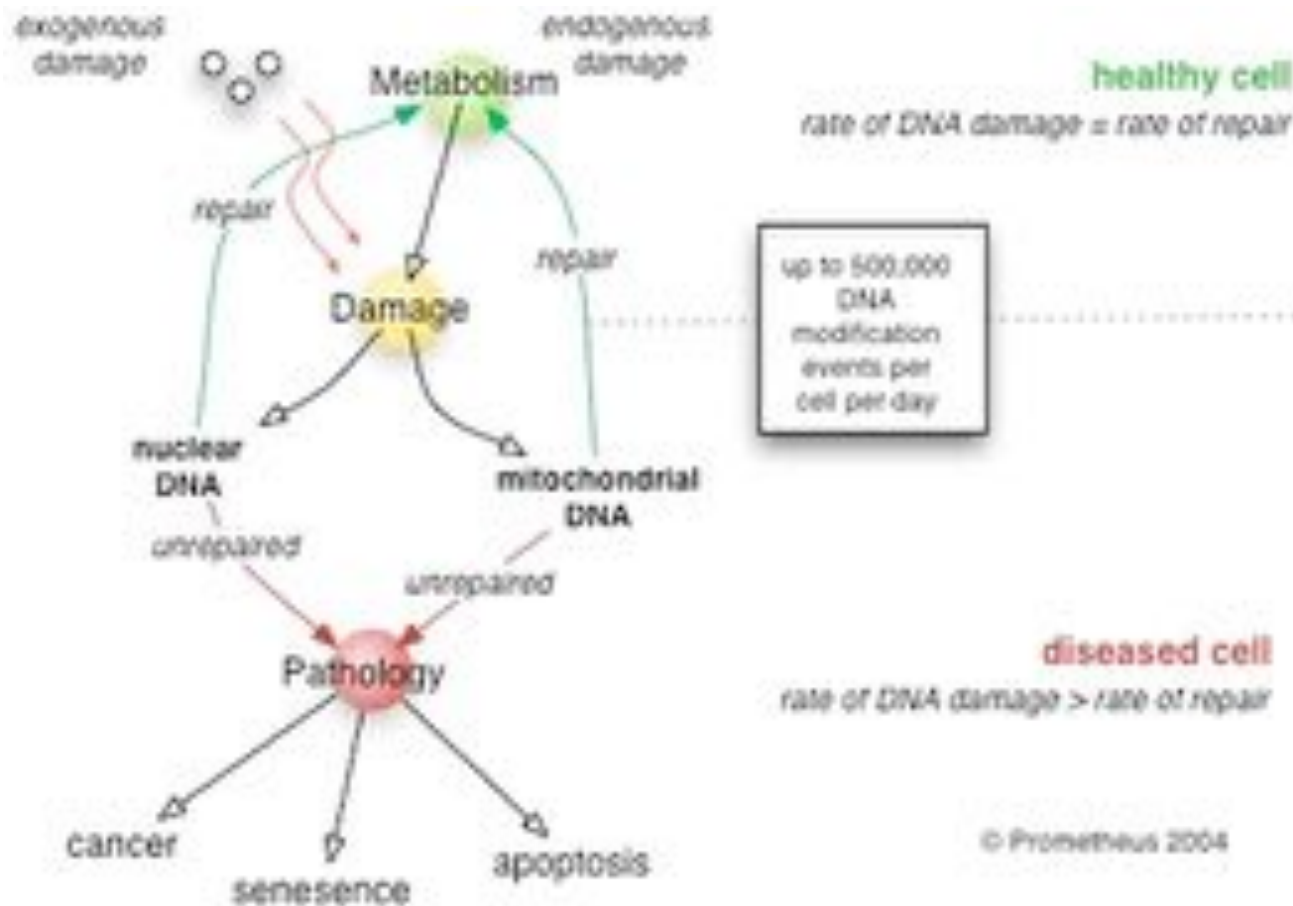
>10,000 known mutations

>17,000 publications

Surfing the p53 network

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

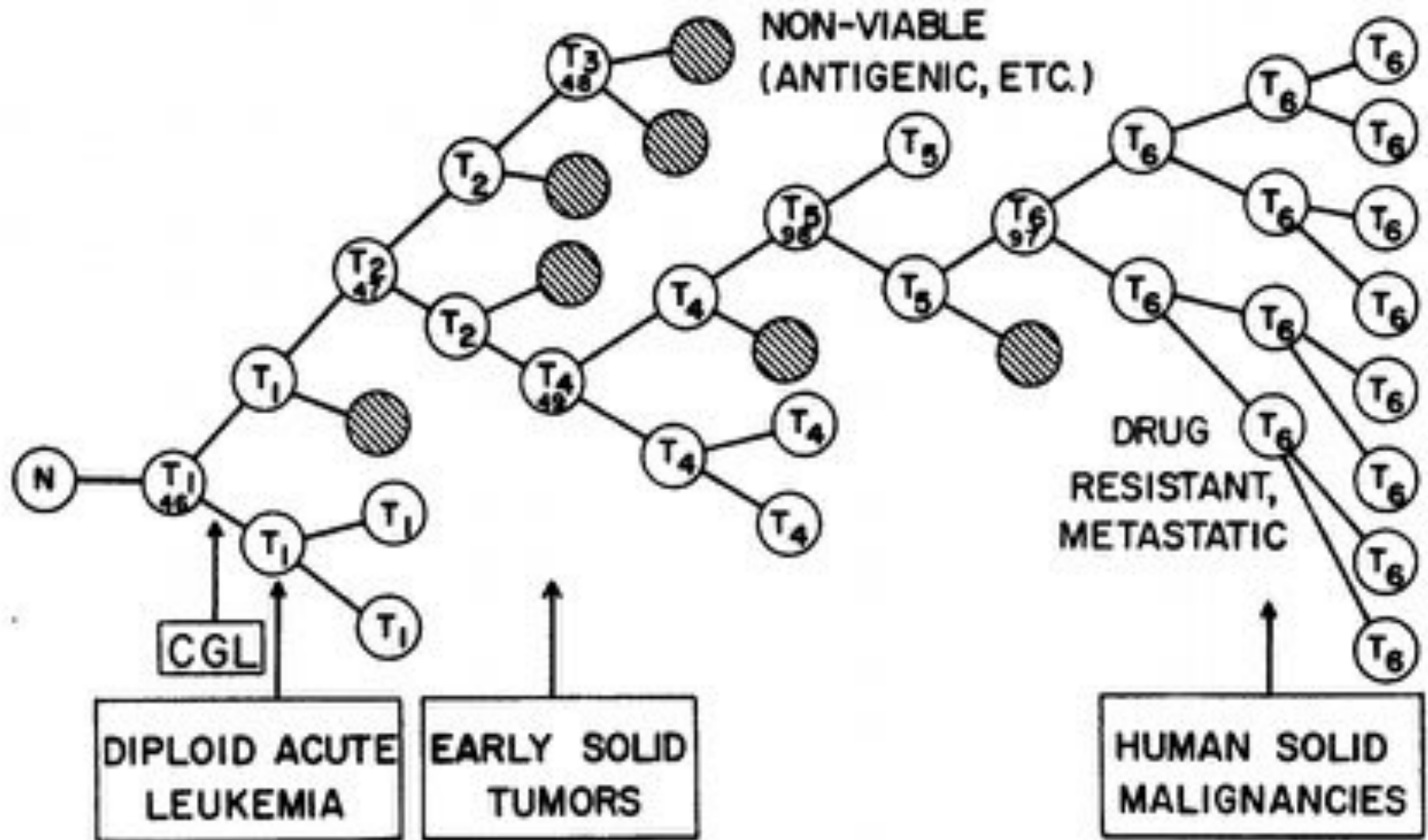
DNA Repair Genes



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

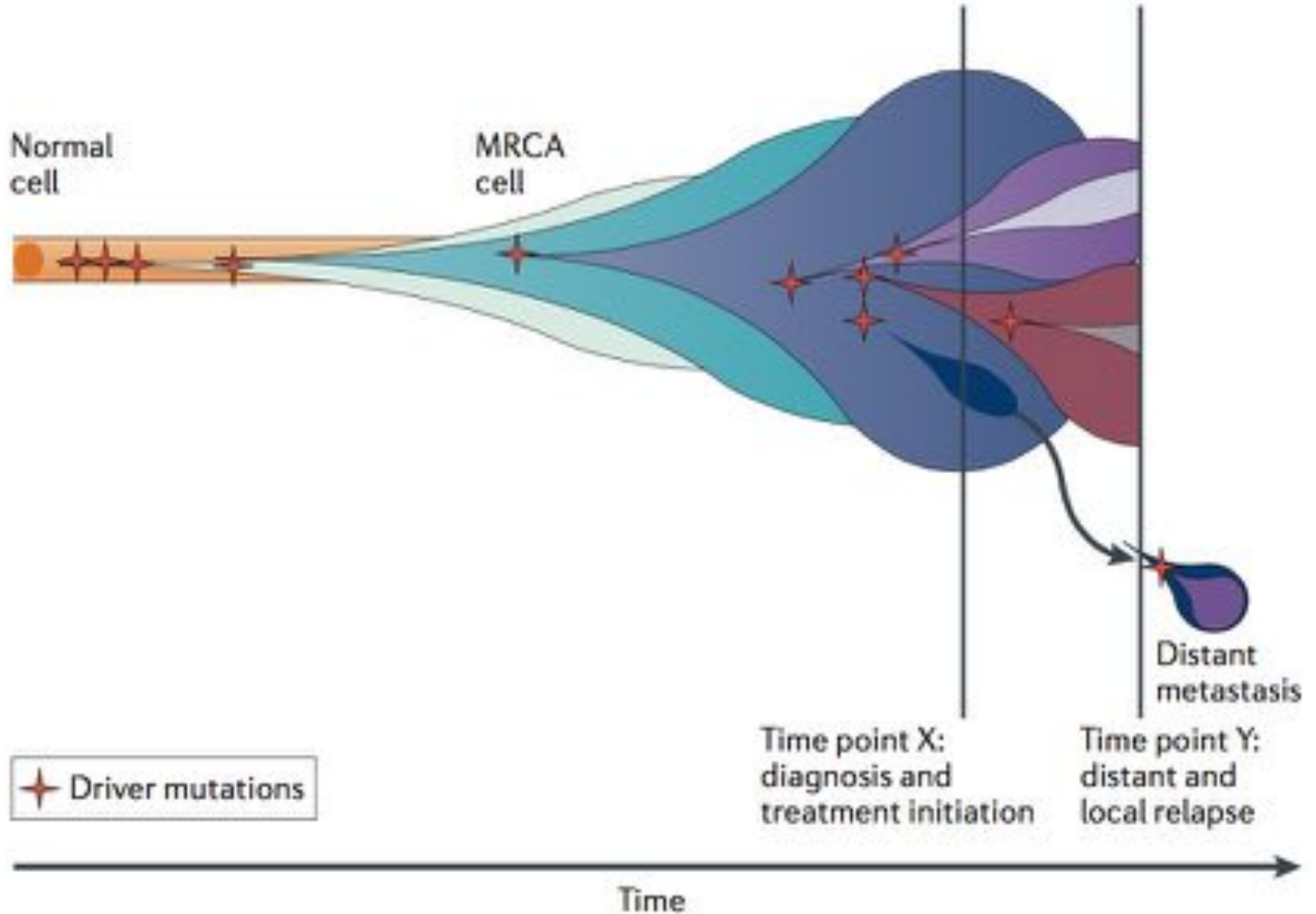
Tumor Evolution



The Clonal Evolution of Tumor Cell Populations

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

Tumor Evolution

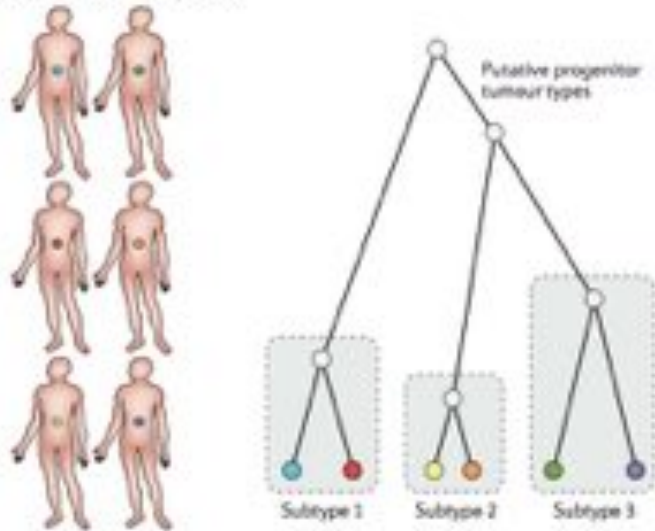


Evolution of the cancer genome

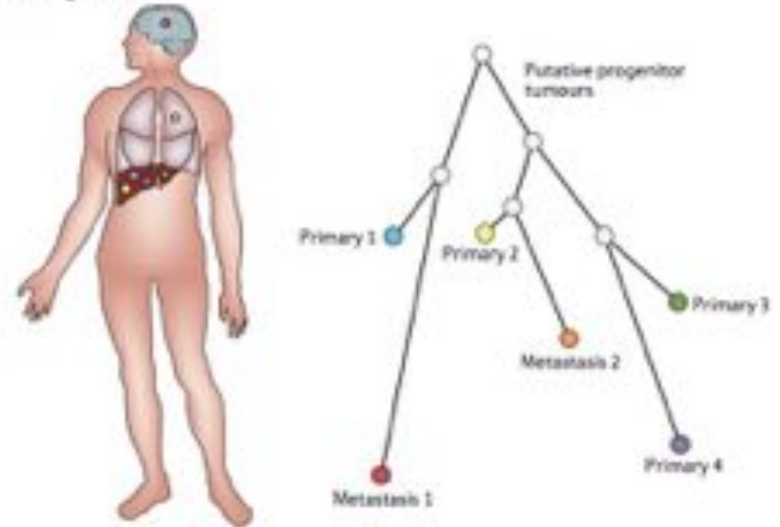
Yates & Campbell (2012) Nature Review Genetics. doi:10.1038/nrg3317

