Lecture 23. Cancer Genomics & Beyond

Michael Schatz

November 18, 2024

Applied Comparative Genomics

Class Schedule

~~M Nov 4 Midterm!~~
~~W Nov 6 Human evolution Final Report Assigned~~

~~M Nov 11 Metagenomics~~
~~W Nov 13 No Class (BIODATA24)~~

M Nov 18 Cancer Genomics
W Nov 20 Project Presentation 1

M Nov 25 Thanksgiving Break
W Nov 27 Thanksgiving Break

M Dec 2 Project Presentation 2
W Dec 4 Project Presentation 3

M Dec 16 Project Report Due

Preliminary Report
Due Monday November 11

Preliminary Project Report

Assignment Date: October 28, 2024
Due Date: Monday, November 11, 2024 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Monday November 11

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 (typically a summary of the data you have identified for your project)
- 5+ References to relevant papers and data

The preliminary report must 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 November 25. You will also submit your final written report (6-8 pages) of your project by Dec 16

Please use Piazza if you have any questions!

Schedule of Presentations

Slot	Date	Start	Team Name	Team Members	Project Title
1	11/20	3:00	Two single-cells, one big problem	Kevin Meza Landeros, YunZhou Liu	Cluster-based single-cell RNA-seq variant detection
2	11/20	3:12	Team Yuxiang Li	Yuxiang Li	Contrastive Learning Approach to Integrate Single-Cell scRNA-seq and scATAC-seq for Mechanistic Understanding of Gene Regulation
3	11/20	3:24	Team Roujin An	Roujin An	Cell Type-Specific SNP-to-Splicing Variants Mapping Using Deep Learning Models
4	11/20	3:36	Team Miller	Logan Miller	Population-Specific Evolutionary Hotspots in Human Genomes
5	11/20	3:48	Team1D	Ben Miller	Comparative Genomic Analysis of NOD and (Simulated) NOR Mouse Genomes to Identify Variants Associated with Type 1 Diabetes
6	11/20	4:00	Genomic Visionaries	Iason Mihalopoulos, Siam Mohammed	AR/VR Visualization of Individual Genomes with AI-driven Insights
1	12/2	3:00	Silent Codebreakers	Cecelia Zhang, Jiarui Yang	Benchmarking Non-Coding Mutation Analysis Schemes on Cancer Genomes
2	12/2	3:12	Team Table	Oce Bohra, Zoe Rudnick	The emerging contribution of non-coding mutations in glioblastoma development
3	12/2	3:24	Team Brady	Brady Bock	DNN analysis of gut microbiomes to predict colorectal cancer disease state
4	12/2	3:36	Variant Visionaries	Alexandra Gorham, Christine Park, Natalie Vallejo	Benchmarking Non-coding Variant Scoring Tools for Cancer Pathogenicity Prediction
5	12/2	3:48	Human to Plants	Xiaojun Gao, Yujia, Yushan Zou	Evaluation of applicability of ChromHMM for Plants in Chromatin States and Gene Expression
1	12/4	3:00	SE Palmeiras	Caleb Hallinan, Jamie Moore, Rafael dos Santos Peixoto	Evaluating cell-type clustering algorithm's robustness to technical artifacts via synthetic spatial transcriptomics data
2	12/4	3:12	Nuclencoder	Amanda Xu, Angela Yang, Jiamin Li	DNA Cryptography: Digital Signatures for Encryption to Facilitate Safe Data Storage
3	12/4	3:24	Geoguessr	Alex Ostrovsky, Nicole Lauren Brown	Investigating geographic and environmental effects on soil metagenomes by correlating GIS data
4	12/4	3:36	Quetzalli Tlalli	Arshana Welivita, Atticus Colwell	Benchmarking Methods for Inferring the Ethnicity of an Individual from Their Genotype
5	12/4	3:48	Team Barbour	Alexis Barbour	Benchmarking non-coding mutation analysis schemes for evaulating Type 1 Diabetes
6	12/4	4:00	All of Us Team	Levon Galstyan, Nitish Aswani, Talia Haller	Genomic Insights into Sleep Patterns, Disease Outcomes, and Biomarker Associations using the All of Us Dataset

https://github.com/schatzlab/appliedgenomics2024/blob/main/project/presentations.md

Presentations!
10 min + 2 min questions

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: How does it compare to other methods/data/ideas?
10. Discussion 1: What did you learn from this study?
11. Discussion 2: What does this mean for the future?
12. Acknowledgements: Who helped you along the way?
13. 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.

Final Report
Due Monday December 16

Final Project Report

Assignment Date: November 6, 2024
Due Date: Monday, December 16, 2024 @ 11:59pm

Each team should submit a PDF of your final project proposal (6 to 9 pages) to GradeScope by 11:59pm on Monday December 16. No late days can be used as grades must be submitted to the registrar that week.

The report should have at least:

- Title of your project
- List of team members and email addresses
- 1-2 paragraph abstract summarizing the project
- 1-2 pages of Introduction: Background, what is the big problem/question you are addressing, overview of data used, summary of results
- 2-3 pages of Methods that you are using: if you are primarily using existing methods, please describe those methods
- 2-3 pages of Results: be sure to describe the data evaluated along with the results of your analysis. If computational time is measured, please list the machine specifications
- 1 page of Discussion: what you have seen or how that relates to other papers
- Please include 4-6 main figures showing your results. If you have more figures, please include them in a supplemental figures section at the end of the PDF.
- 1 paragraph of acknowledgements
- 1/2 to 1 page of references to relevant papers and data

The report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. You can (and should) expand on your preliminary report into the full report.

Please use Piazza if you have any questions!

https://github.com/schatzlab/appliedgenomics2024/blob/main/project/finalreport.md

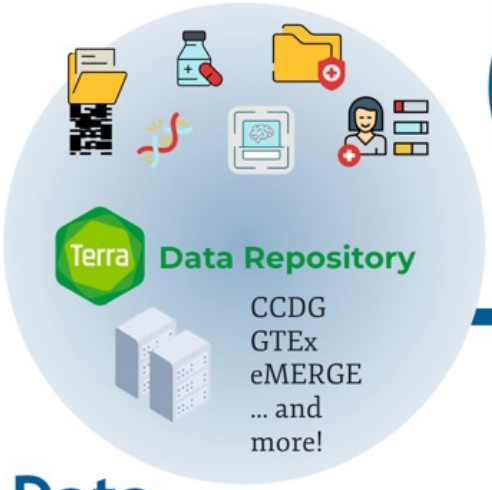
AnVIL Community Conference 2024

November 12-13, 2024

AnVIL

NHGRI AnVIL

Secure data storage, analysis, and sharing platform



Data

- Bring your structured data or use a flexible data schema
- Access public and managed-access data hosted in AnVIL

Community

- Share your analysis and data with collaborators or with the world
- Get support at help.anvilproject.org

Compute

- Leverage elastic, scalable compute
- Bring your own pipelines, find shared
- Use powerful bioinformatics platforms

AnVIL Community Conference Speakers

Tychele Turner

Anshul Kundaje

Benedict Paten

anvilproject.org

2024 Biological Data Science Conference

Arranged by

Elinor Karlsson, UMass Chan Medical School
Michael Schatz, Johns Hopkins University
Catalina Vallejos, University of Edinburgh, UK

#biodata24

2024 Biological Data Science

Tools & Visualization


Robert Carroll


Mentewab Ayalew


Ana Conesa


Jeremy Goecks


Fabio Cumbo


Gamze Gursoy

PopGen & Personalized Med


Jack Bowden


Sohini Ramachandran


Yuchen Zhou


Jackson Killian


Jaehee Kim


Andrew Ghazi


A. Jajoo


Chiraag Gohel

Algorithms


Victoria Popic


Tobias Marschall


Can Firtina


Sandy Kim


Katharine Jenike


Ashton Omdahl

Machine Learning


Genevieve Stein-O'Brien


Gunnar Ratsch


Peter Koo


Jessica L. Zhou


Abdul Muntakim Rafi


Gillian Chu


Yang Lu


Cristina Martin Linares

Spatial and Imaging


Christopher Mason


Michal Levo


Hyeyeon Hwang


Daniel Stein


Moritz Schaefer


Alyssa Obermayer

Single Cell & Functional Genomics


Athma Pai


Yoav Gilad


Robin Andersson


Radhika Jangi


Benjamin Parks


Xuan Li

Kasper Hansen

Irika Sinha

294 attendees + 66 virtual attendees | 6 Sessions | 2 Keynotes + 42 Talks | 184 posters

Ben Neale

Katie Pollard

Biological Data Science

Tools & Visualization

Robert Carroll

Mentewab Ayalew

Ana Conesa

Jeremy Goecks

Fabio Cumbo

Gamze Gursoy

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

#biodata24

nature methods

Brief Communication

https://doi.org/10.1038/s41592-024-02229-2

SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms

Received: 1 June 2023

Accepted: 1 March 2024

Published online: 20 March 2024

Check for updates

Francisco J. Pardo-Palacios[1,2,8], Angeles Arzalluz-Luque[1,2,8], Liudmyla Kondratova[3,4], Pedro Salguero[2], Jorge Mestre-Tomás[1], Rocío Amorín[4,5], Eva Estevan-Morió[1], Tianyuan Liu[1], Adalena Nanni[6], Lauren McIntyre[4,6], Elizabeth Tseng[7] & Ana Conesa[1]

SQANTI3 is a tool designed for the quality control, curation and annotation of long-read transcript models obtained with third-generation sequencing technologies. Leveraging its annotation framework, SQANTI3 calculates quality descriptors of transcript models, junctions and transcript ends. With this information, potential artifacts can be identified and replaced with reliable sequences. Furthermore, the integrated functional annotation feature enables subsequent functional iso-transcriptomics analyses.

Long-read sequencing, driven by biotechnology companies such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), was recognized as the method of the year 2022 by *Nature Methods*[1–3] for providing single-molecule reads spanning thousands of bases and advancing genomics and transcriptomics research. When applied to gene expression analysis, long-read RNA sequencing (lrRNA-seq) has the potential to capture full-length transcripts and elucidate isoform diversity in both normal and disease conditions[4–6]. Software tools have been developed for lrRNA-seq-based transcript identification[7–9], quantification[10,11], differential splicing analysis[6] and functional interpretation[12].

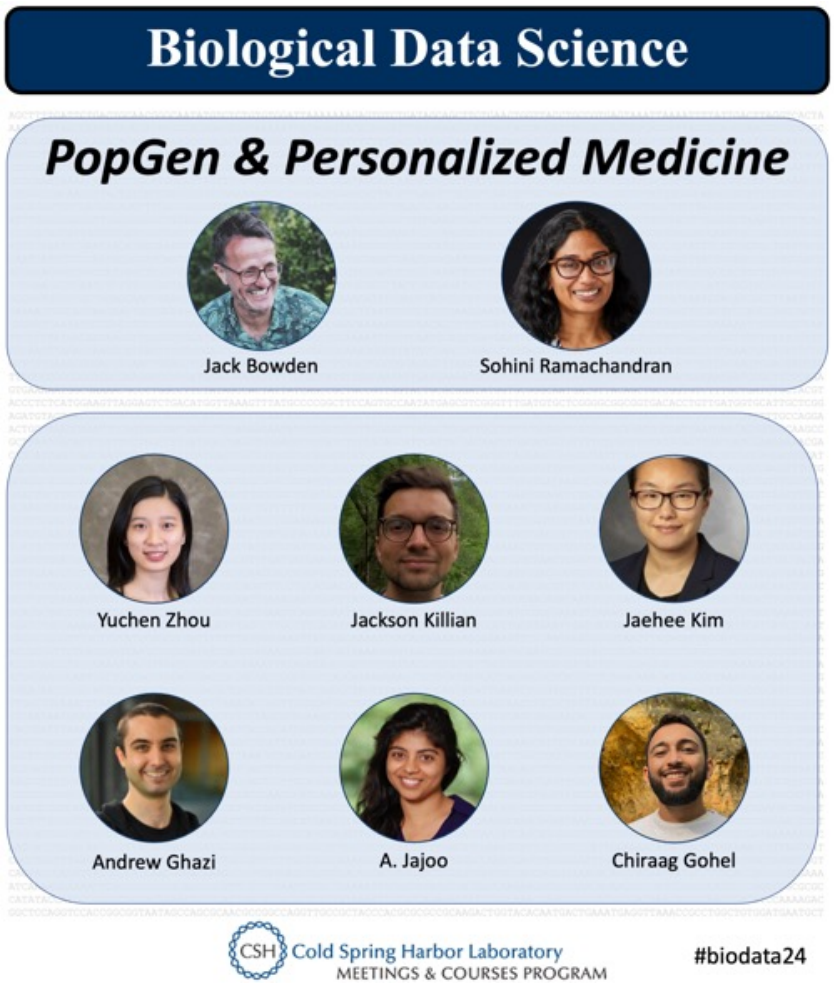
One of the most striking results in lrRNA-seq studies is the identification of thousands of novel transcripts, even in well-annotated genomes[5,10,13]. However, long-read technologies are error prone, and biases due to RNA degradation, library preparation issues and sequencing errors, as well as read mapping and transcript reconstruction inaccuracies, often lead to false transcript identification. Several studies have evaluated the accuracy of lrRNA-seq methods and algorithms[14–17]. These works have consistently highlighted significant disagreements between experimental and computational approaches at identifying transcripts from long-read data, especially for novel transcripts not present in the reference annotations. Disagreements involve the annotation of splice junctions and the definition of transcription start sites (TSS) and transcription termination sites (TTS)[16], which are particularly difficult to discriminate from RNA degradation in the sequenced samples. Given the large number of novel isoforms reported by most lrRNA-seq studies, quality control and curation of the data are crucial steps in long read-based transcriptome definition.