Tumor Heterogeneity

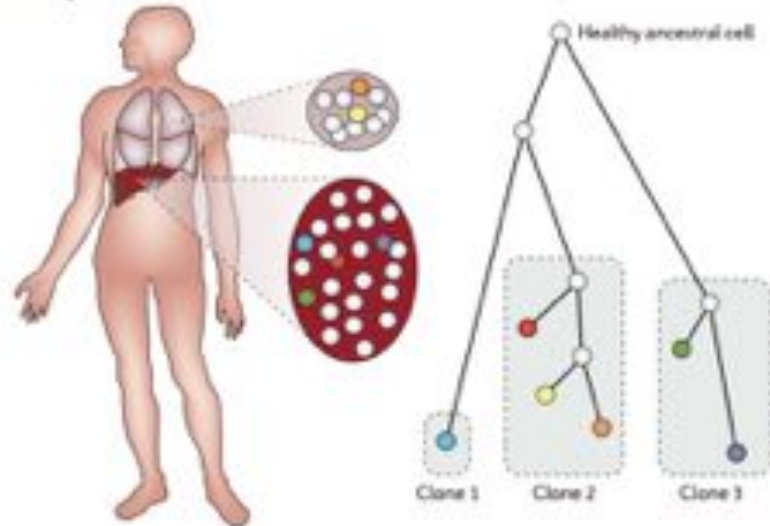
a Cross-sectional (oncogenetic)



b Regional bulk



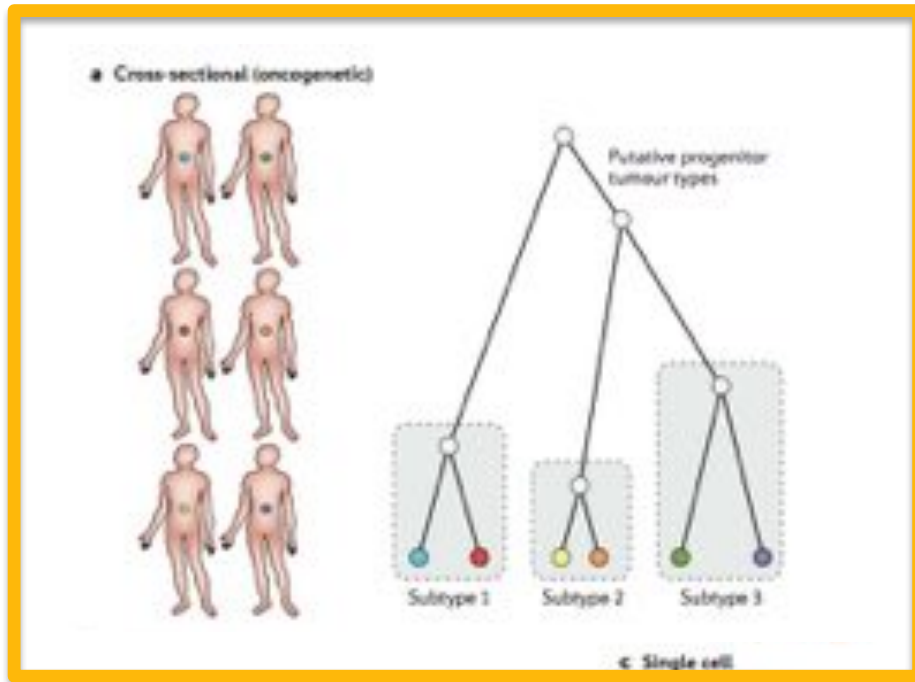
c Single cell



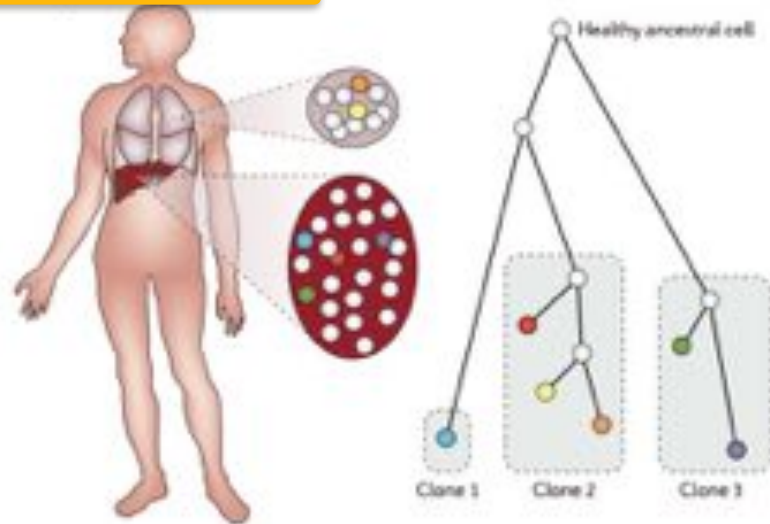
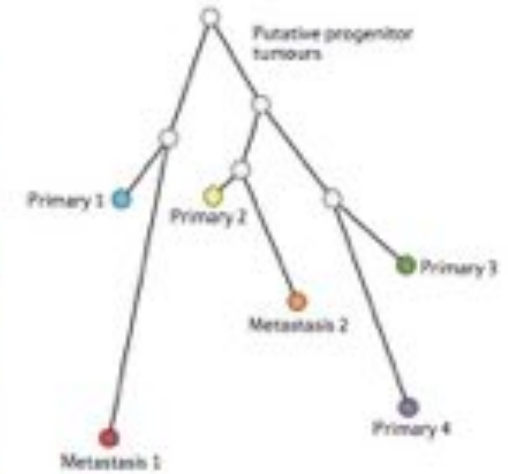
The evolution of tumour phylogenetics: principles and practice

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

Tumor Heterogeneity



b Regional bulk

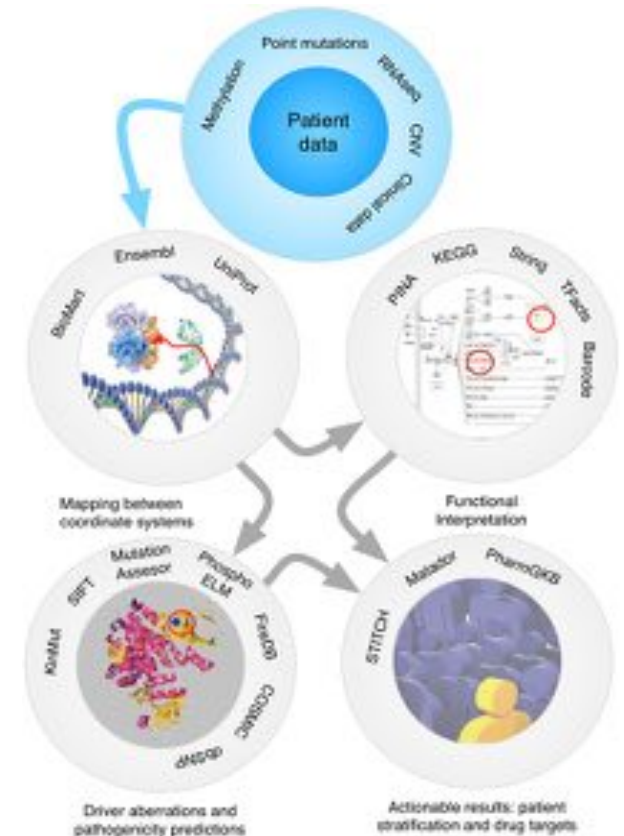


The evolution of tumour phylogenetics: principles and practice

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 on 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

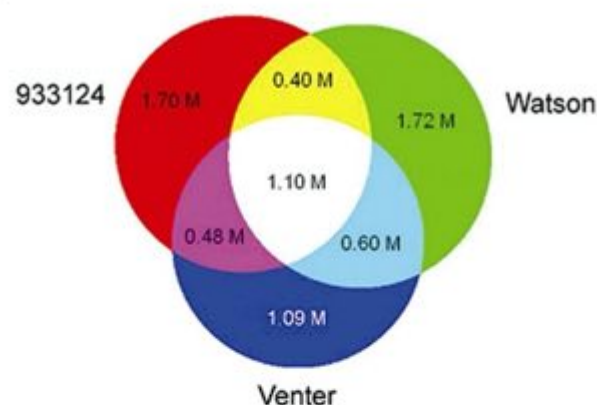
DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley^{1,2,3,4*}, Elaine R. Mardis^{2,4*}, Li Ding^{2,3}, Bob Fulton¹, Michael D. McLellan³, Ken Chen³, David Dooling³, Brian H. Dunford-Shore³, Sean McGrath³, Matthew Hickenbotham³, Lisa Cook³, Rachel Abbott³, David E. Larson³, Dan C. Koboldt³, Craig Pohl³, Scott Smith³, Amy Hawkins³, Scott Abbott³, Devin Locke³, LaDeana W. Hillier^{3,4}, Tracie Miner³, Lucinda Fulton³, Vincent Magrini^{2,3}, Todd Wylie³, Jarret Glasscock³, Joshua Conyers³, Nathan Sander³, Xiaoqi Shi³, John R. Osborne³, Patrick Minx³, David Gordon³, Asif Chinwalla³, Yu Zhao¹, Rhonda E. Ries¹, Jacqueline E. Payton³, Peter Westervelt^{1,4}, Michael H. Tomasson^{1,4}, Mark Watson^{3,4,5}, Jack Baty⁶, Jennifer Ivanovich^{6,7}, Sharon Heath^{1,4}, William D. Shannon^{1,4}, Rakesh Nagarajan^{8,5}, Matthew J. Walter^{1,4}, Daniel C. Link^{1,4}, Timothy A. Graubert^{1,4}, John F. DiPersio^{1,4} & Richard K. Wilson^{2,3,4}

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 Cancer Genome

A.



B.

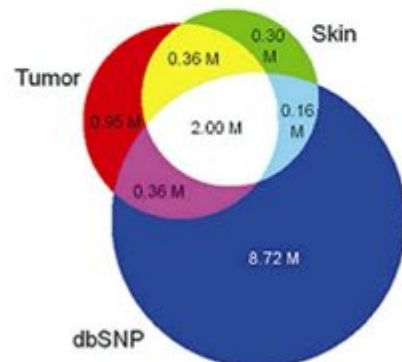


Figure 1. Overlap of SNPs detected in 933124 and other genomes

(A) Venn diagram of overlap between SNPs detected in the 933124 tumor genome and the genomes of Watson and Venter. (B) Venn Diagram of overlap among 933124 tumor genome, skin genome, and dbSNP (ver. 127). Single nucleotide variants were defined with a MAQ SNP quality ≥ 15 .

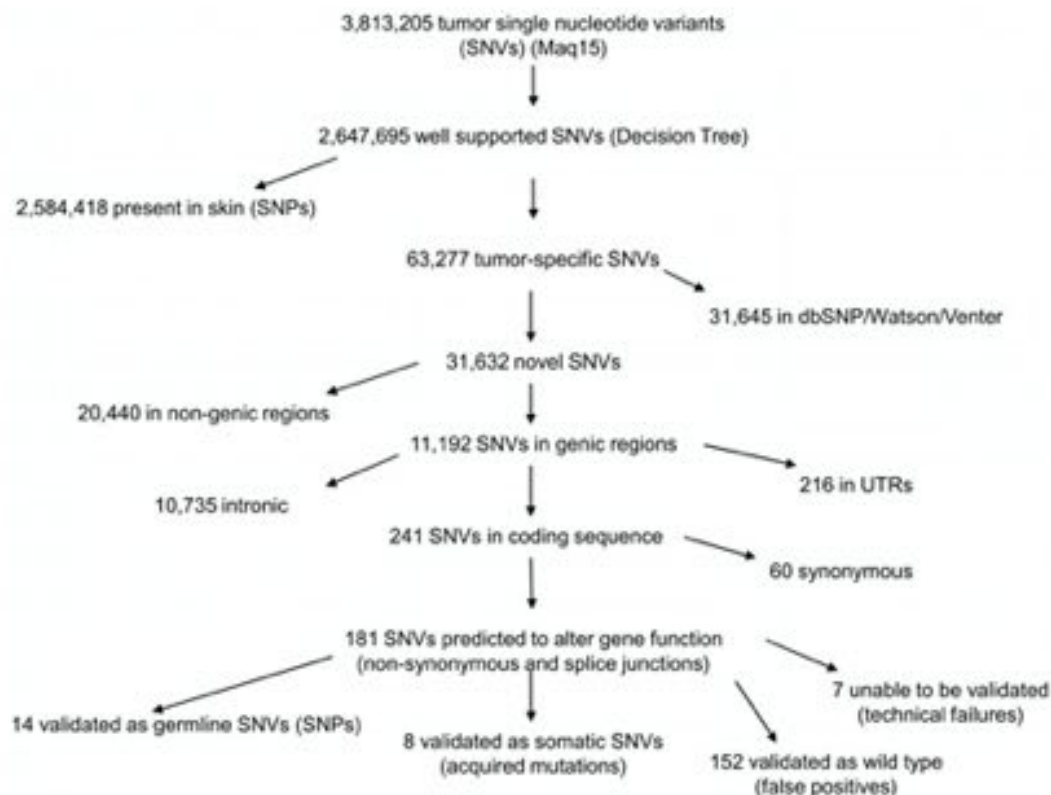
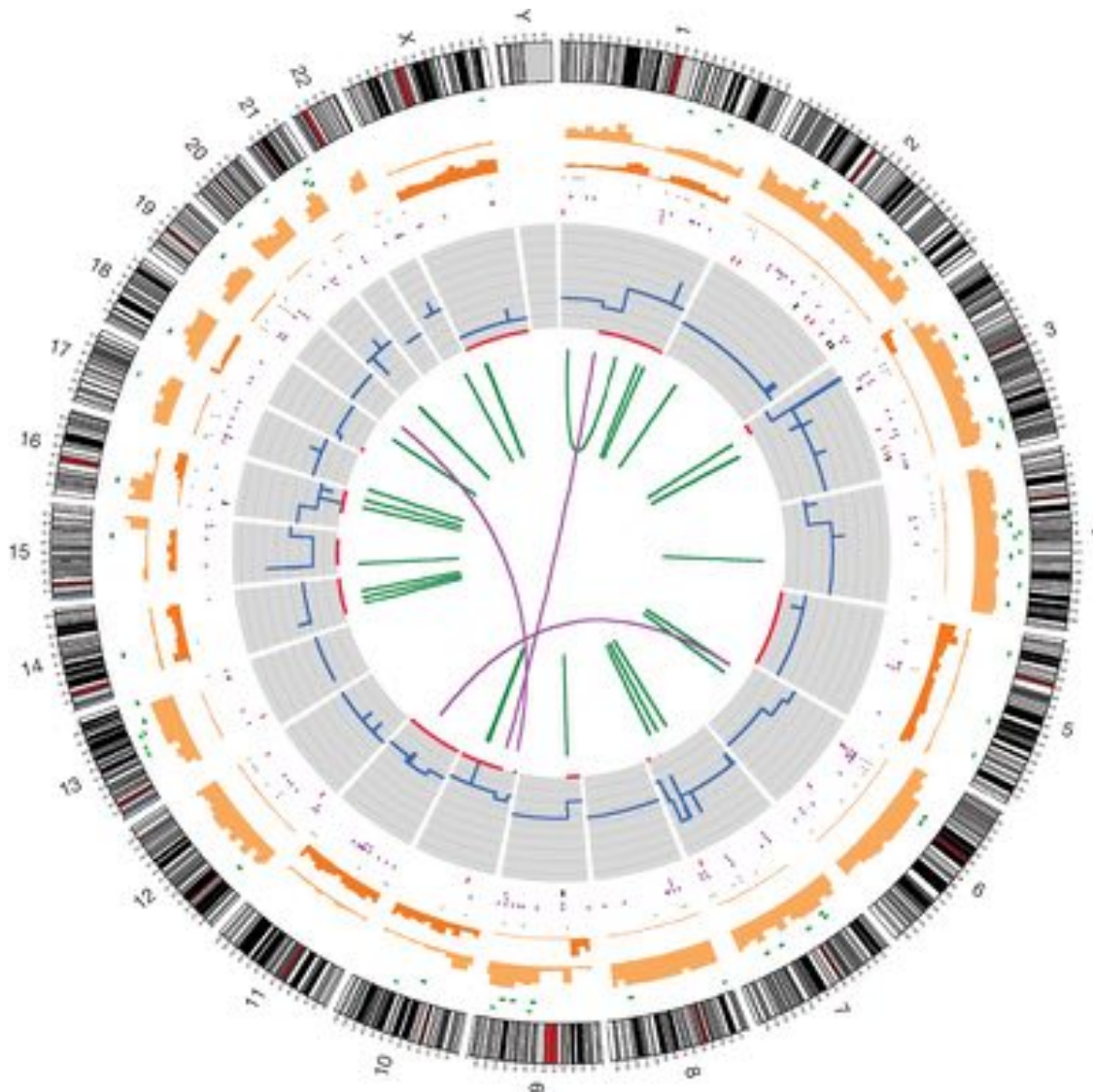


Figure 2. Filters used to identify somatic point mutations in the tumor genome
See text for details.

First Melanoma Genome

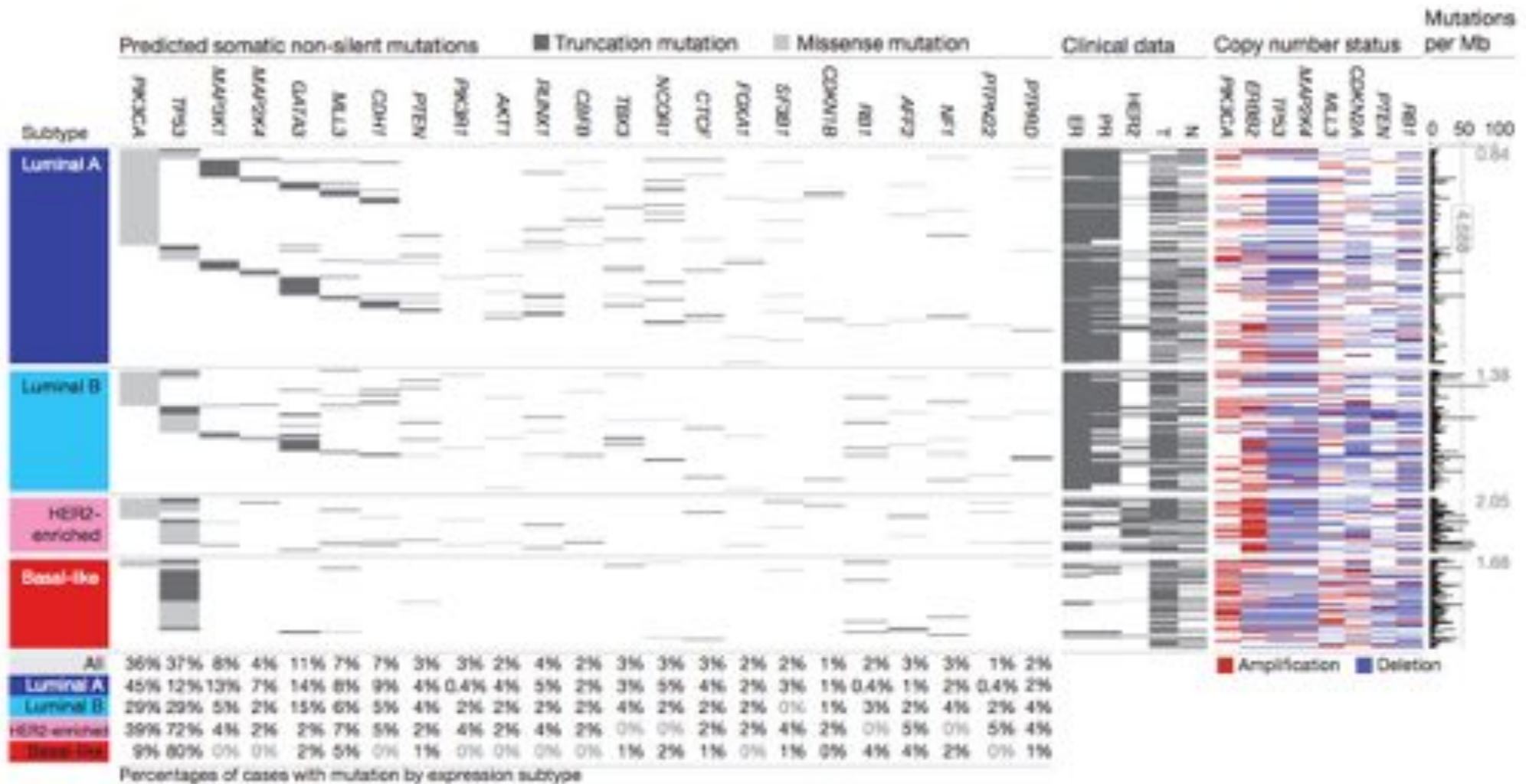


- Insertions (light-green rectangles);
- Deletions (dark-green rectangles);
- Heterozygous (light-orange bars) and Homozygous (dark-orange bars) Substitutions
- Coding substitutions (coloured squares: silent in grey, missense in purple, nonsense in red and splice site in black);
- Copy number (blue lines);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

A comprehensive catalogue of somatic mutations from a human cancer genome

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

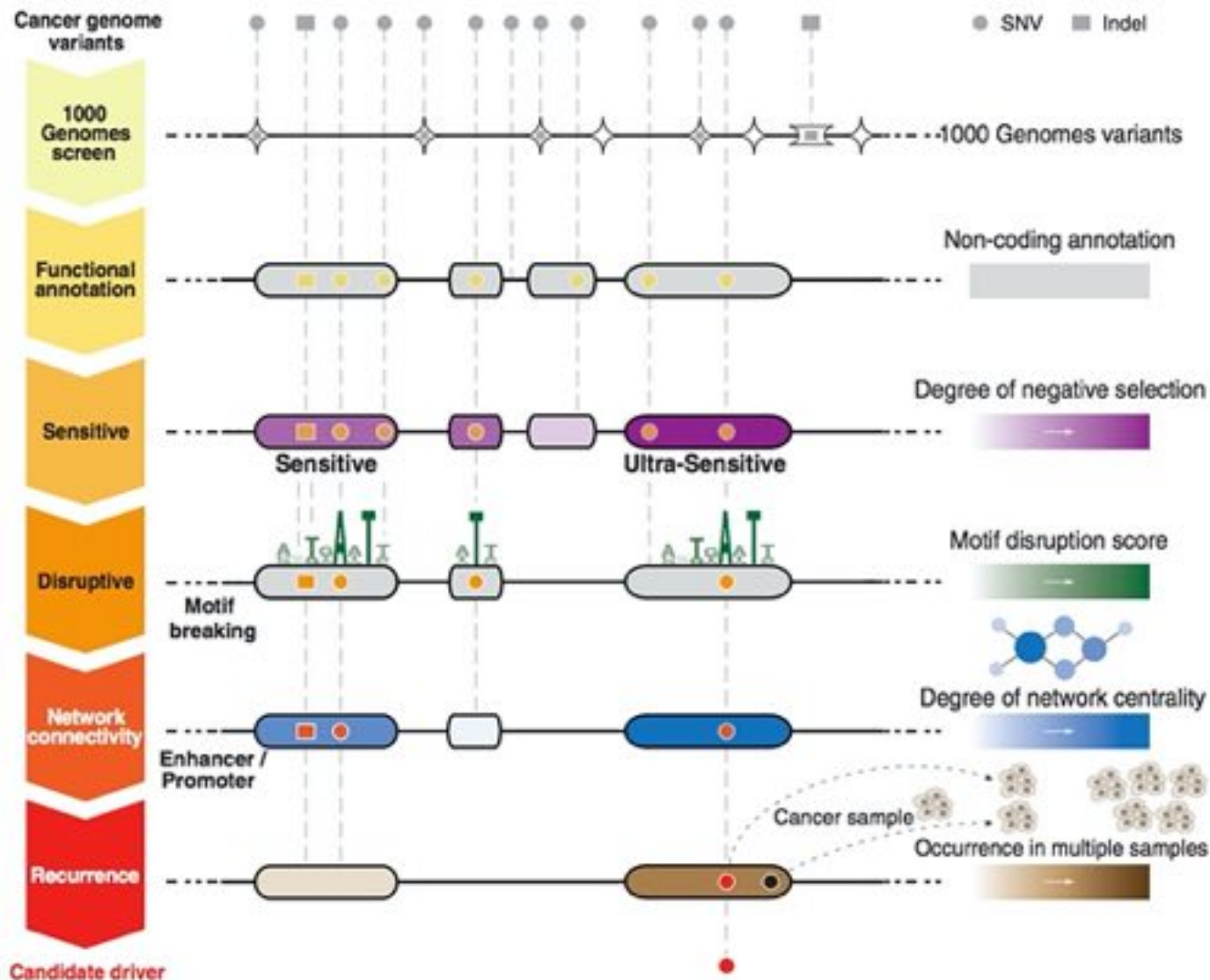
Mutations in Breast Cancer



Comprehensive molecular portraits of human breast tumours

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

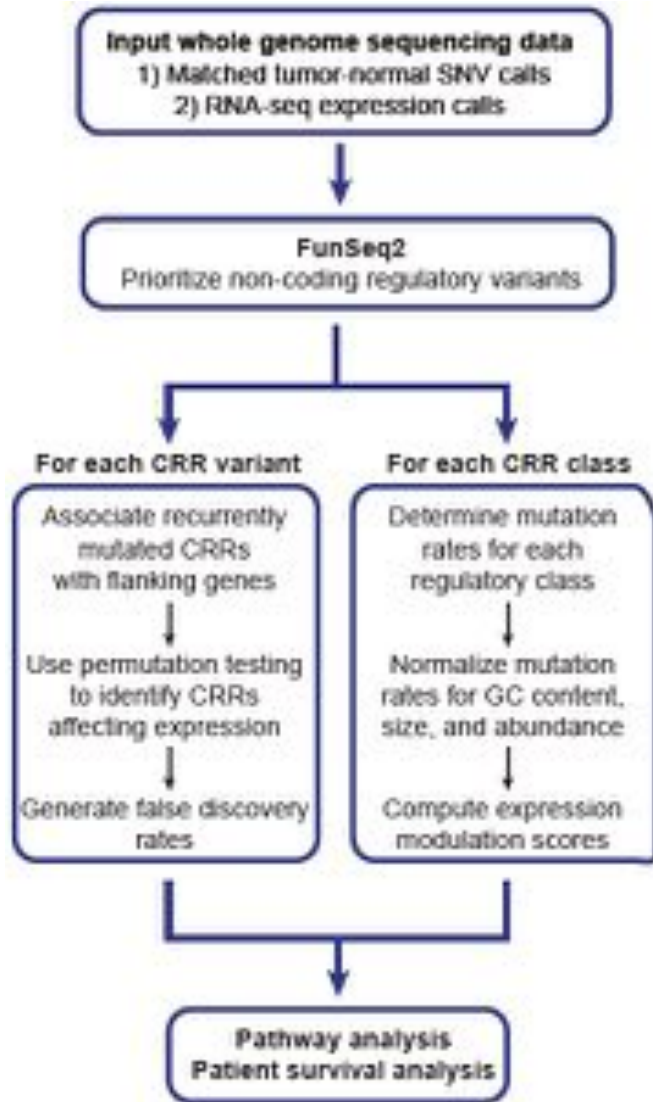
Finding Driving Mutations



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

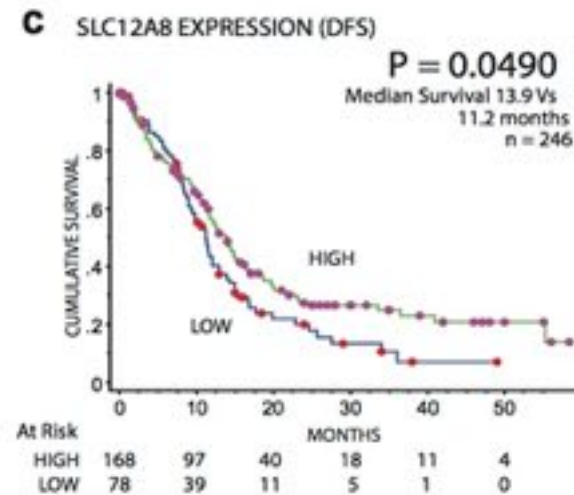
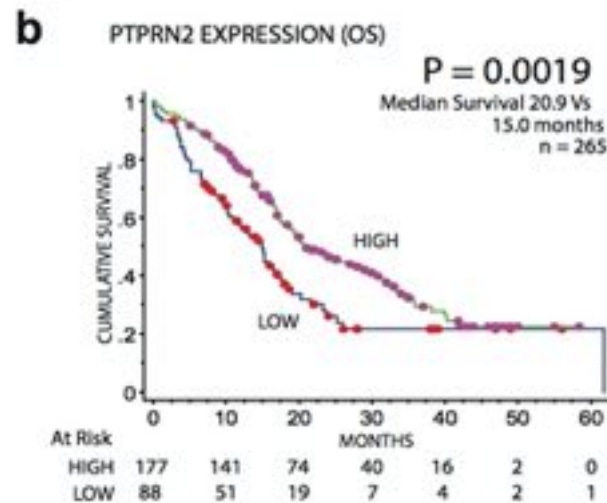
Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma

Feigin, M, Garvin, T et al. (2017) Nature Genetics. doi:10.1038/ng.3861

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

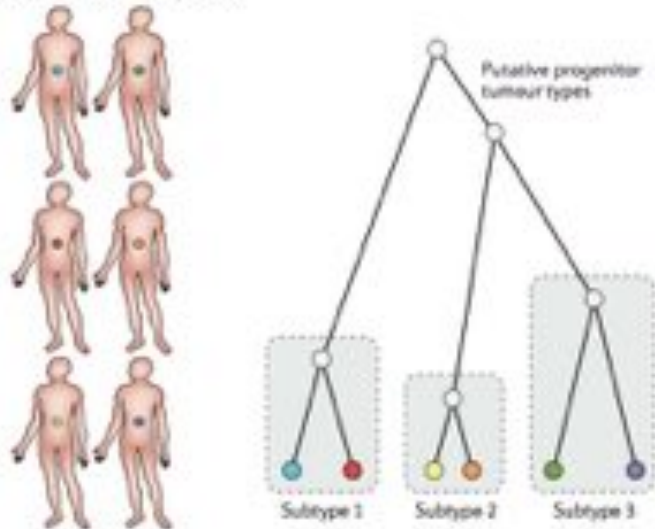


Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma

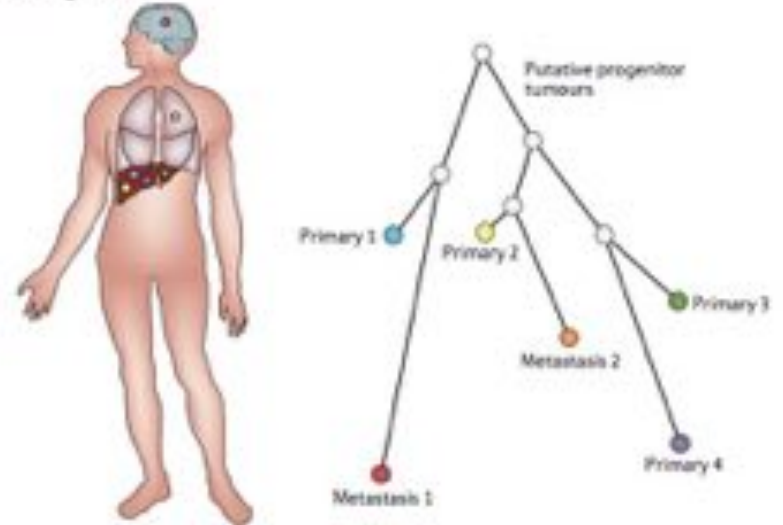
Feigin, M, Garvin, T et al. (2017) Nature Genetics. doi:10.1038/ng.3861