We hereby present SQANTI3, a tool for the evaluation of long-read transcript models used as an evaluation engine in the LRGASP (Long-read RNA-seq Genome Annotation Assessment Project)[16]. SQANTI3 builds on SQANTI[18], a widely used tool for quality control of lrRNA-seq data (a comparison of SQANTI and SQANTI3 functionality is given in Supplementary Note 1).

The SQANTI3 workflow consists of three modules (Fig. 1a). First, quality control (QC) classifies long-read transcript models according to SQANTI3 structural categories, which consist of the SQANTI splice junction-based transcript classes: full-splice-match (FSM); incomplete-splice-match (ISM); novel-in-catalog (NIC); novel-not-in-catalog (NNC); antisense; fusion; genic genomic; and intergenic (Fig. 1b); and novel subcategories based on TSS and TTS annotations (Fig. 1c).

Of these, 'reference match' is defined as an FSM transcript in which both the 3′ and 5′ ends are within 50 bp of the reference transcript's

[1]Institute for Integrative Systems Biology, Spanish National Research Council, Paterna, Valencia, Spain. [2]Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Valencia, Valencia, Spain. [3]Horticultural Sciences Department, University of Florida, Gainesville, FL, USA. [4]Genetics Institute, University of Florida, Gainesville, FL, USA. [5]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL, USA. [6]Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, USA. [7]Pacific Biosciences, Menlo Park, CA, USA. [8]These authors contributed equally: Francisco J. Pardo-Palacios, Angeles Arzalluz-Luque. e-mail: ana.conesa@csic.es



Genetics identifies obesity as a shared risk factor for co-occurring multiple long-term conditions.

Short title: Genetics pinpoints obesity as a shared risk factor between diseases

Ninon Mounier [1], Bethany Voller [1], Jane AH Masoli [1,2], João Delgado [1], Frank Dudbridge [3], Luke C Pilling [1], Timothy M Frayling [1,4], Jack Bowden [1,5], on behalf of the GEMINI Consortium

Affiliations

1. Department of Clinical and Biomedical Sciences, Faculty of Health and Life Sciences, University of Exeter, UK
2. Royal Devon University Healthcare NHS Foundation Trust, Barrack Road, Exeter, EX2 5DW, UK
3. Department of Population Health Sciences, University of Leicester, UK
4. Department of Genetic Medicine and Development, Faculty of Medicine, 1 rue Michel-Servet, CH-1211 Genève 4, Switzerland
5. Novo Nordisk Research Centre, Roosevelt Drive, Headington, Oxford, UK

Corresponding author

Prof Jack Bowden (J.Bowden2@exeter.ac.uk)

Biological Data Science

Algorithms

Victoria Popic

Tobias Marschall

Can Firtina

Sandy Kim

Katharine Jenike

Ashton Omdahl

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

#biodata24

Solanum pan-genomics and pan-genetics reveal paralogs as contingencies in crop engineering

Matthias Benoit[1,*,†], Katharine M. Jenike[2,3*], James W. Satterlee[1,4,o], Srividya Ramakrishnan[3,o], Iacopo Gentile[5,o], Anat Hendelman[1,4,o], Michael J. Passalacqua[5], Hamsini Suresh[5], Hagai Shohat[4], Gina M. Robitaille[1,4], Blaine Fitzgerald[1,4], Michael Alonge[3,†], Xingang Wang[4,†], Ryan Santos[4,†], Jia He[1,4], Shujun Ou[3,†], Hezi Golan[6], Yumi Green[7], Kerry Swartwood[7], Gina P. Sierra[8], Andres Orejuela[9], Federico Roda[8], Sara Goodwin[4], W. Richard McCombie[4], Elizabeth B. Kizito[10], Edeline Gagnon[11,12,†], Sandra Knapp[13], Tiina E. Särkinen[12], Amy Frary[14], Jesse Gillis[4,15,#], Joyce Van Eck[7,16,#], Michael C. Schatz[2,3,#], Zachary B. Lippman[1,4,5,#]

1. Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
2. Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA
3. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
4. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
5. School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
6. SiteKicks.ai, Setauket, NY, USA
7. Boyce Thompson Institute, Ithaca, NY, USA
8. Max Planck Tandem Group, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia
9. Departamento de Biología, Facultad de Ciencias Exactas y Naturales, Universidad de Cartagena, Cartagena de Indias, Colombia
10. Faculty of Agricultural Sciences, Uganda Christian University, Mukono, Uganda
11. Department of Integrative Biology, University of Guelph, Ontario, Canada
12. Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK
13. Natural History Museum, London, UK
14. Department of Biological Sciences, Mount Holyoke College, South Hadley, MA, USA

Biological Data Science

Machine Learning

Genevieve Stein-O'Brien

Gunnar Ratsch

Peter Koo

Jessica L. Zhou

Abdul Muntakim Rafi

Gillian Chu

Yang Lu

Cristina Martin Linares

Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

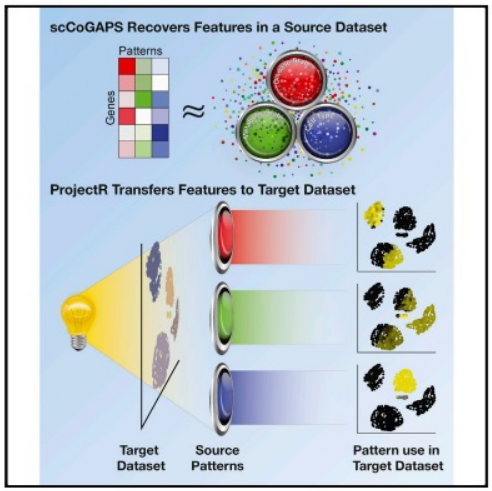
#biodata24

Cell Systems

Article

Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species

Graphical Abstract



Authors

Genevieve L. Stein-O'Brien, Brian S. Clark, Thomas Sherman, ..., Seth Blackshaw, Loyal A. Goff, Elana J. Fertig

Correspondence

ejfertig@jhmi.edu

In Brief

We present tools and workflows for latent space exploration across datasets. scCoGAPS is an implementation of NNMF that is specifically suited for large, sparse scRNA-seq datasets. ProjectR implements a transfer-learning framework that rapidly projects new data into learned latent spaces. We demonstrate the utility of this approach for *de novo* annotation of new datasets, cross-species analysis, linking genomic regulatory and transcriptional signatures, and exploration of features across a catalog of cell types.

Highlights

- Latent spaces provide greater insight into biological systems than marker genes alone
- scCoGAPS learns biologically meaningful latent spaces from sparse scRNA-Seq data
- Transfer learning (TL) enables discovery across experimental systems and species
- ProjectR is a TL framework to rapidly explore latent spaces across independent datasets

Stein-O'Brien et al., 2019, Cell Systems *8*, 395–411
May 22, 2019
https://doi.org/10.1016/j.cels.2019.04.004

CellPress

Biological Data Science

Spatial and Imaging

Christopher Mason

Michal Levo

Hyeyeon Hwang

Daniel Stein

Moritz Schaefer

Alyssa Obermayer

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

#biodata24

Biological Data Science

Single Cell & Functional Genomics

Athma Pai

Yoav Gilad

Robin Andersson

Radhika Jangi

Benjamin Parks

Xuan Li

Kasper Hansen

Irika Sinha

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

#biodata24

Cell-type and dynamic state govern genetic regulation of gene expression in heterogeneous differentiating cultures

Joshua M. Popp[1†], Katherine Rhodes[2†], Radhika Jangi[3], Mingyuan Li[3], Kenneth Barr[2], Karl Tayeb[4], Alexis Battle[1,5,6*], Yoav Gilad[2,7,8*]

[1] Department of Biomedical Engineering, Johns Hopkins University; Baltimore, MD, 21218.

[2] Department of Medicine, University of Chicago; Chicago, IL, 60637.

[3] Department of Biology, Johns Hopkins University; Baltimore, MD, 21218.

[4] Committee on Genetics, Genomics, and Systems Biology, University of Chicago; Chicago, IL, 60637.

[5] Department of Computer Science, Johns Hopkins University; Baltimore, MD, 21218.

[6] Department of Genetic Medicine, Johns Hopkins University; Baltimore, MD, 21218.

[7] Department of Human Genetics, University of Chicago; Chicago, IL, 60637.

[8] Lead contact.

*Corresponding authors. Email: ajbattle@jhu.edu and gilad@uchicago.edu

† These authors contributed equally to this work

Abstract

Identifying the molecular effects of human genetic variation across cellular contexts is crucial for understanding the mechanisms underlying disease-associated loci, yet many cell-types and developmental stages remain underexplored. Here we harnessed the potential of heterogeneous differentiating cultures (**HDCs**), an *in vitro* system in which pluripotent cells asynchronously differentiate into a broad spectrum of cell-types. We generated HDCs for 53 human donors and collected single-cell RNA-sequencing data from over 900,000 cells. We identified expression quantitative trait loci in 29 cell-types and characterized regulatory dynamics across diverse differentiation trajectories. This revealed novel regulatory variants for genes involved in key developmental and disease-related processes while replicating known effects from primary tissues, and dynamic regulatory effects associated with a range of complex traits.

Introduction

Decoding the molecular consequences of genetic variation is a central goal in human genetics. With the advent of genome-wide association studies, a vast array of genetic variants associated with diseases have been uncovered. These predominantly lie in non-coding regions of the genome, suggesting primarily regulatory mechanisms [1]. This insight has spurred a surge in mapping expression quantitative trait loci (eQTLs) to understand how the disease-associated genetic variants influence gene expression levels. Despite significant strides made by several large-scale projects such as the GTEx Consortium to map eQTLs [2–7], a comprehensive understanding of the molecular impacts of disease-associated loci remains elusive, in part due to the context-dependent and dynamic nature of gene regulation [8–11].