Tumor Heterogeneity

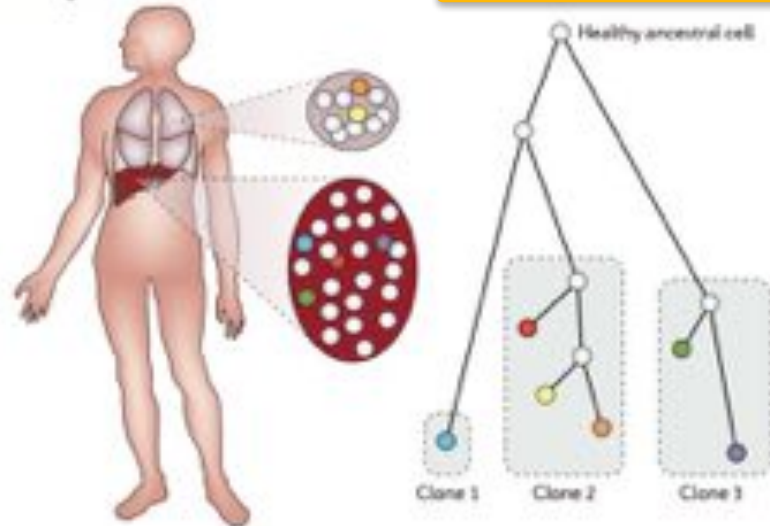
a Cross-sectional (oncogenetic)



b Regional bulk



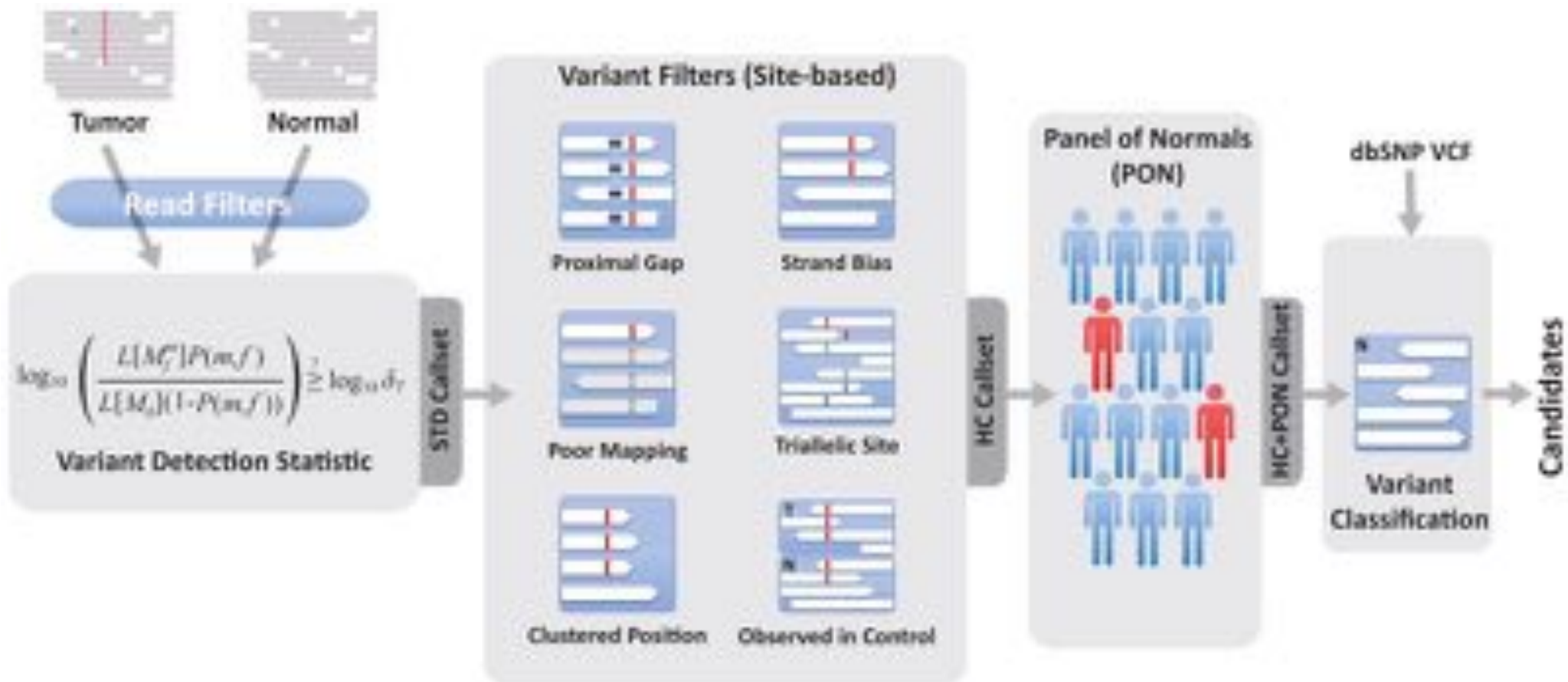
c Single cell



The evolution of tumour phylogenetics: principles and practice

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

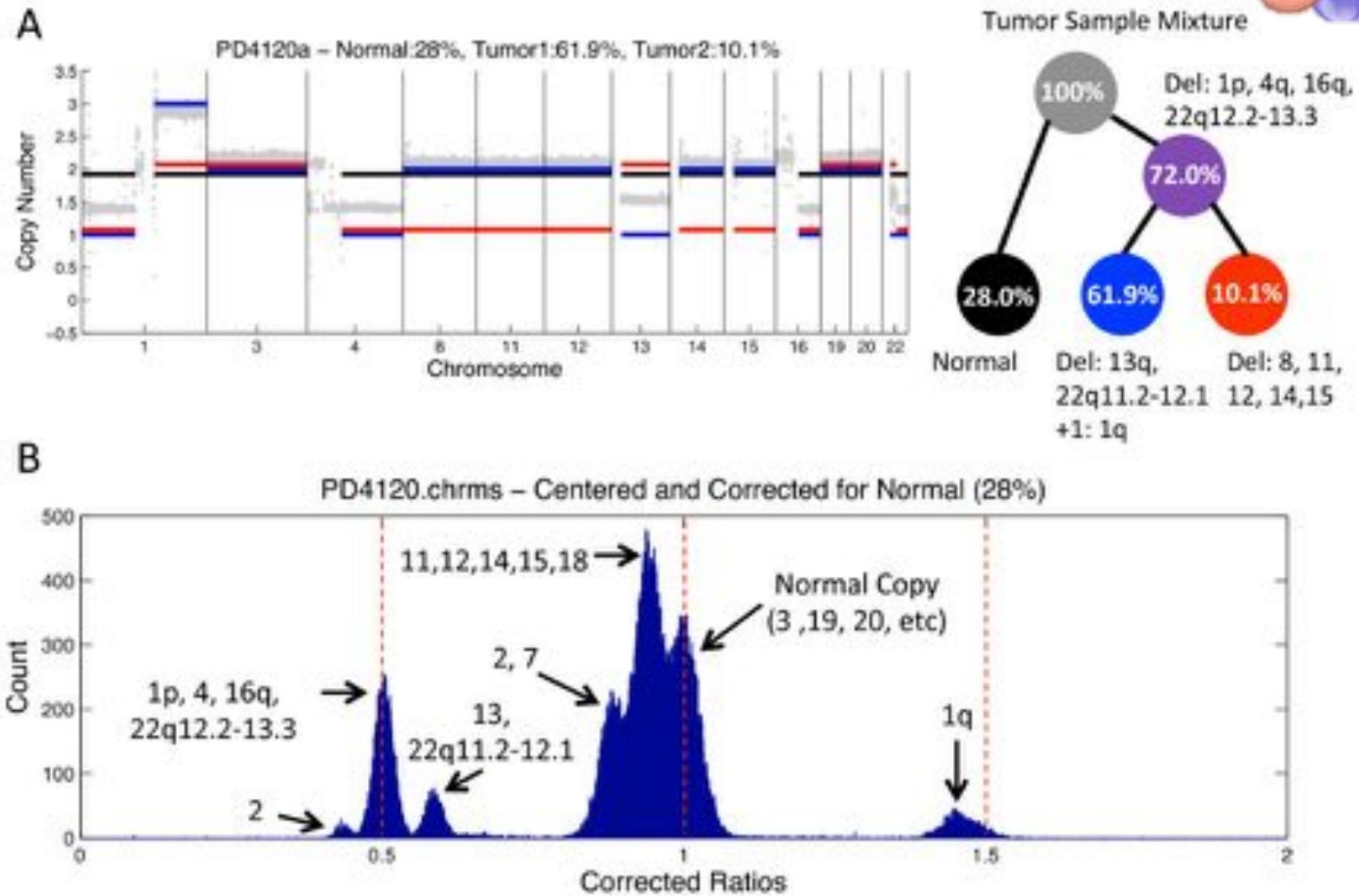
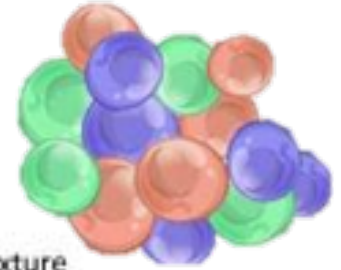
Tumor-Normal Pairs



Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples

Cibulskis et al (2013) Nature Biotech. doi:10.1038/nbt.2514

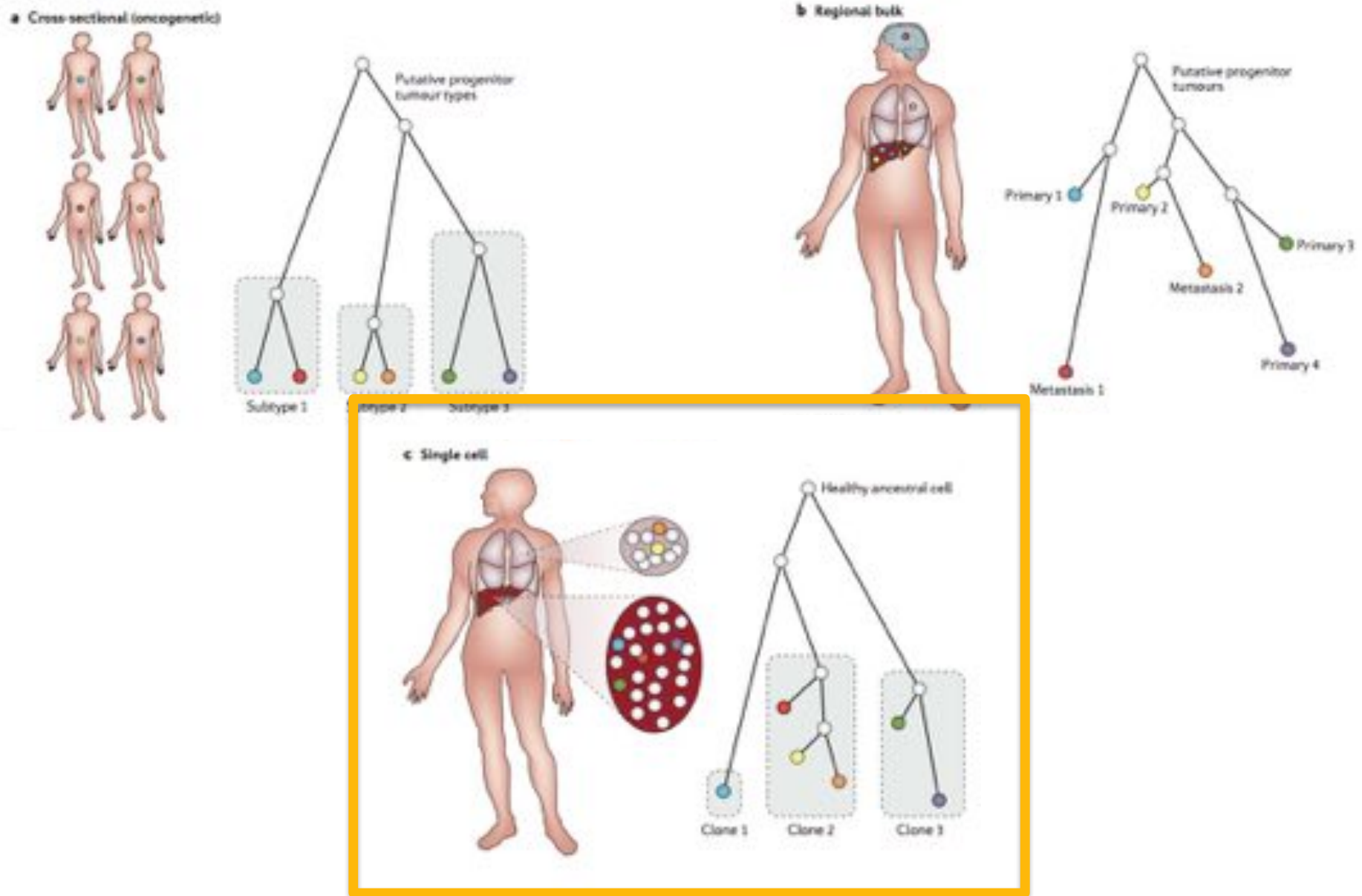
Bulk Heterogeneity



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

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

Tumor Heterogeneity

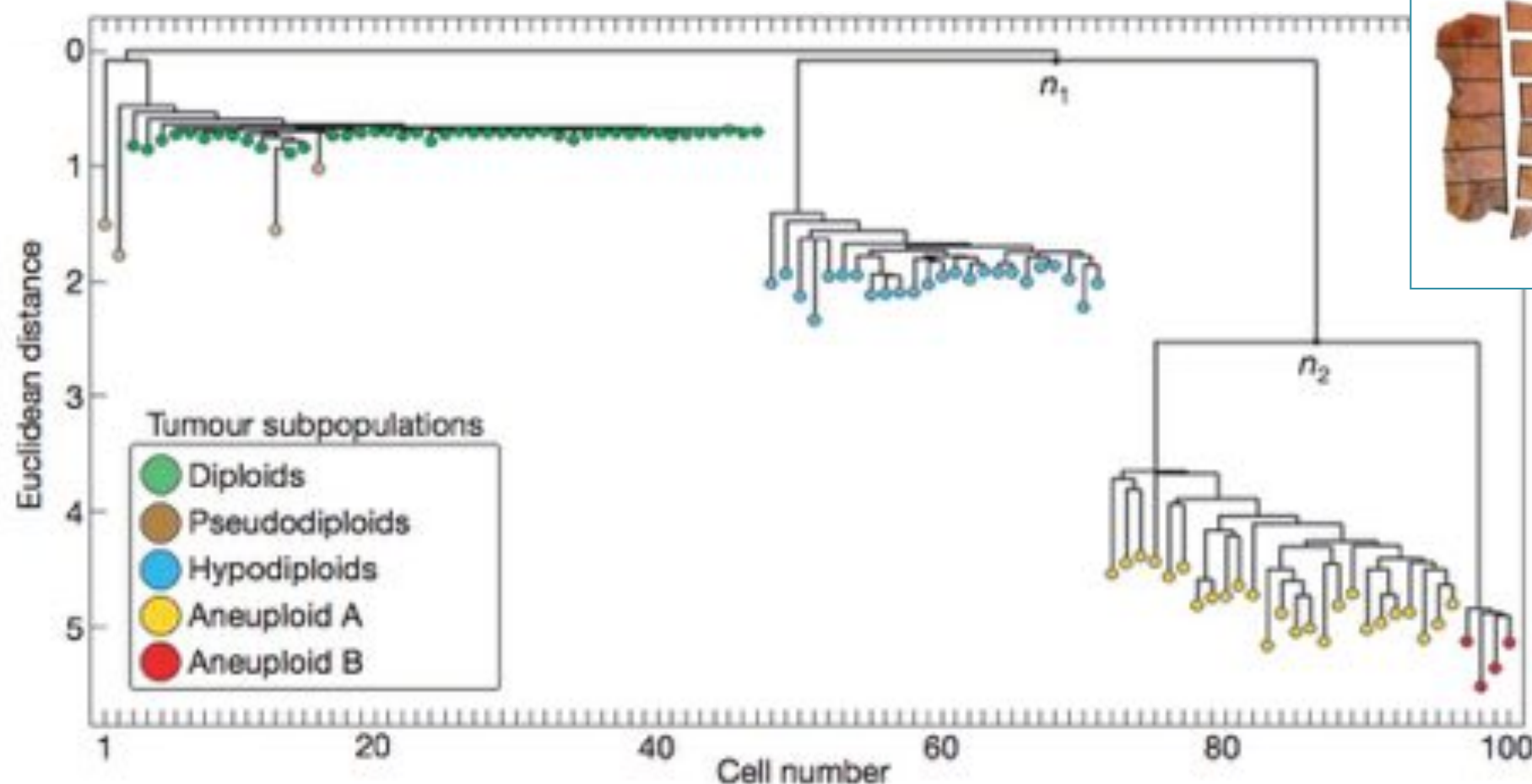


The evolution of tumour phylogenetics: principles and practice

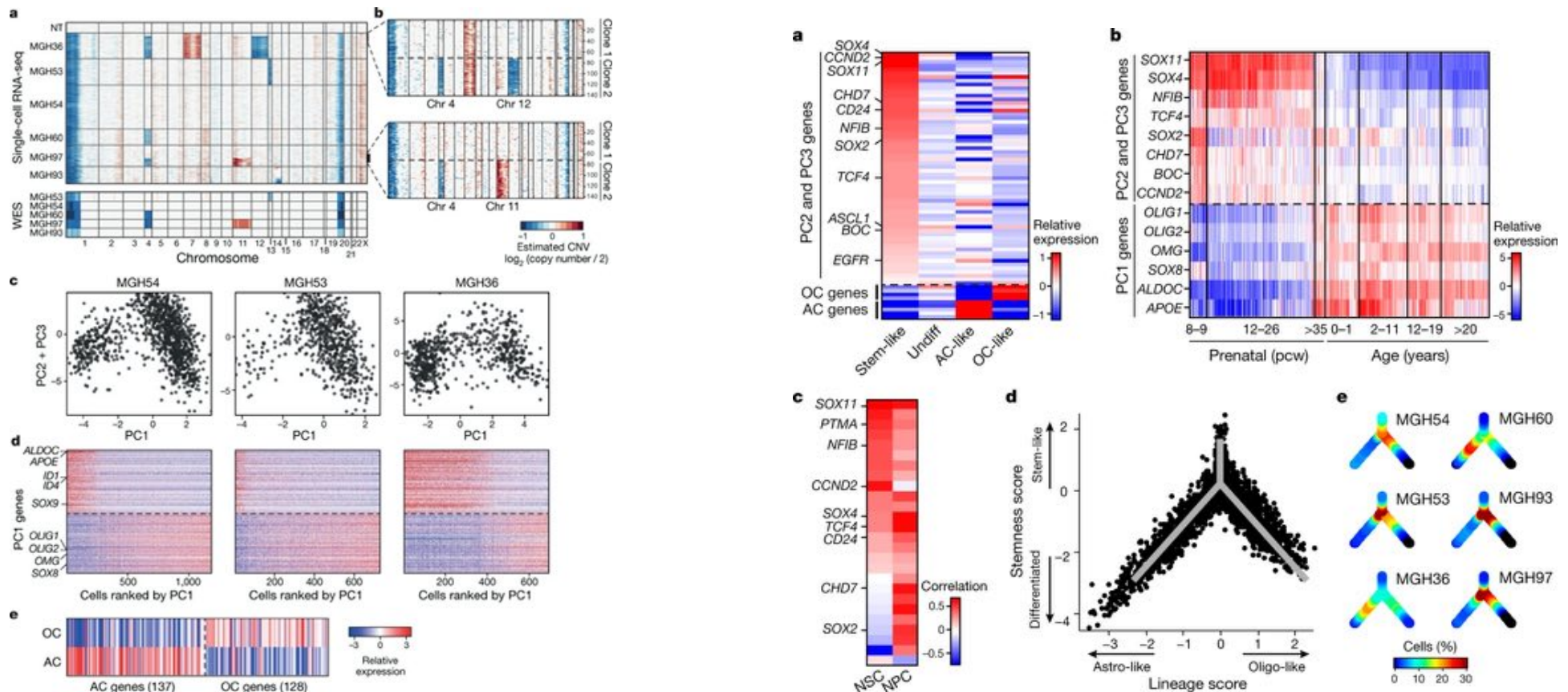
Schwarz and Schaffer (2017) *Nature Reviews Genetics*. doi:10.1038/nrg.2016.170

Tumour evolution inferred by single-cell sequencing

Nicholas Navin^{1,2}, Jude Kendall¹, Jennifer Troge¹, Peter Andrews¹, Linda Rodgers¹, Jeanne McIndoo¹, Kerry Cook¹, Asya Stepanisky¹, Dan Levy¹, Diane Esposito¹, Lakshmi Muthuswamy³, Alex Krasnitz¹, W. Richard McComble¹, James Hicks¹ & Michael Wigler¹