Ben Neale, Ph.D.
Associate Professor
Broad Institute
Massachusetts General Hospital

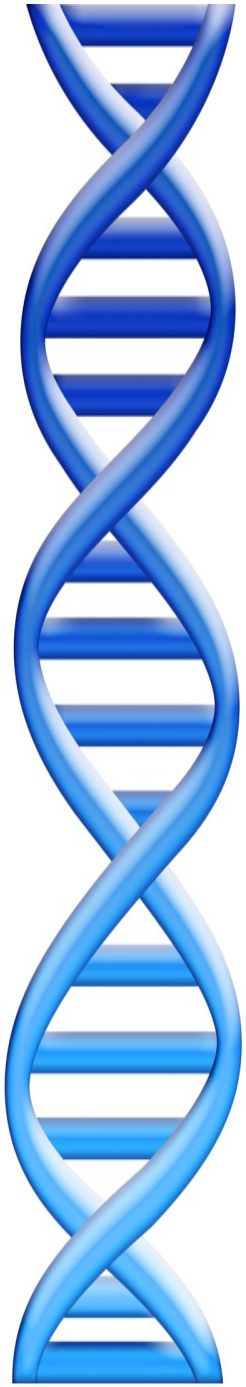
"Genetic Analysis at Scale"
Thursday @ 4pm ET

Katie Pollard, Ph.D.
Professor & Director
Gladstone Institutes, UCSF
Chan Zuckerberg Biohub

"Sequence-to-activity modeling"
Friday @ 4pm ET

#biodata24 Reflections

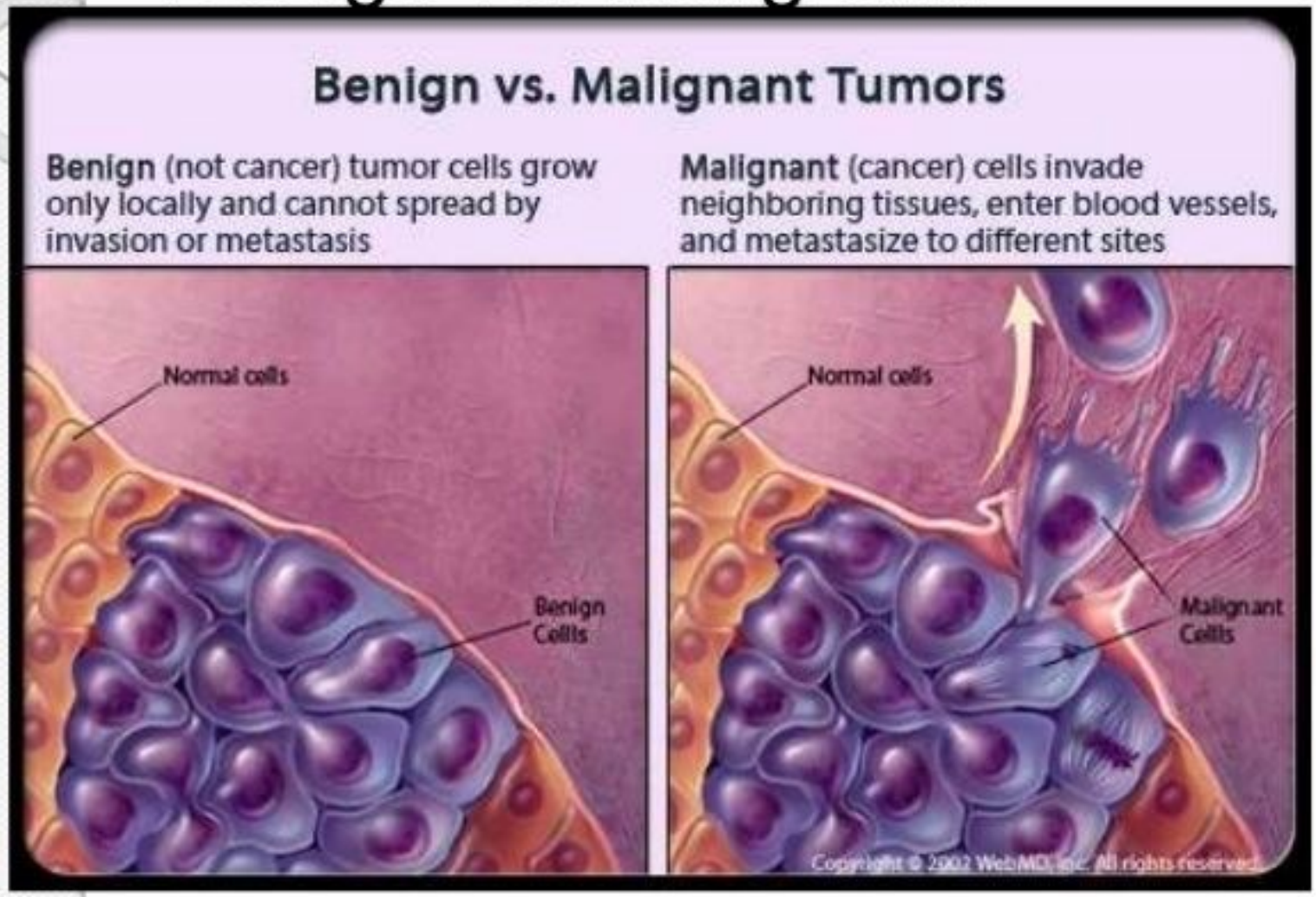
1. Deep learning has arrived. Every session and many talks & posters featured the use of deep learning
2. Healthy debate on the use of foundation models vs more specialized models. Foundation models are potentially more flexible but are harder to train and interpret; specialized models require more feature engineering and thought into the architecture but can be more efficient and more robust for their dedicated tasks
3. Widespread interest to use LLM to make analysis more accessible: prompt an LLM in English to generate code to explore a dataset; use an LLM to autogenerate descriptions of genes, genomes, workflows, etc. Just make sure to validate the outputs – human experts are still needed :-)
4. Core methods for "standard analysis" are mature (genome sequencing, bulk & scRNAseq, ATAC-seq, etc). Focus on dynamics and interpretation: mutations (SNPs & SVs), expression, & epigenetics over time/space/cell types/patient outcomes, etc
5. The unreasonable effectiveness of data still prevails – we can (often) overcome noisy datasets by increasing their size. Many of the most exciting results came from data integration over enormous scales (biobanks, cell atlases, etc)
6. Working at scale is hard. Need to optimize the data structures to shrink the data footprint and use streaming algorithms and federation to perform analyses without loading the entire dataset into RAM
7. Cloud resources are becoming more mainstream (Galaxy, All of Us, AnVIL, etc) but cost remains a concern (except Galaxy!)
8. As if there was ever any doubt - this meeting showcased time after time how critical data science is to biology and medicine
9. The #biodata24 community is incredible! Together we have strength beyond what any one person can contribute; Together we will have the strength to persevere through any challenge
10. See you all at #biodata26 on November 3-7, 2026 at CSHL!



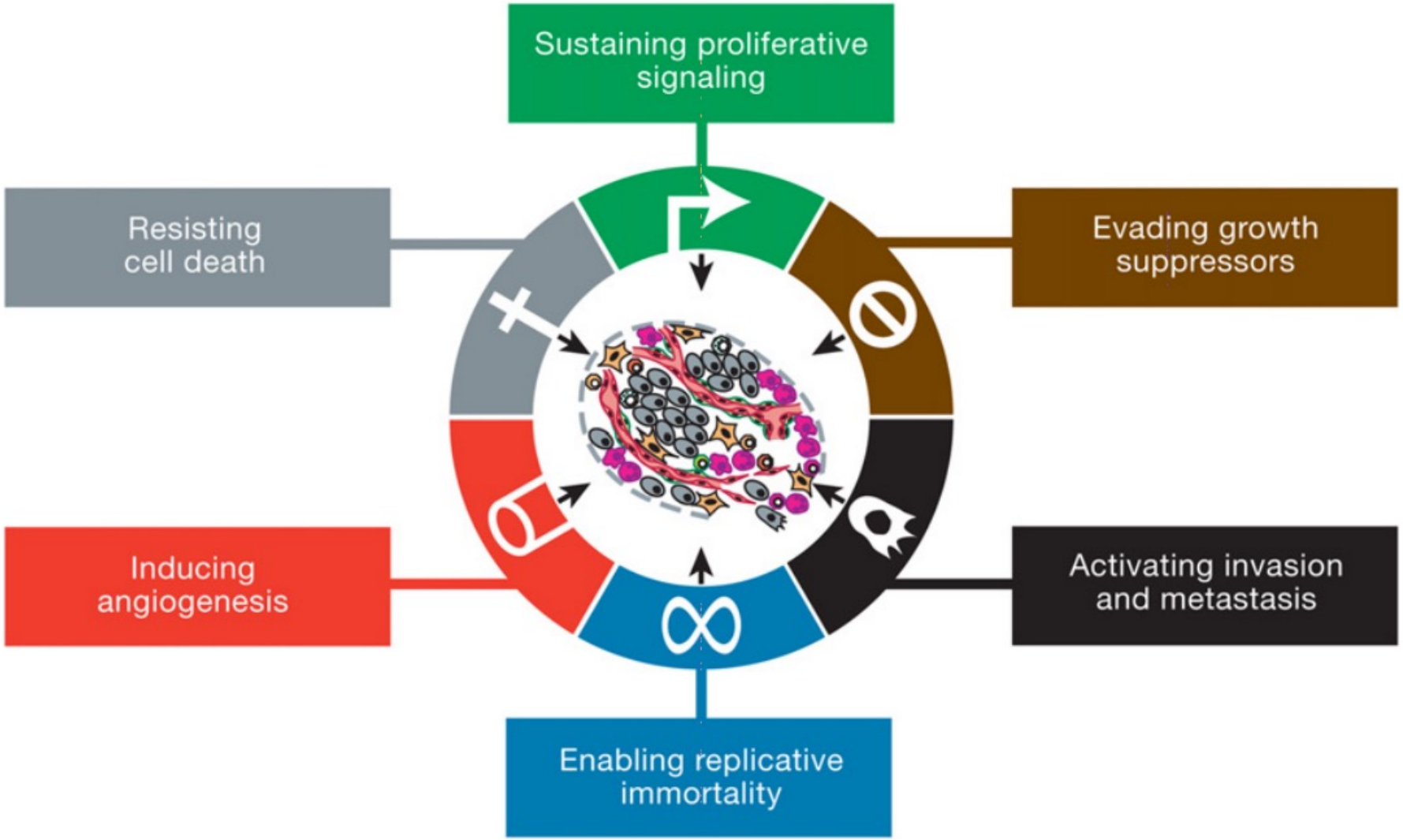
Cancer Genetics & Genomics

A tumor removed by surgery in 1689.