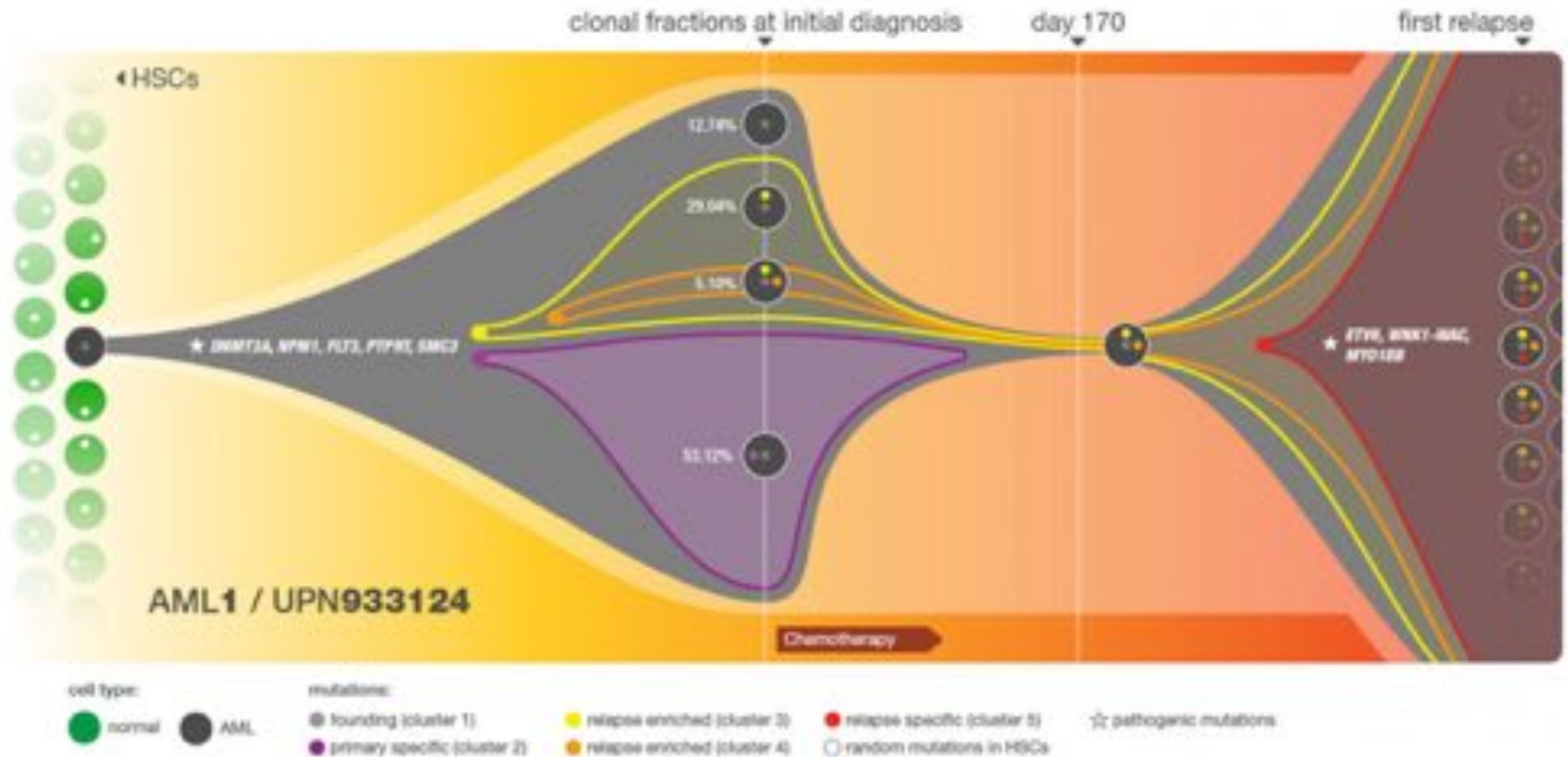


Single Cell RNA-seq of Cancer



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

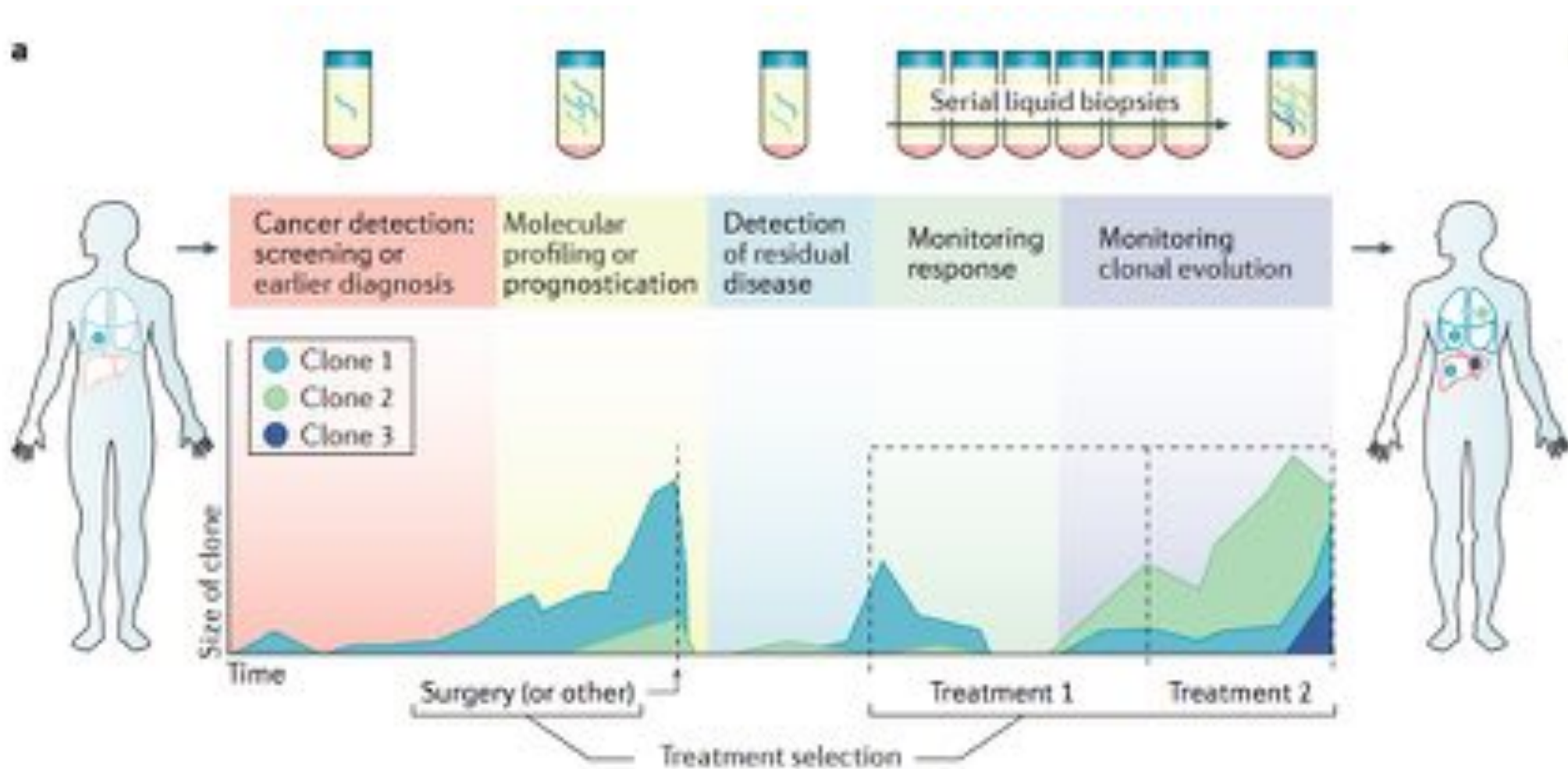
Tumor Heterogeneity and Treatment



Clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing

Ding et al (2012) Nature. doi:10.1038/nature10738

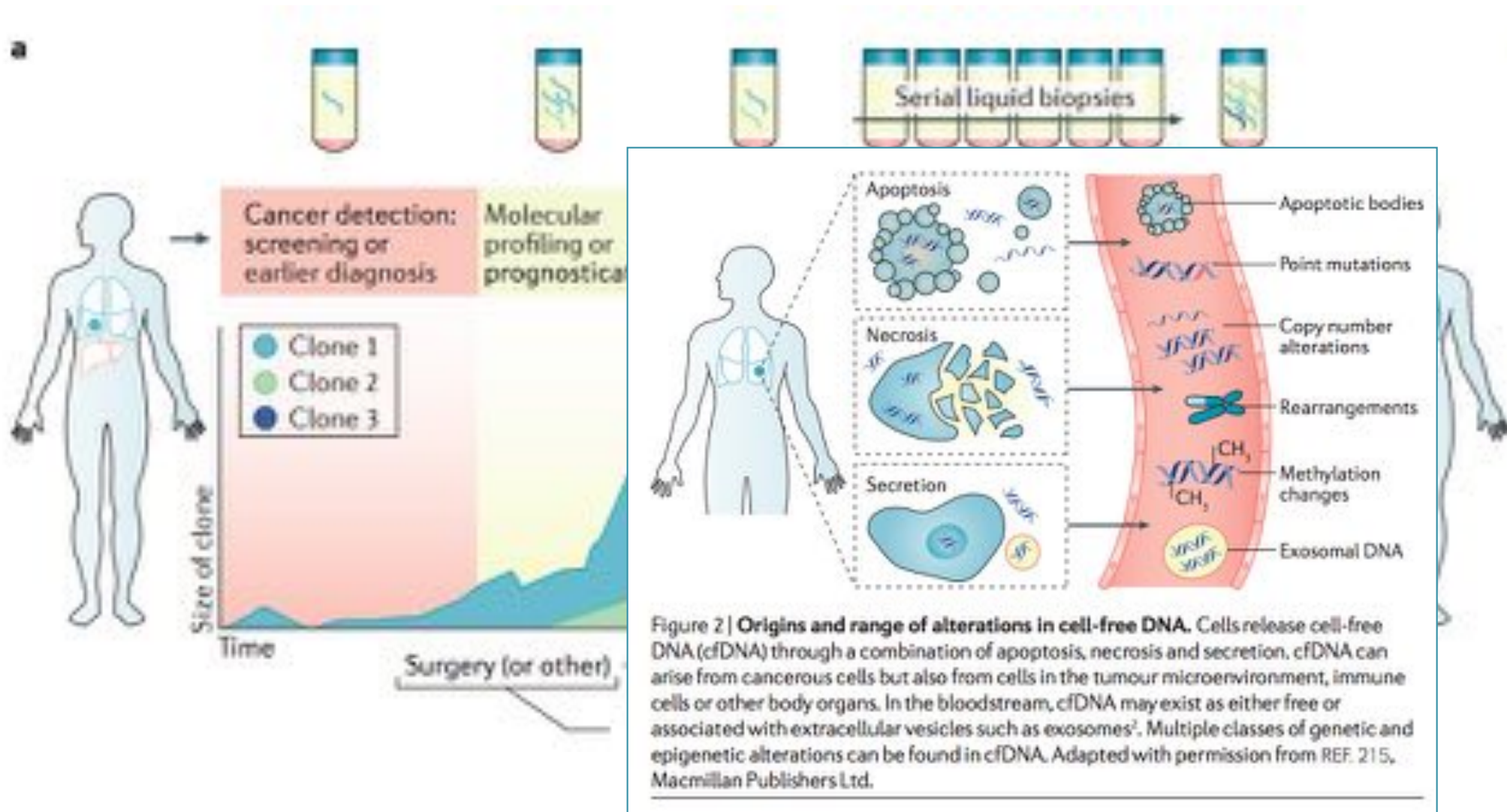
Liquid Biopsies



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

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

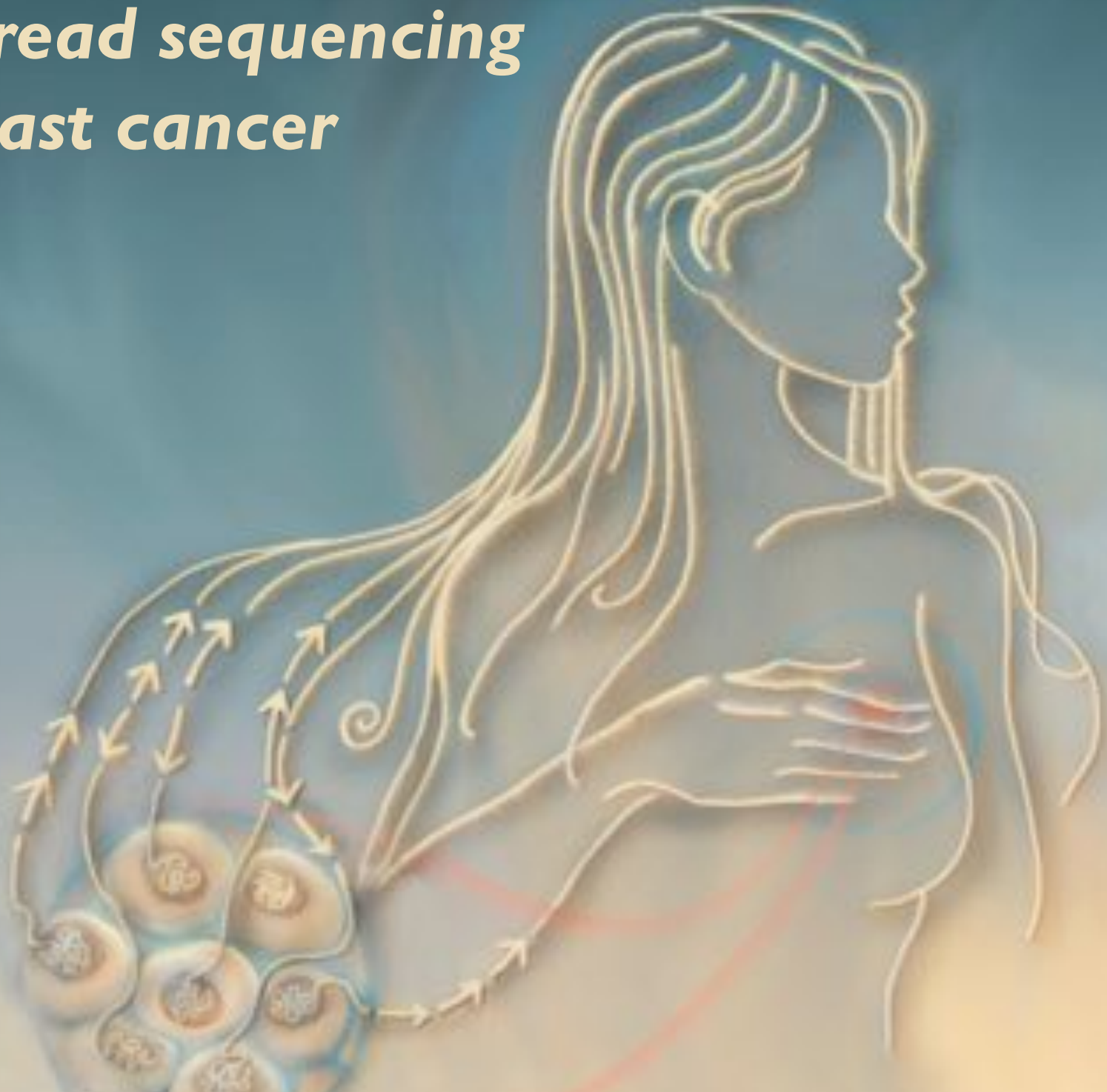
Liquid Biopsies



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

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

Long-read sequencing of breast cancer

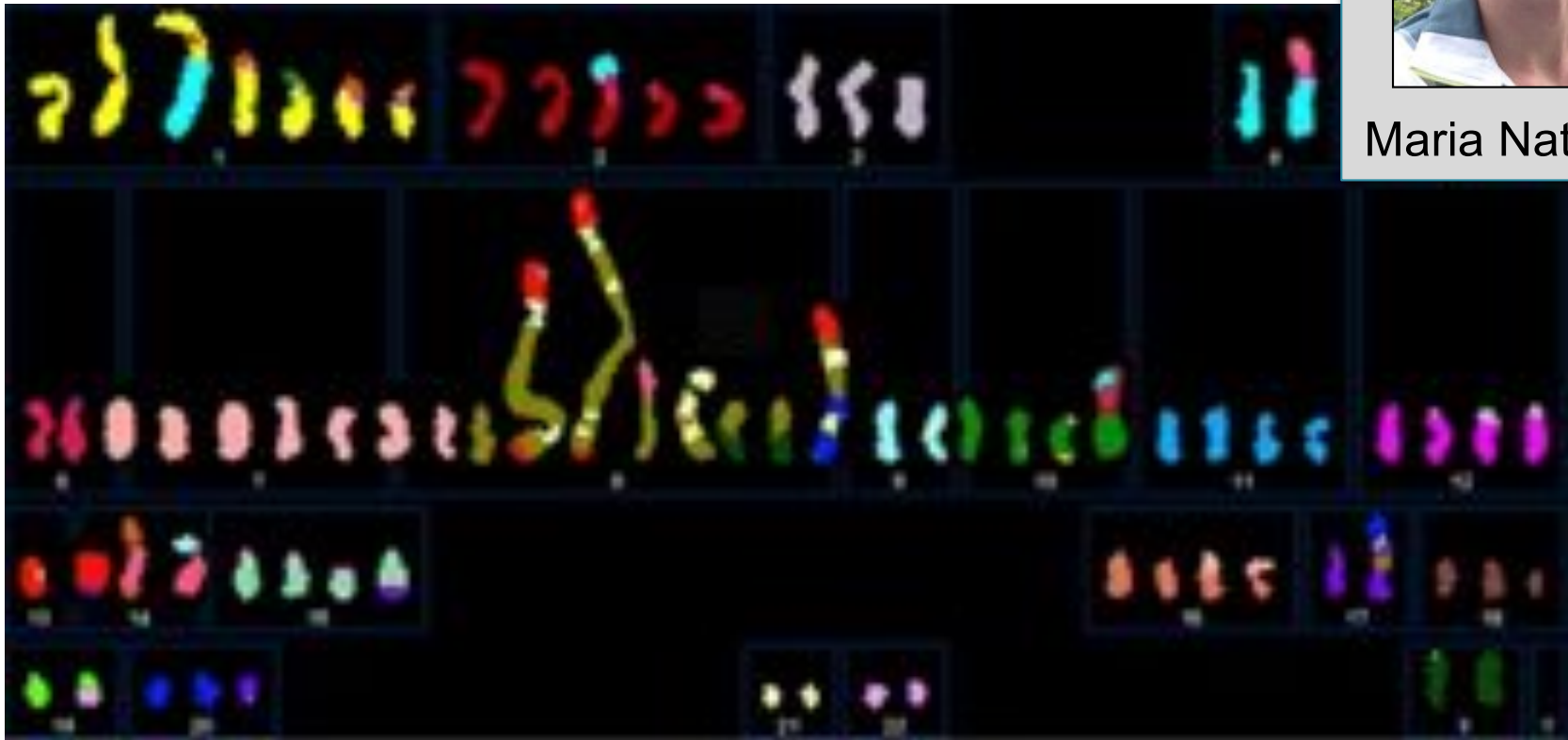


SK-BR-3

Most commonly used Her2-amplified breast cancer



Maria Nattestad



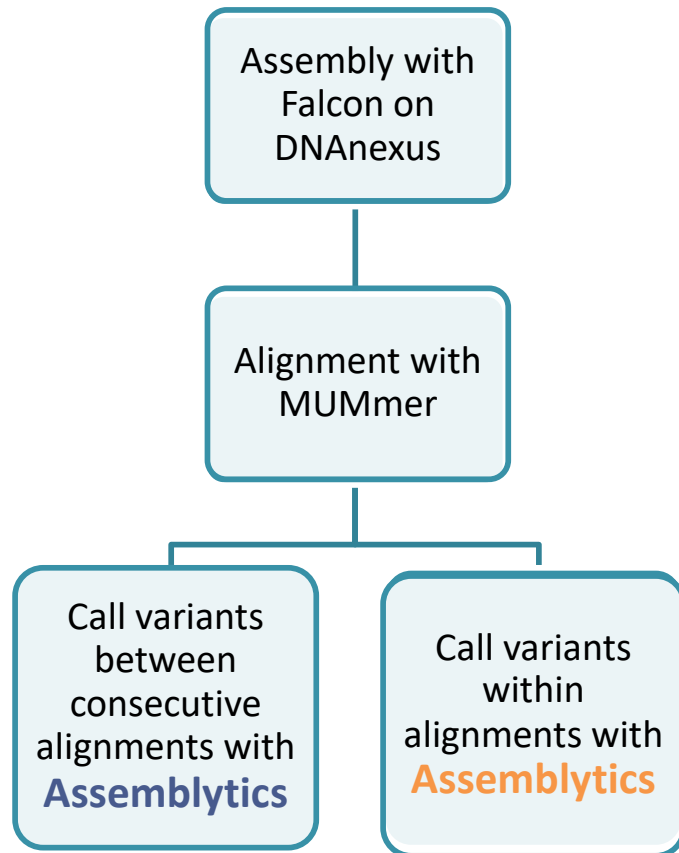
(Davidson et al, 2000)

Can we resolve the complex structural variations, especially around Her2?

Recent collaboration between JHU, CSHL and OICR to *de novo* assemble and analyze the complete cell line genome with PacBio long reads

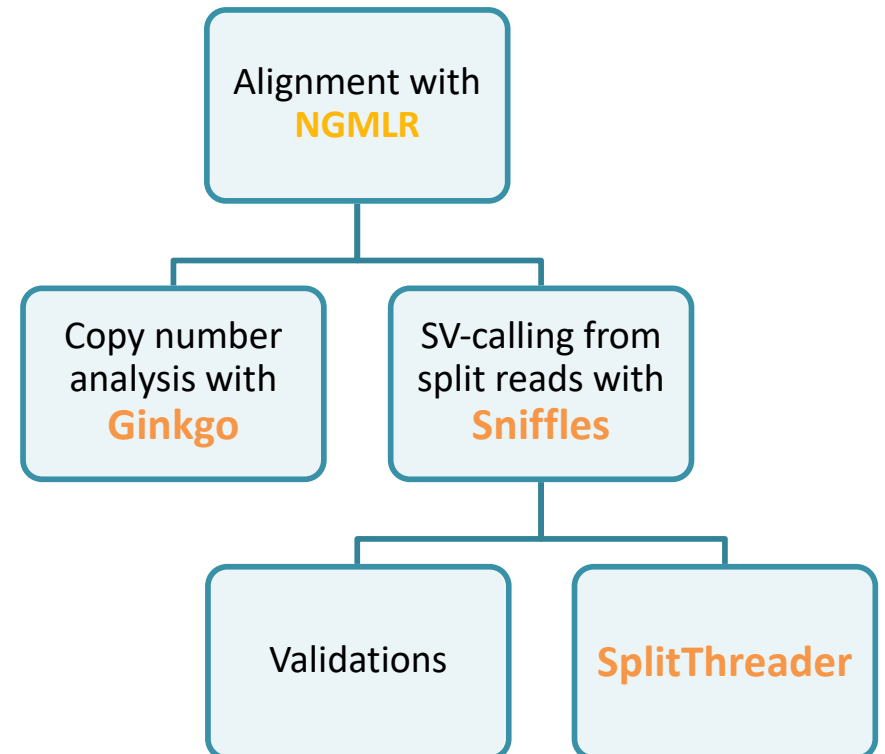
Structural Variation Analysis

Assembly-based



~ 11,000 structural variants
50 bp to 10 kbp

Split-Read based



~ 20,000 structural variants
Including many inter-chromosomal
rearrangements

NGMLR + Sniffles

BWA-MEM:



NGMLR:



NGMLR: Convex scoring model to accommodate many small gaps from sequencing errors along with less frequent but larger SVs

Accurate detection of complex structural variations using single molecule sequencing

Sedlazeck, Rescheneder et al (2018) *Nature Methods*. doi:10.1038/s41592-018-0001-7

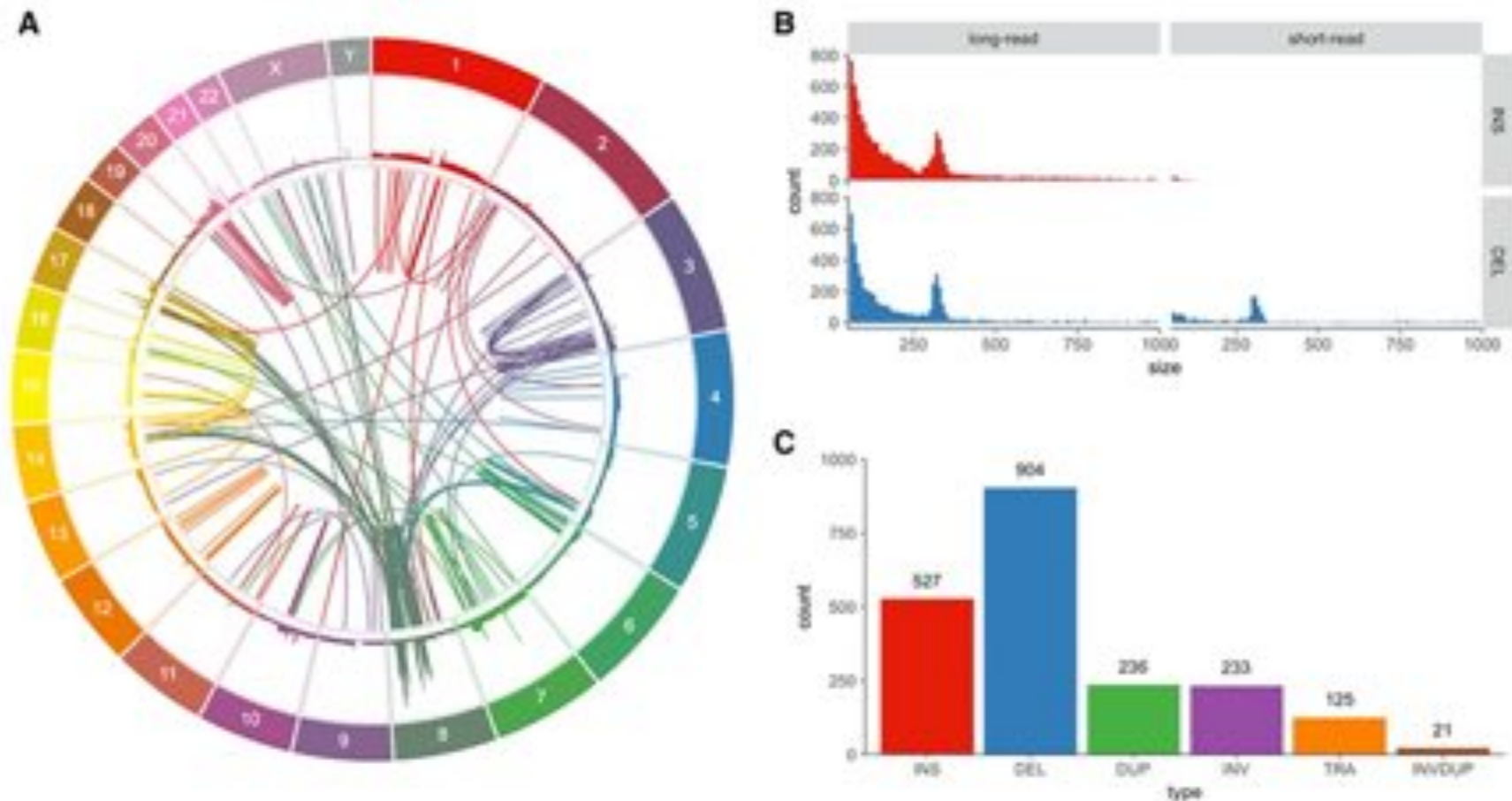


Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (A) Circos (Kryzwiniski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the outer track. (B) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (SURVIVOR 2-caller consensus) variant calling, showing similar size distributions for insertions and deletions from long reads but not for short reads, where insertions are greatly underrepresented. (C) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Nattestad et al. (2018) *Genome Research*. doi: 10.1101/gr.231100.117

Highlights

- Finding 10s of thousands of additional variants
- PCR validation confirms high accuracy of long reads
- Detect many novel gene fusions
- Identify early vs late mutations in the cancer

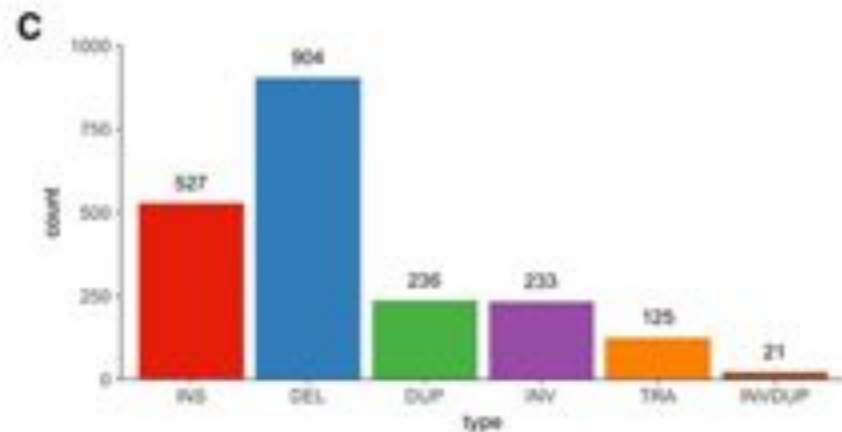
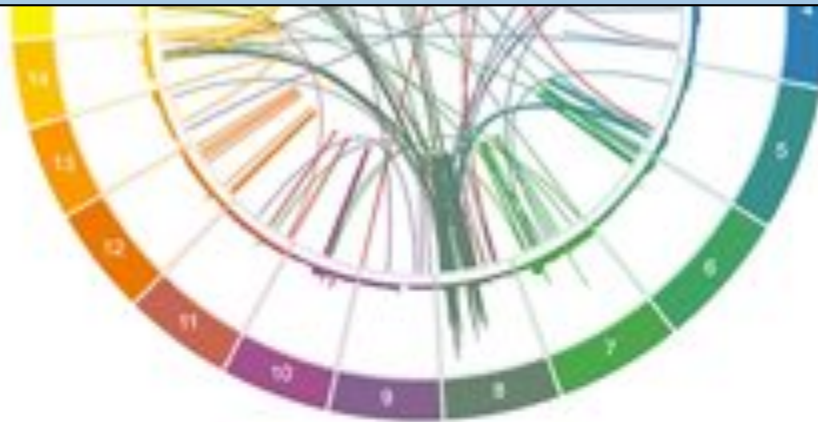
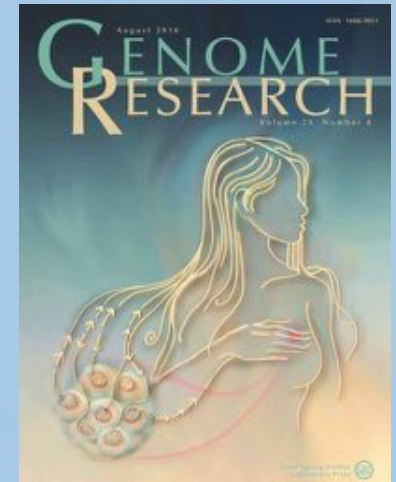
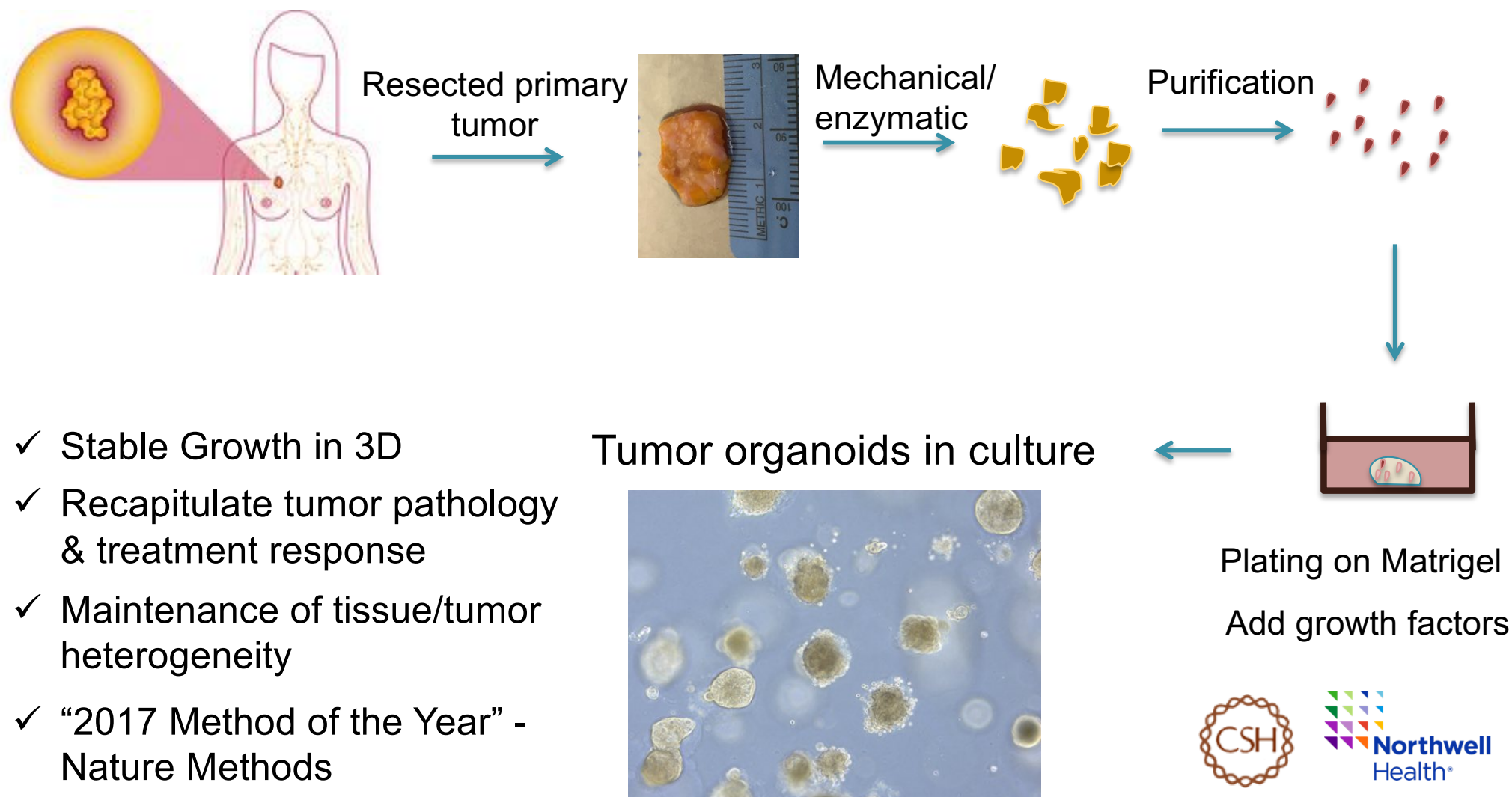


Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (A) Circos (Kryzwiniski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the outer track. (B) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (SURVIVOR 2-caller consensus) variant calling, showing similar size distributions for insertions and deletions from long reads but not for short reads, where insertions are greatly underrepresented. (C) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Nattestad et al. (2018) *Genome Research*. doi: 10.1101/gr.231100.117

Taking Long Read Sequencing into the Clinic



- ✓ Stable Growth in 3D
- ✓ Recapitulate tumor pathology & treatment response
- ✓ Maintenance of tissue/tumor heterogeneity
- ✓ “2017 Method of the Year” - Nature Methods