Benign vs. Malignant

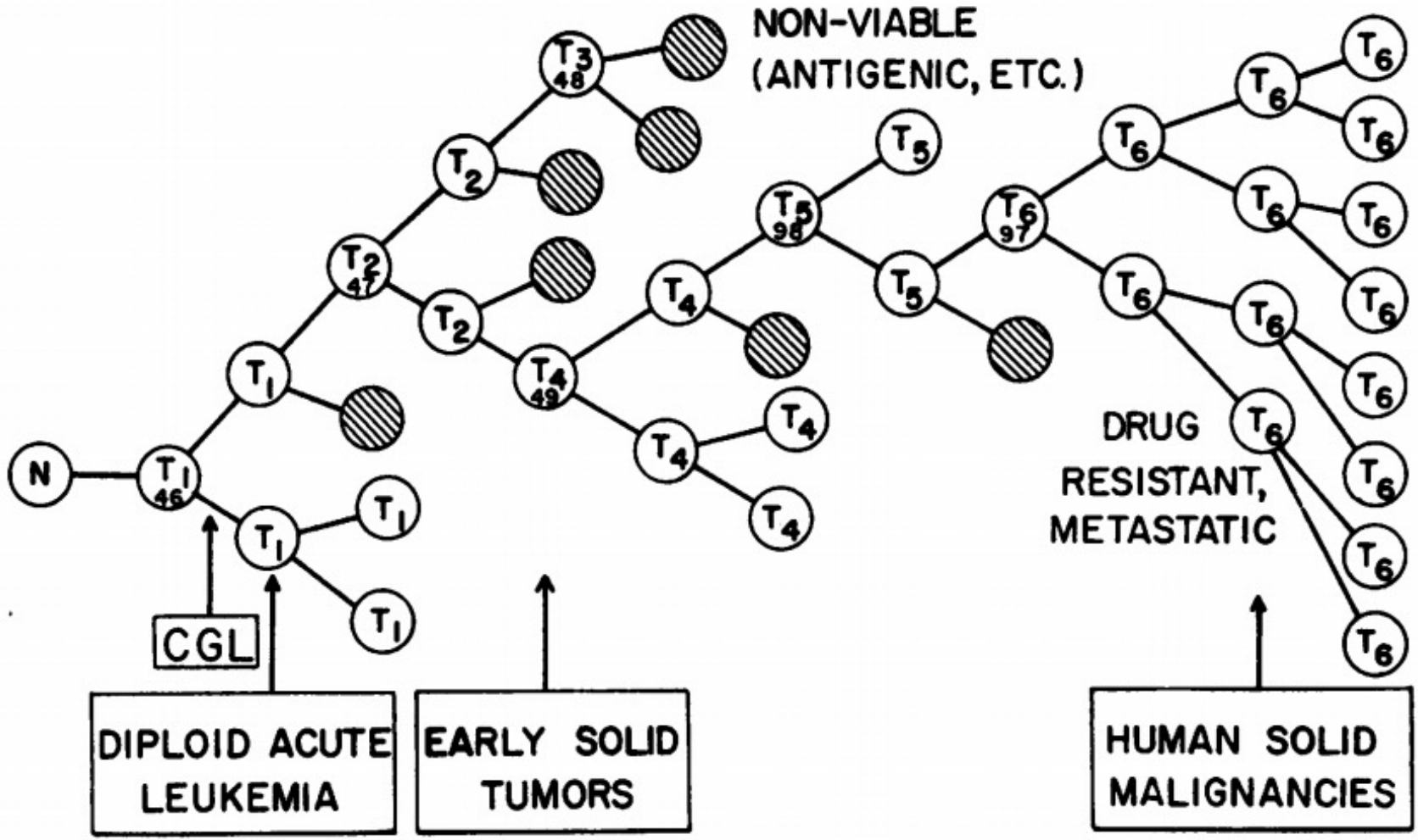


The Six Hallmarks of Cancer



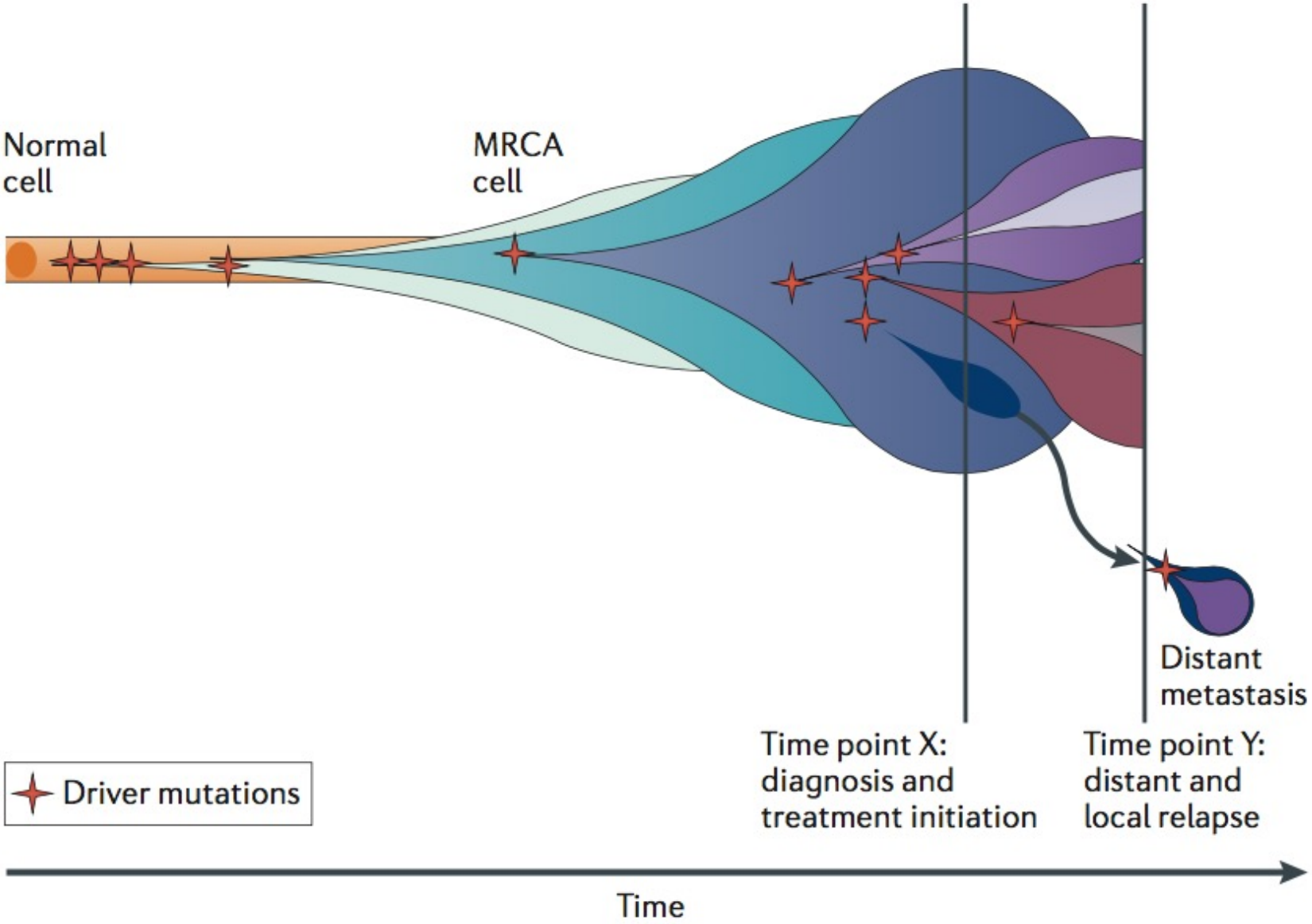
Hallmarks of Cancer
Hanahan and Weinberg (2000) Cell. http://doi.org/10.1016/S0092-8674(00)81683-9

Tumor Evolution



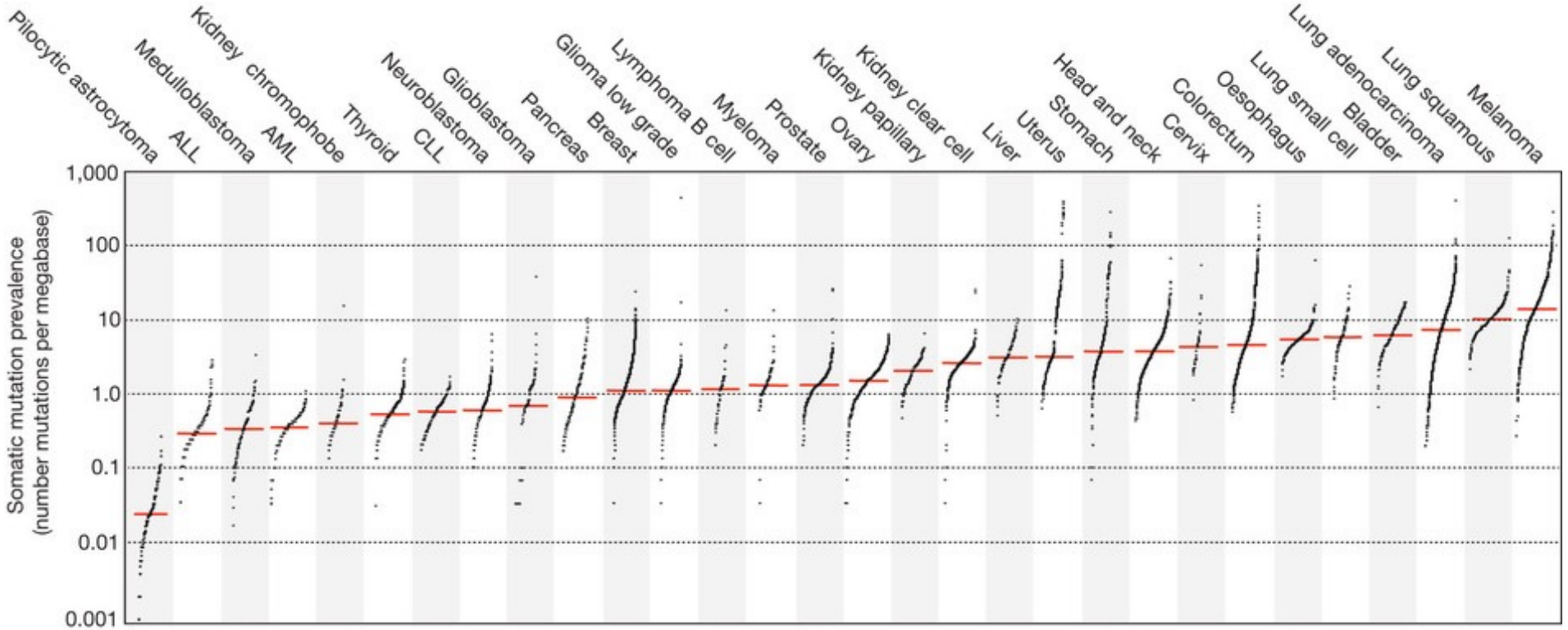
The Clonal Evolution of Tumor Cell Populations
Peter C. Nowell (1976) *Science*. 194(4260):23-28 DOI: 10.1126/science.959840

Tumor Evolution



Evolution of the cancer genome
Yates & Campbell (2012) Nature Review Genetics. doi:10.1038/nrg3317

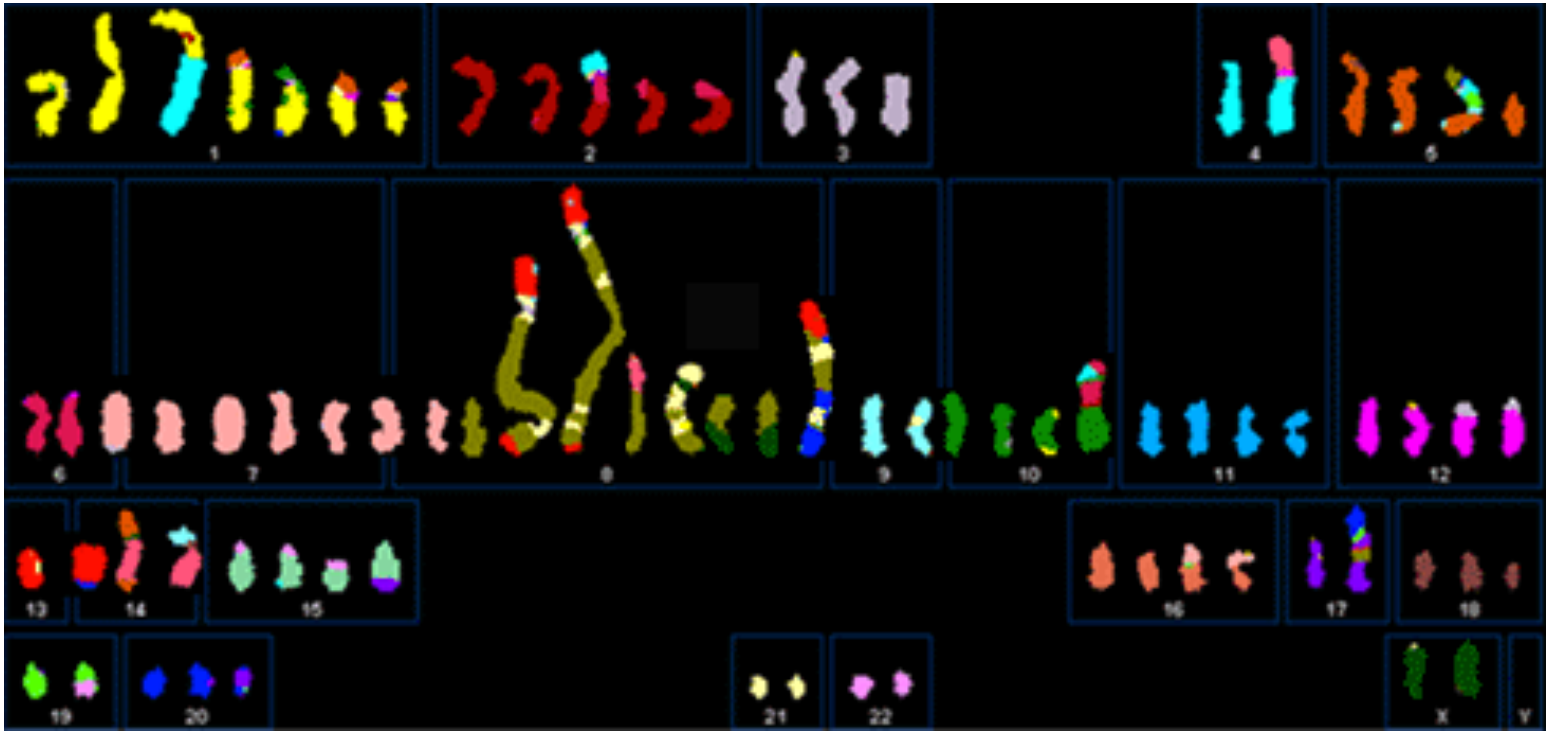
Somatic Mutations In Cancer



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

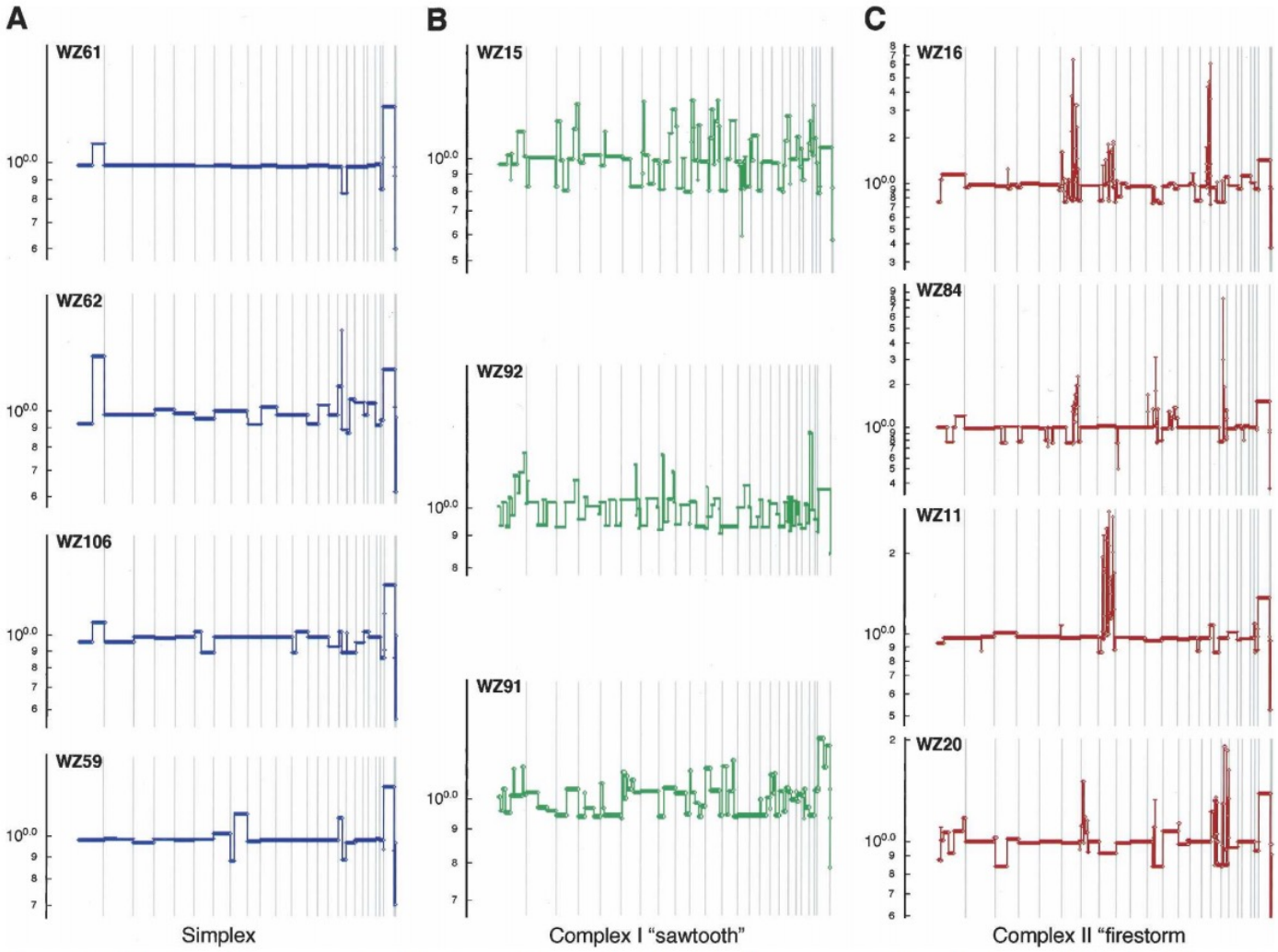
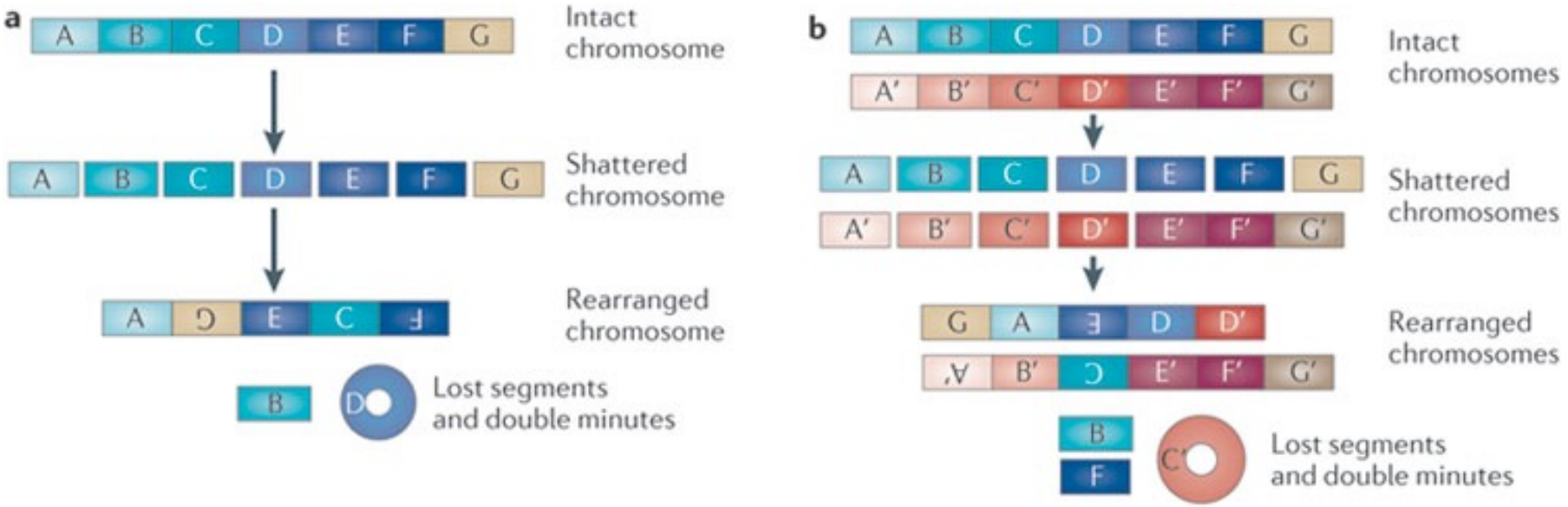


Figure 2. Major types of tumor genomic profiles. Segmentation profiles for individual tumors representing each category: (*A*) simplex; (*B*) complex type I or sawtooth; (*C*) complex type II or firestorm. Scored events consist of a minimum of six consecutive probes in the same state. The *y*-axis displays 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[3–5] and changes in methylation[6–8] 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[8,12–30]. The results of these studies have varied; depending on the techniques and systems used, an increase[12–19], decrease[20–24], or no change[25–29] 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.

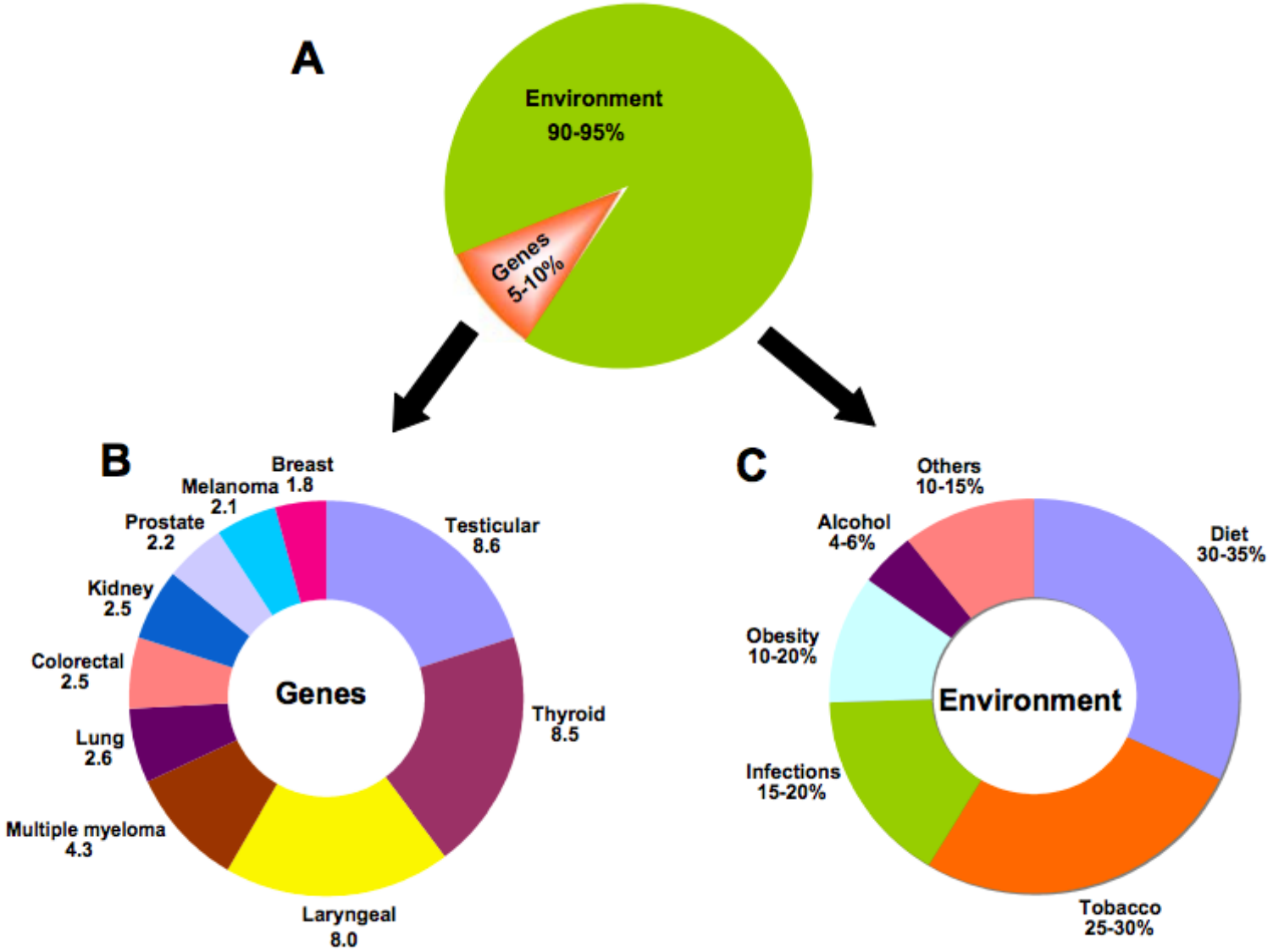
The first cancer studied was a grade D (ref. 43), moderately well differentiated adenocarcinoma of the colon from a 67-yr-old male. Tissue was obtained from the cancer itself and also from colonic mucosa stripped from the colon at a site just outside the histologically proven tumour margin. Figure 1 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	*Hpa*II	<10	35	—
			*Hha*I	<10	39	—
		γ-Globin	*Hpa*II	<10	52	—
			*Hha*I	<10	39	—
		α-Globin	*Hpa*II	<10	<10	—
			*Hha*I	<10	<10	—
2	Colon	HGH	*Hpa*II	<10	76	—
			*Hha*I	<10	85	—
		γ-Globin	*Hpa*II	<10	58	—
			*Hha*I	<10	23	—
		α-Globin	*Hpa*II	<10	<10	—
			*Hha*I	<10	<10	—
3	Colon	HGH	*Hpa*II	<10	41	—
			*Hha*I	<10	38	—
		γ-Globin	*Hpa*II	<10	50	—
			*Hha*I	<10	22	—