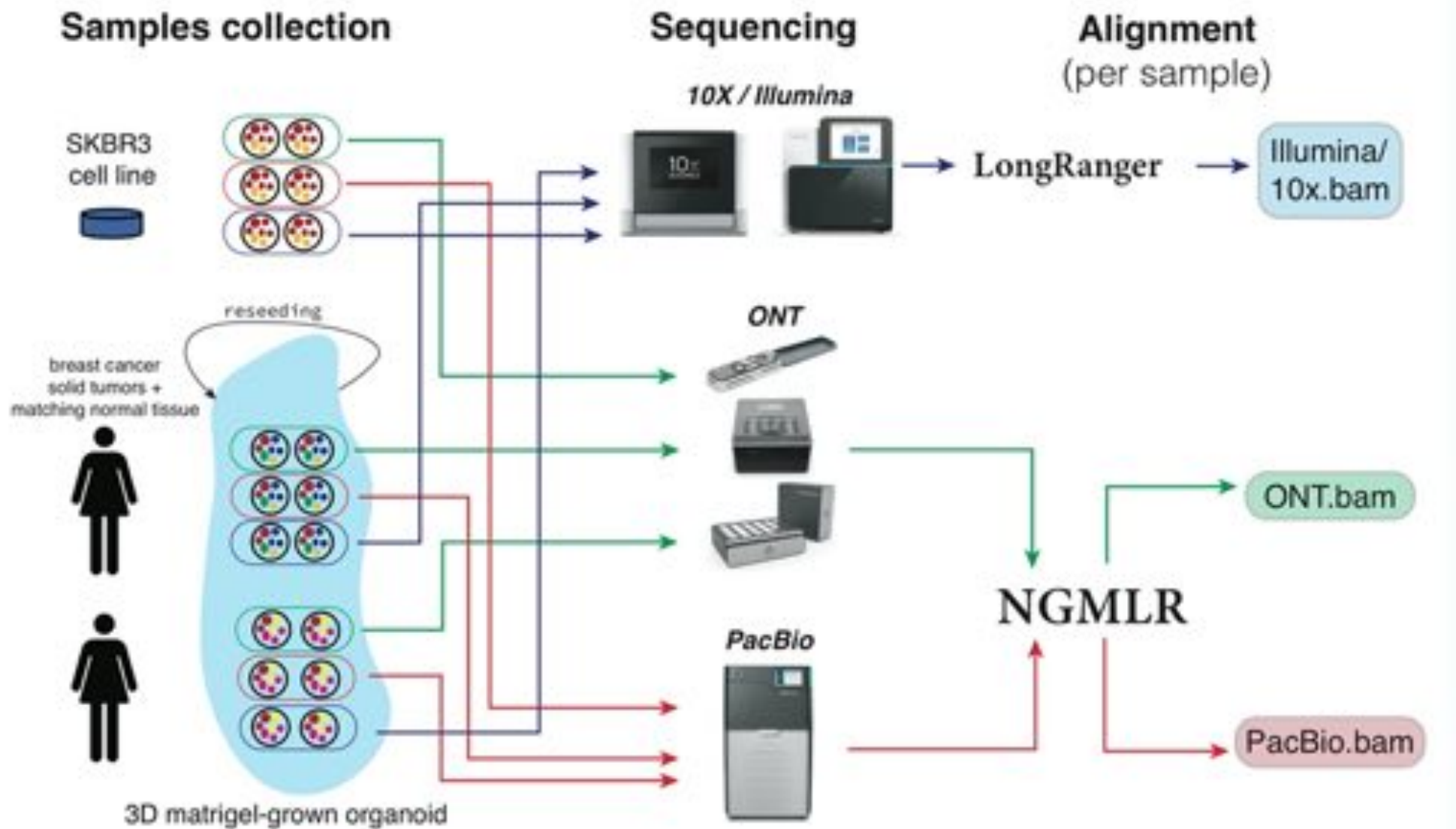


David Spector



Karen Kostroff

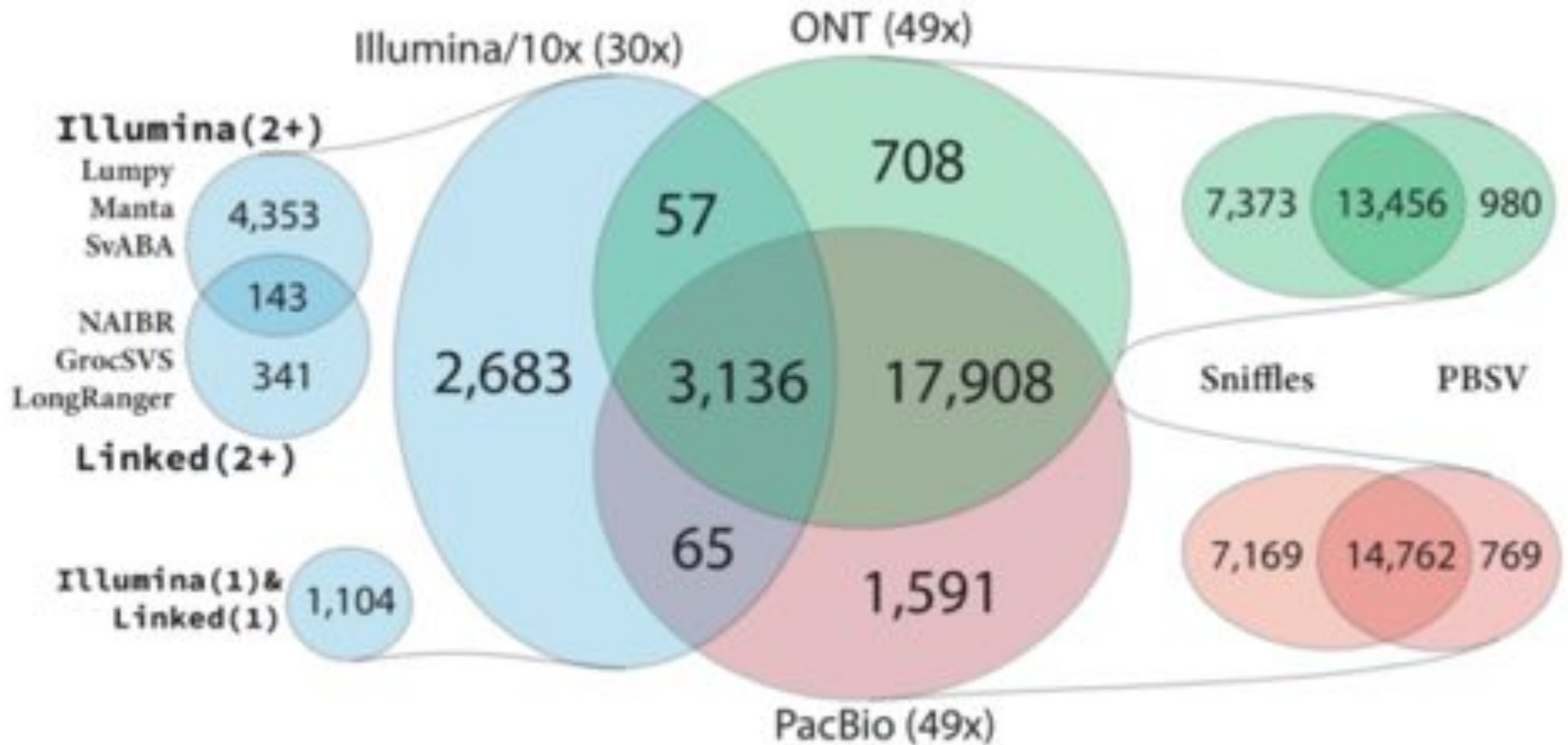
Data Production



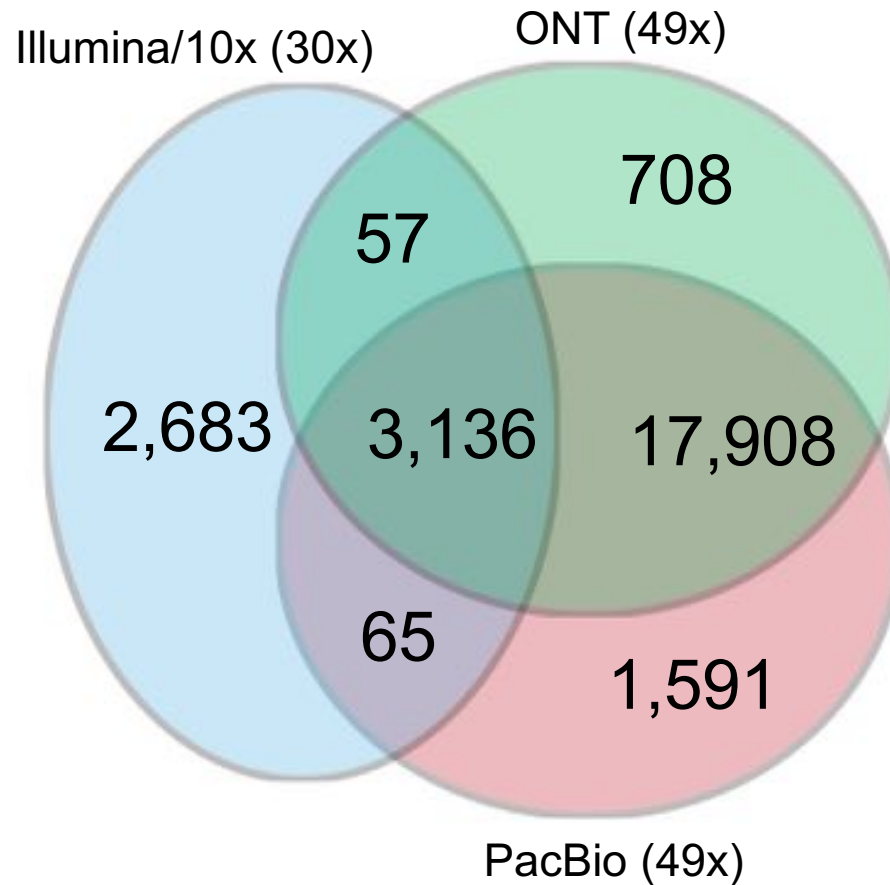
Comprehensive analysis of structural variants in breast cancer genomes using single molecule sequencing

Aganezov, S et al. (2020) *Genome Research*

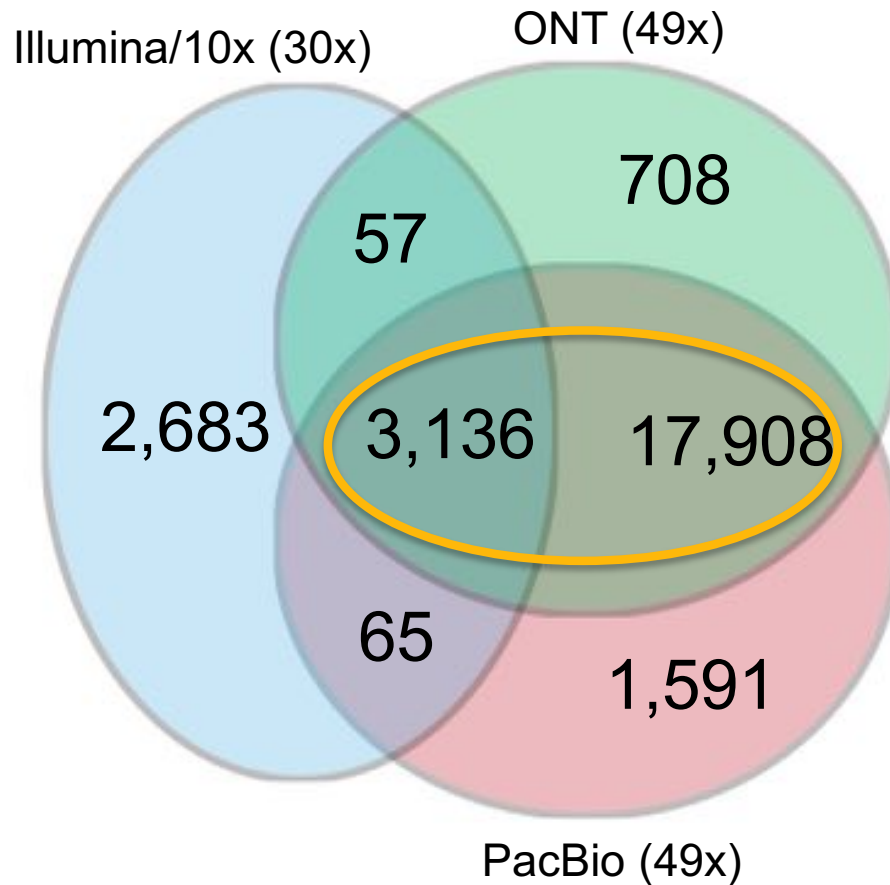
Structural Variation Consistency



Structural Variation Consistency

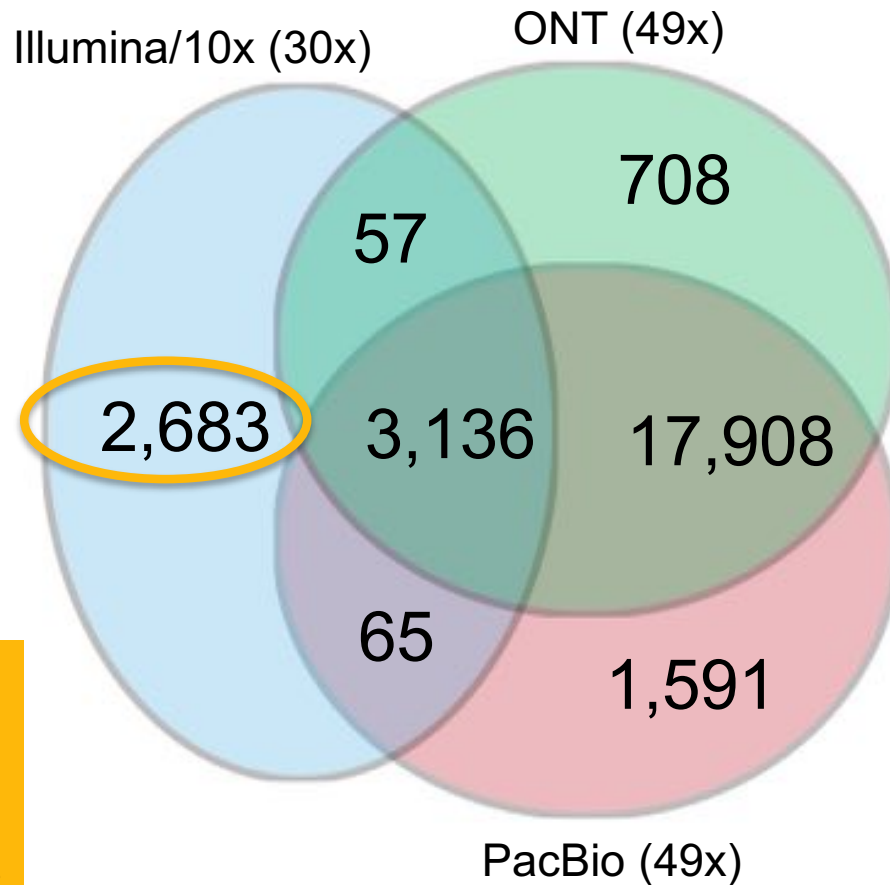


Structural Variation Consistency



- Very strong concordance between long read platforms
- Substantially more variants than detected by short reads

Structural Variation Consistency

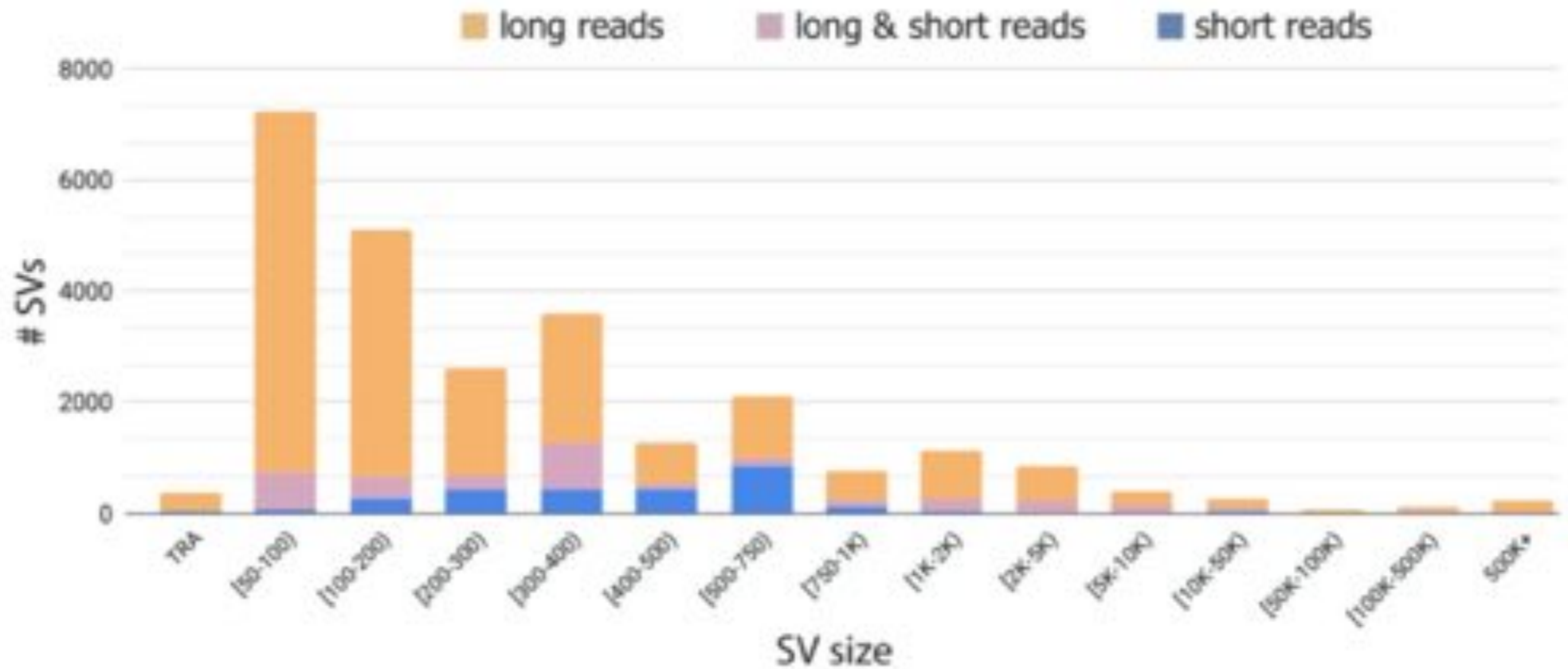


- PCR validation shows most Illumina-only calls are false positives

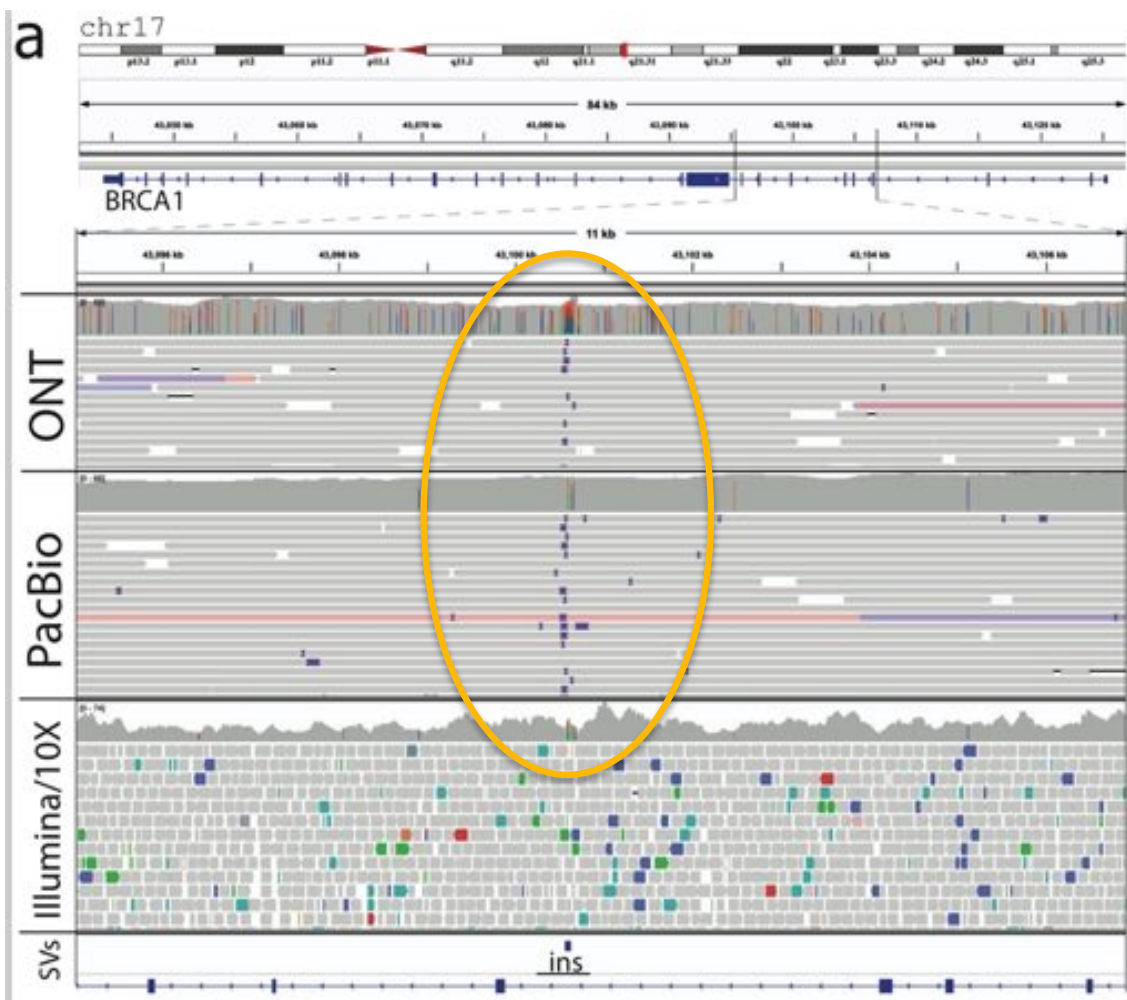
Structural Variation Consistency



Structural Variation Identification



Hidden Variants in Breast Cancer Genes



62bp repeat expansion in BRCA1 detected in normal tissue that is undetectable using a panel or short read sequencing

What causes “outlier” families?