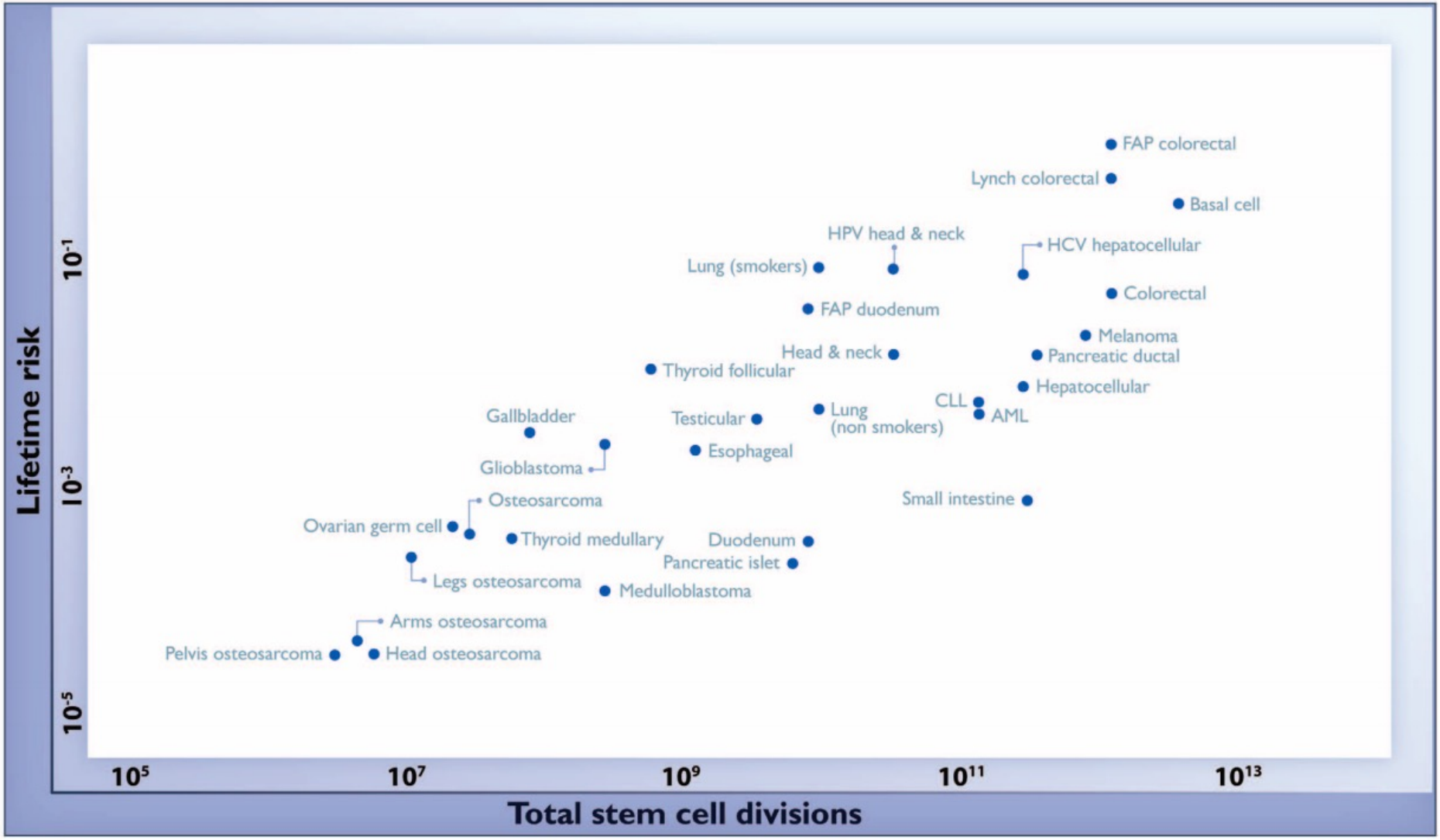
Nature 301, 89 - 92 (06 January 1983); doi:10.1038/301089a0

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

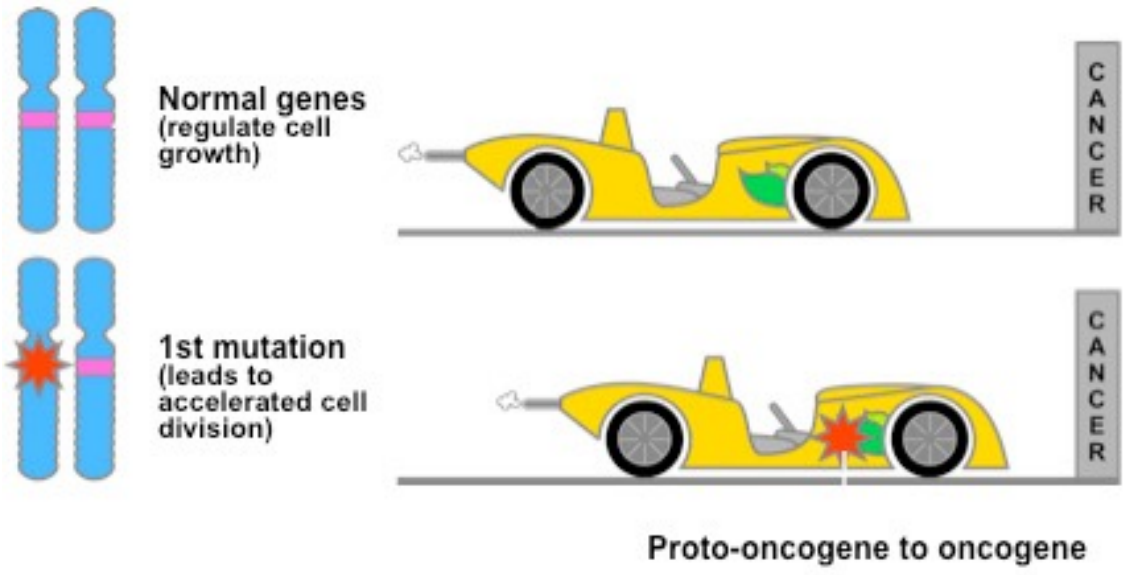


FAP = Familial Adenomatous Polyposis ◆ HCV = Hepatitis C virus ◆ HPV = Human papillomavirus ◆ CLL = Chronic lymphocytic leukemia ◆ AML = Acute myeloid leukemia

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.

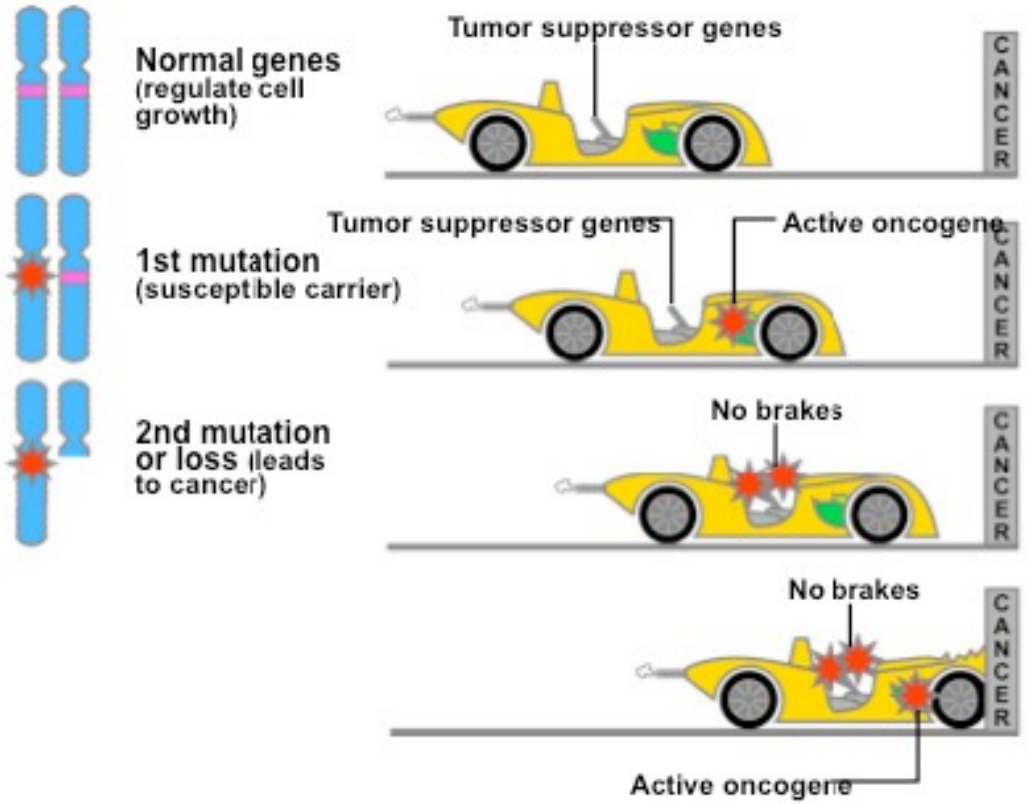
Variation in cancer risk among tissues can be explained by the number of stem cell divisions
Tomasetti and Vogelstein (2015) Science. DOI: 10.1126/science.1260825

Oncogenes



- ***HER-2/neuHER-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-tumorogenic Akt activation.

TP53: The first and most important tumor suppressor

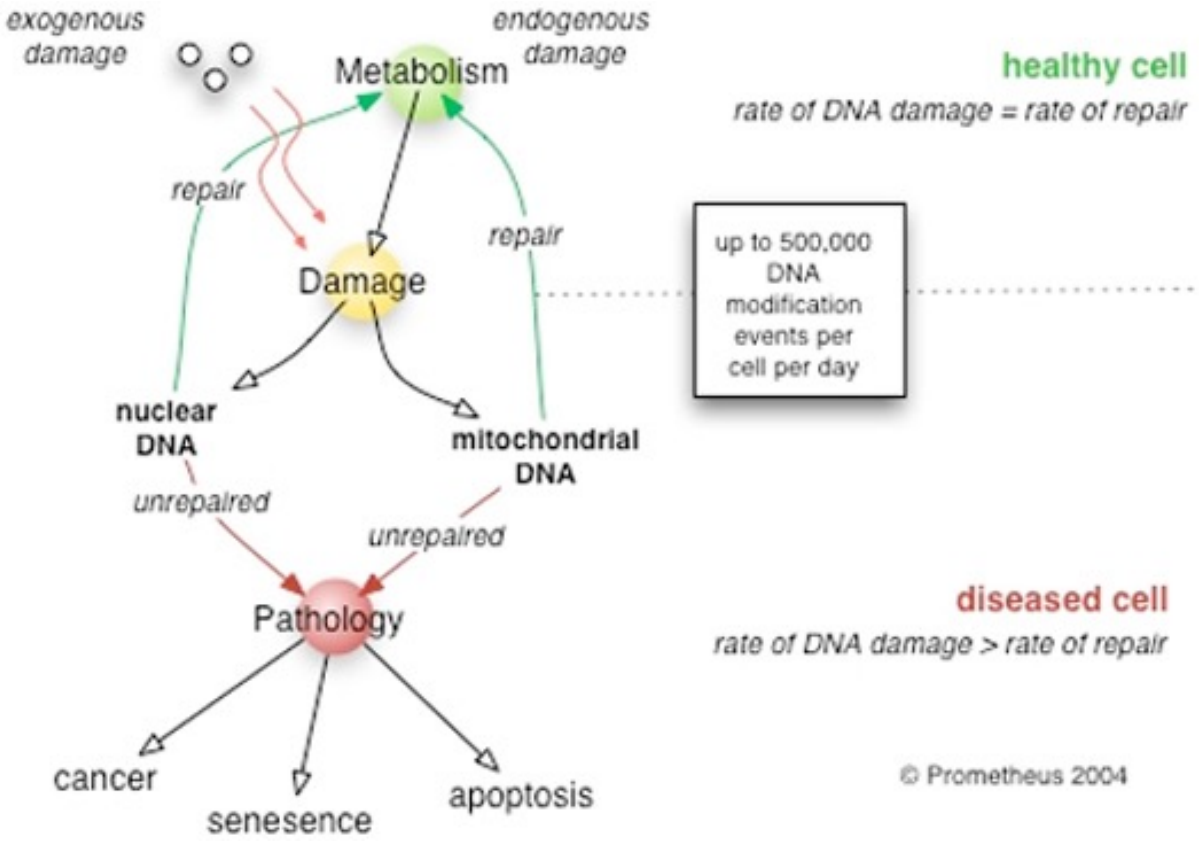
Mechanism of inactivating p53	Typical tumours	Effect of inactivation
Amino-acid-changing mutation in the DNA-binding domain	Colon, breast, lung, bladder, brain, pancreas, stomach, oesophagus and many others	Prevents p53 from binding to specific DNA sequences and activating the adjacent genes
Deletion 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, expecially 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

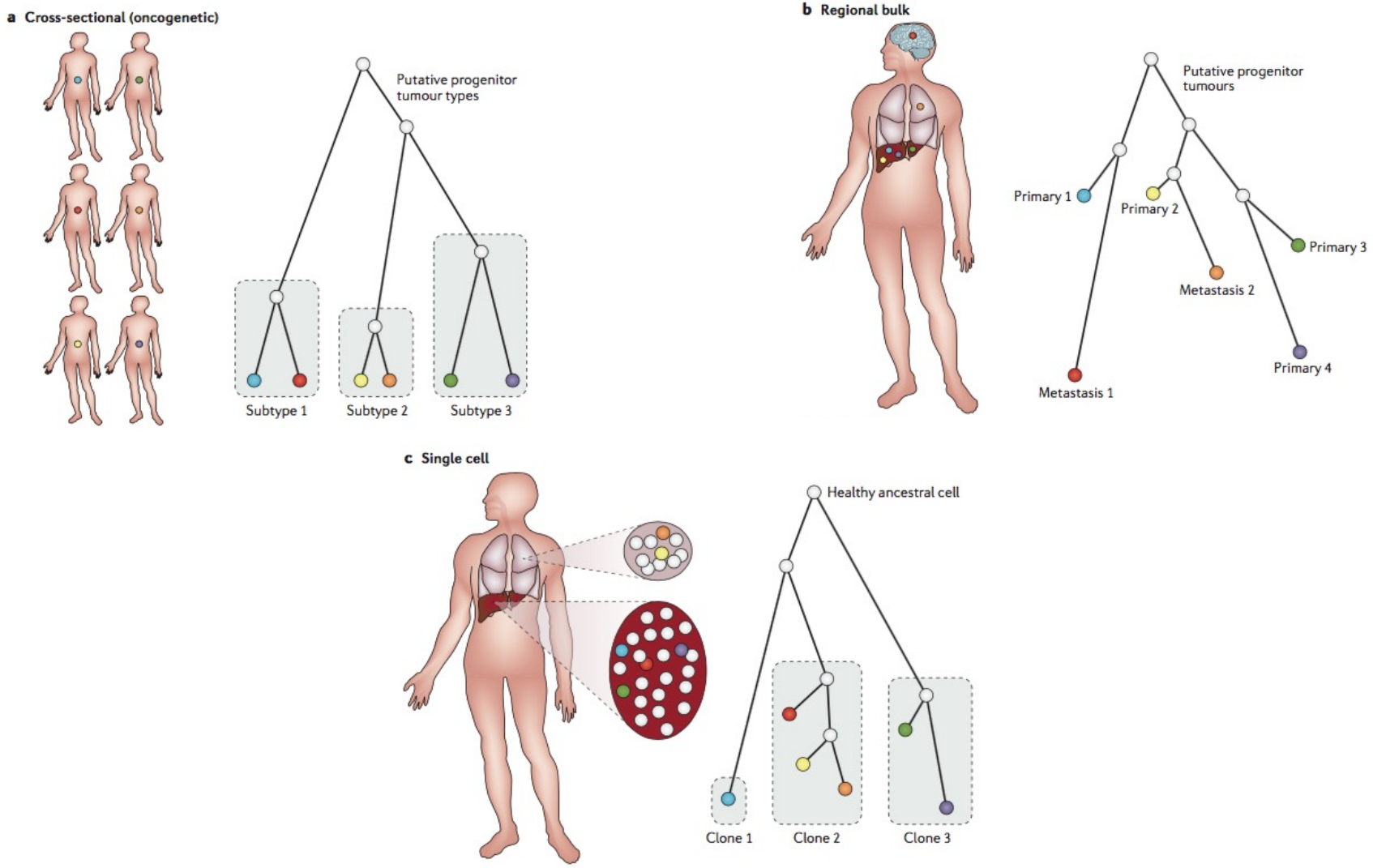
Surfing the p53 network
Volgelstein et al (2000) Nature. DOI: 10.1038/35042675

DNA Repair Genes



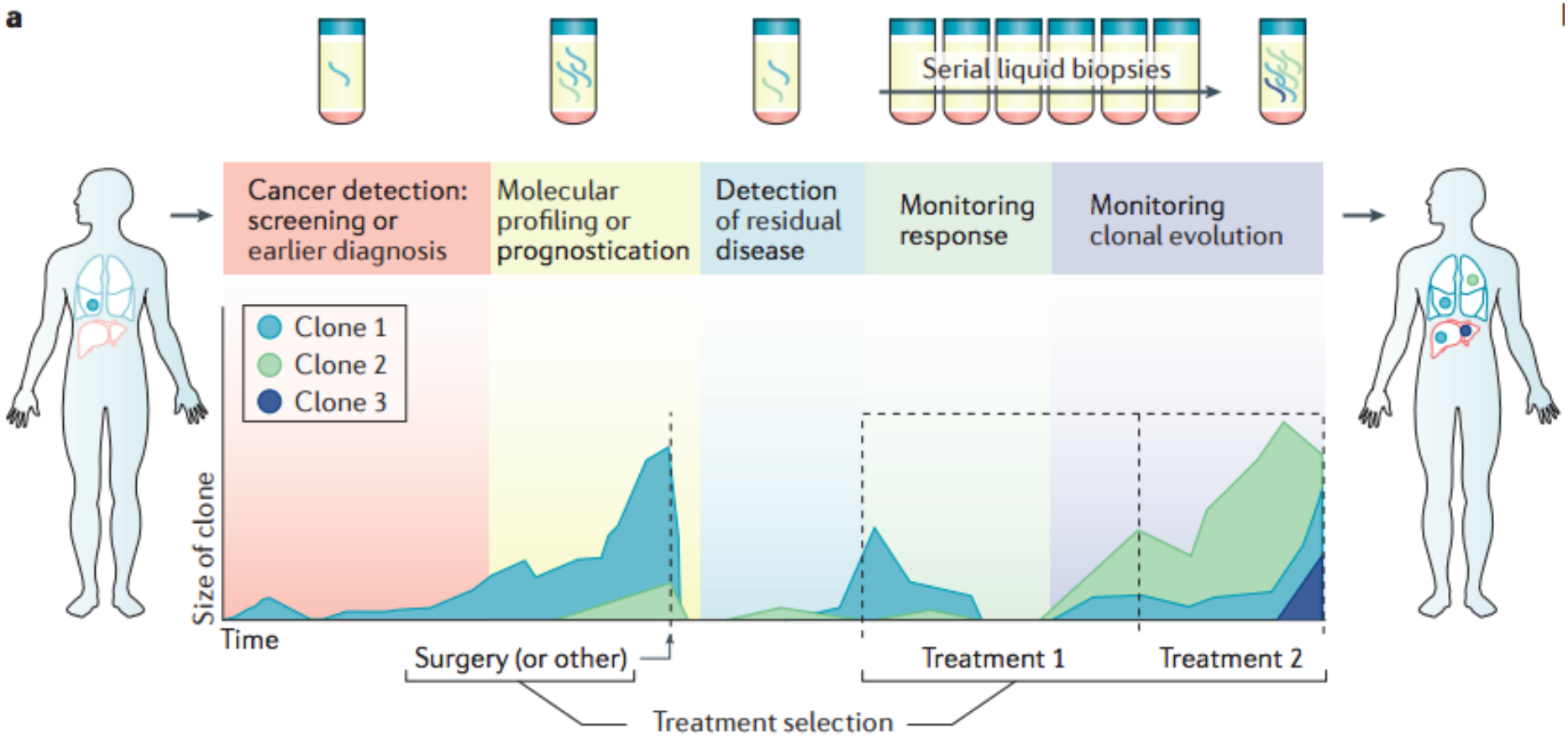
BRCA1 and BRCA2 (breast cancer type 1/2 susceptibility genes)
Normally expressed in the cells of breast and other tissue, where they help repair damaged DNA, or destroy cells if DNA cannot be repaired. They are involved in the repair of chromosomal damage with an important role in the error-free repair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

Tumor Heterogeneity



The evolution of tumour phylogenetics: principles and practice
Schwarz and Schaffer (2017) *Nature Reviews Genetics*. *doi:10.1038/nrg.2016.170*

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

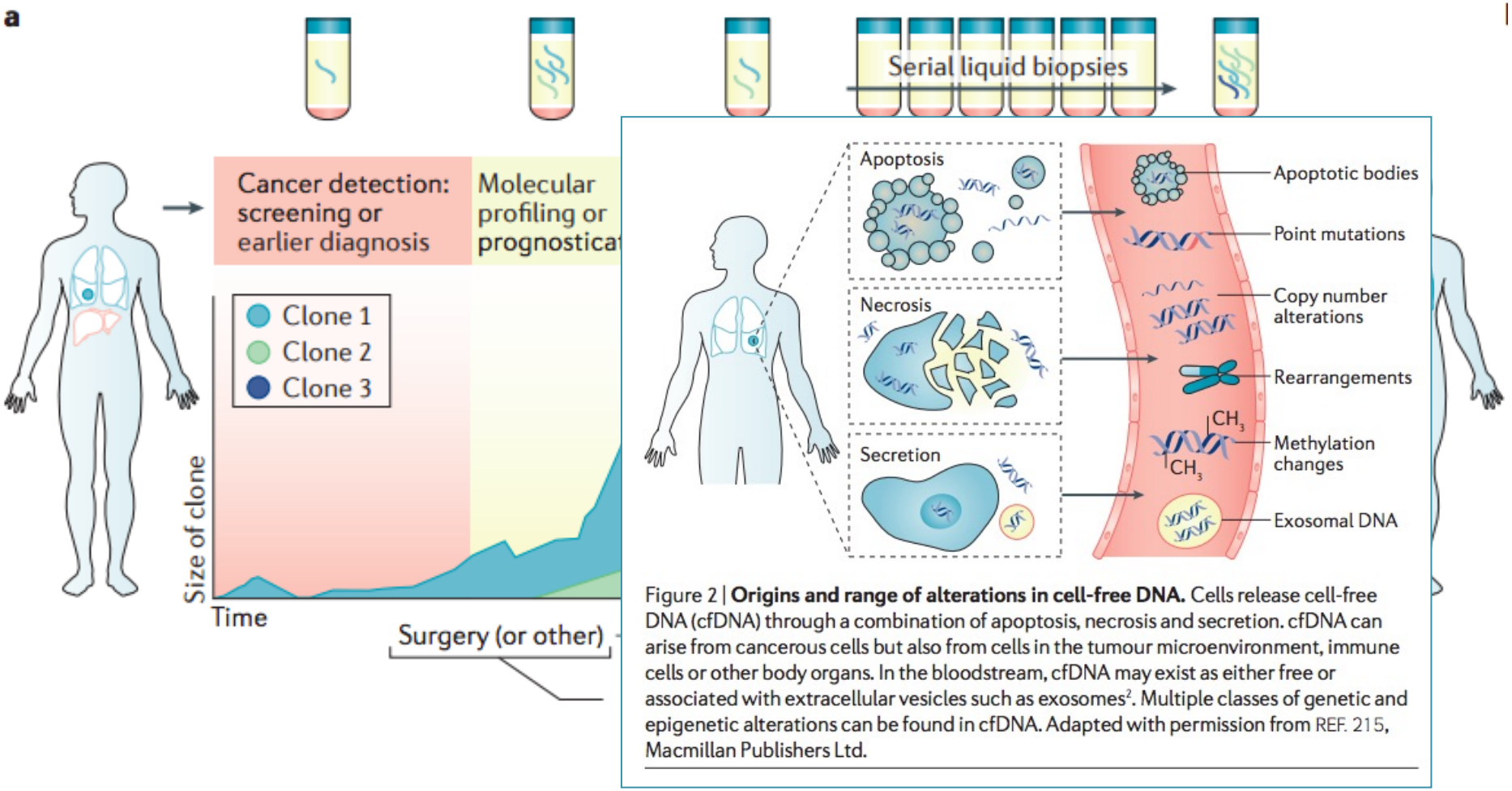


Figure 2 | **Origins and range of alterations in cell-free DNA.** Cells release cell-free DNA (cfDNA) through a combination of apoptosis, necrosis and secretion. cfDNA can arise from cancerous cells but also from cells in the tumour microenvironment, immune cells or other body organs. In the bloodstream, cfDNA may exist as either free or associated with extracellular vesicles such as exosomes[2]. Multiple classes of genetic and epigenetic alterations can be found in cfDNA. Adapted with permission from REF. 215, Macmillan Publishers Ltd.

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 cancer genomes

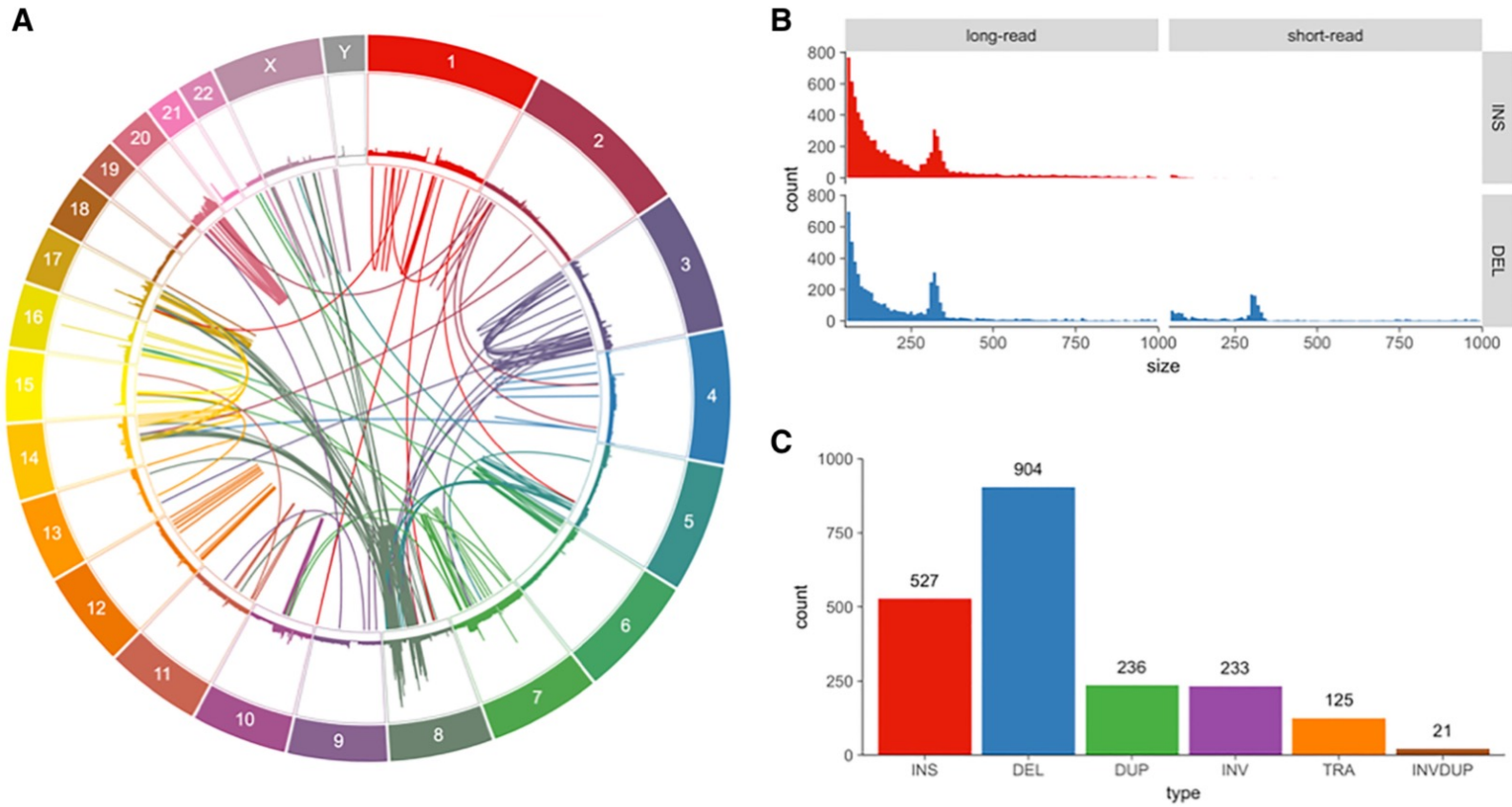


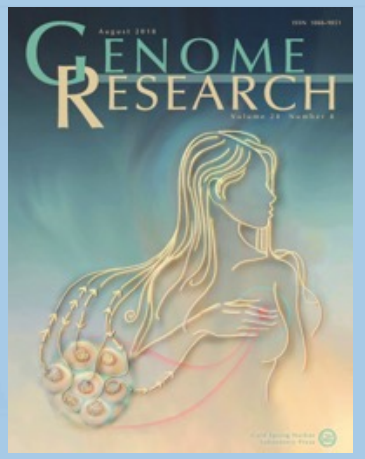
Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (*A*) Circos (Krzywinski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the *outer* track. (*B*) Variant size histogram 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



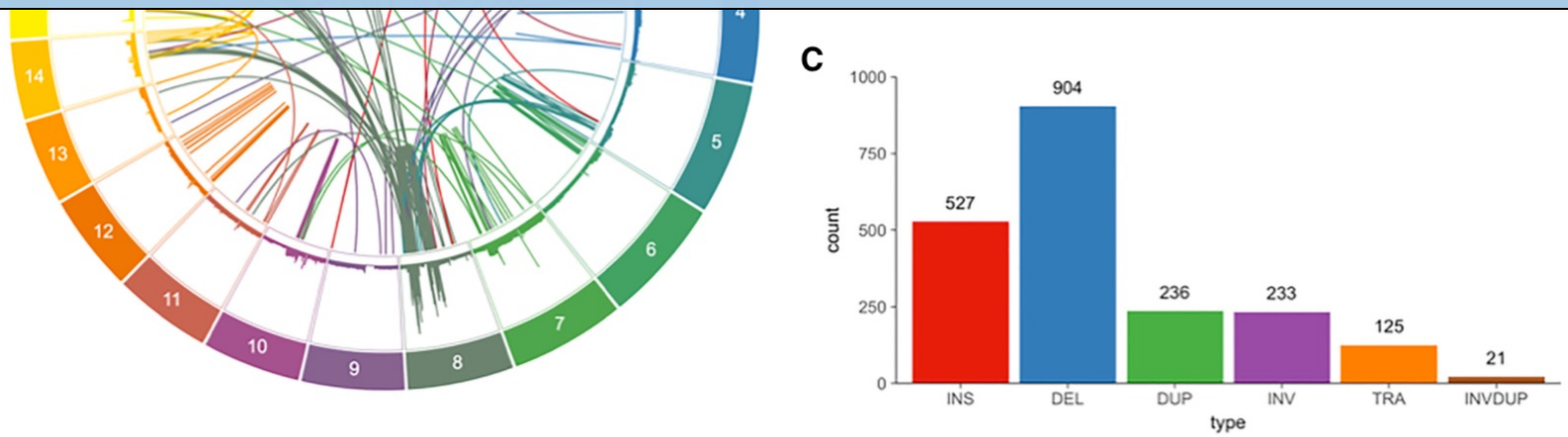
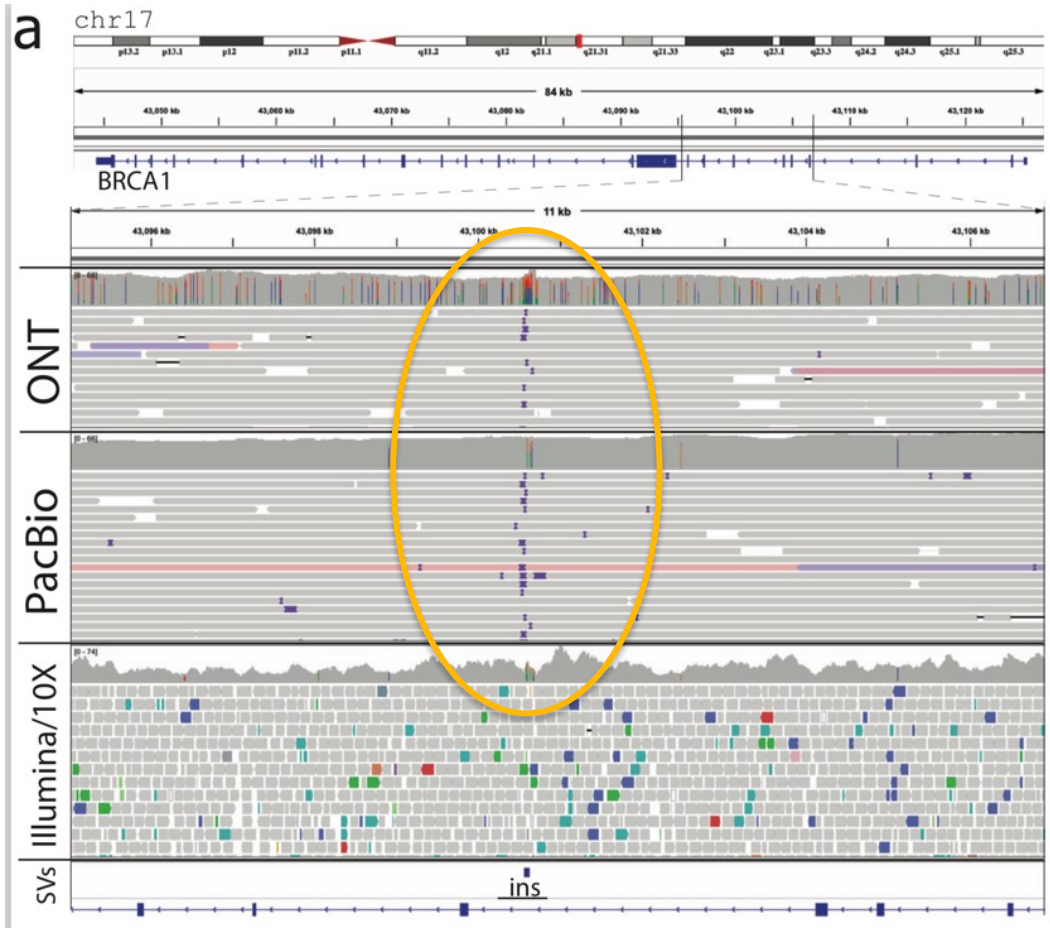


Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (*A*) Circos (Krzywinski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the *outer* track. (*B*) Variant size histogram 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

Hidden Variants in Breast Cancer Genes

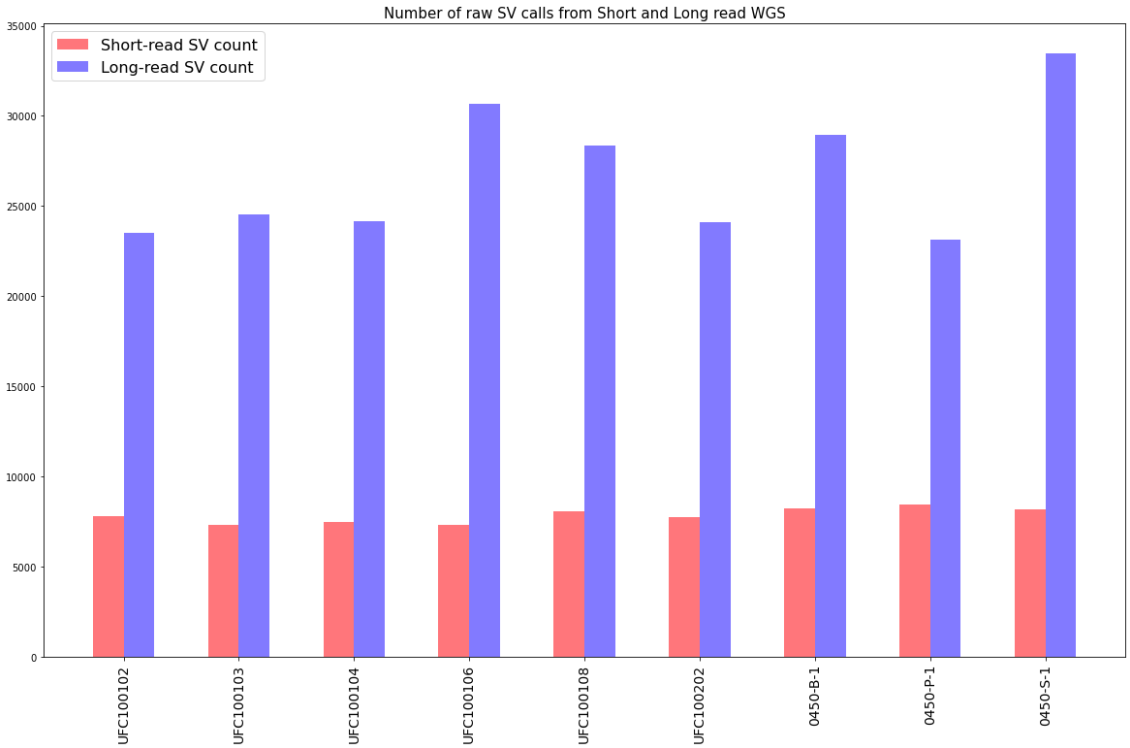


62bp repeat expansion in BRCA1 detected in normal tissue that is undetectable using a panel or short read sequencing

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

Aganezov, et al (2020) Genome Research doi: 10.1101/gr.260497.119

Structural variant analysis in cancer patient genomes using short and long reads



Melanie Kirsche

Van Allen Lab

Jasmine: Population-scale structural variant comparison and analysis
Kirsche, et al (2023) Nature Methods. https://doi.org/10.1038/s41592-022-01753-3

Long read analysis of pancreatic cancer risk

JOHNS HOPKINS MEDICINE
THE SIDNEY KIMMEL COMPREHENSIVE CANCER CENTER

Alison Klein

Winston Timp

COMMONWEALTH FOUNDATION FOR CANCER RESEARCH

JHU DISCOVERY AWARD

LUSTGARTEN FOUNDATION

Thank you!