Article

Pan-cancer analysis of whole genomes

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

Nature 2020

Bicna Song

The pan-cancer analysis of whole genomes

- The pan-cancer analysis of whole genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in 2,658 cancer whole-cancer genomes across 38 tumour types from the PCAWG Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA).

- The project produced large amount data with many types including simple somatic mutations (SNVs, MNVs and small INDELs), large-scale somatic structural variations, copy number alterations, germline variations, RNA expression profiles, gene fusions, and phenotypic annotations.

- PCAWG data have been imported, processed and made available in the following major online resources for download and exploration by the cancer researchers worldwide.

Box 1

Online resources for data access, visualization and analysis

The PCAWG landing page (http://docs.icgc.org/pcawg) provides links to several data resources for interactive online browsing, analysis and download of PCAWG data and results (Supplementary Table 4).

Direct download of PCAWG data

Aligned PCAWG read data in BAM format are also available at the European Genome Phenome Archive (EGA; https://www.ebi.ac.uk/ega/search/site/pcawg under accession number EGAS00001001692). In addition, all open-tier PCAWG genomics data, as well as reference datasets used for analysis, can be downloaded from the ICGC Data Portal at http://docs.icgc.org/pcawg/data/. Controlled-tier genomic data, including SNVs and indels that originated from TCGA projects (in VCF format) and aligned reads (in BAM format) can be downloaded using the Score (https://www.overture.bio/) software package, which has accelerated and secure file transfer, as well as BAM slicing facilities to selectively download defined regions of genomic alignments.

PCAWG computational pipelines

The core alignment, somatic variant-calling, quality-control and variant consensus-generation pipelines used by PCAWG have each been packaged into portable cross-platform images using the Dockstore system[84] and released under an Open Source licence that enables unrestricted use and redistribution. All PCAWG Dockstore images are available to the public at https://dockstore.org/organizations/PCAWG/collections/PCAWG.

ICGC Data Portal

The ICGC Data Portal[85] (https://dcc.icgc.org) serves as the main entry point for accessing PCAWG datasets with a single uniform web interface and a high-performance data-download client. This uniform interface provides users with easy access to the myriad of PCAWG sequencing data and variant calls that reside in many repositories and compute clouds worldwide. Streaming technology[86] provides users with high-level visualizations in real time of BAM and VCF files stored remotely on the Cancer Genome Collaboratory.

UCSC Xena

UCSC Xena[87] (https://pcawg.xenahubs.net) visualizes all PCAWG primary results, including copy-number, gene-expression, gene-fusion and promoter-usage alterations, simple somatic mutations, large somatic structural variations, mutational signatures and phenotypic data. These open-access data are available through a public Xena hub, and consensus simple somatic mutations can be loaded to the local computer of a user via a private Xena hub. Kaplan–Meier plots, histograms, box plots, scatter plots and transcript-specific views offer additional visualization options and statistical analyses.

The Expression Atlas

The Expression Atlas (https://www.ebi.ac.uk/gxa/home) contains RNA-sequencing and expression microarray data for querying gene expression across tissues, cell types, developmental stages and/or experimental conditions[88]. Two different views of the data are provided: summarized expression levels for each tumour type and gene expression at the level of individual samples, including reference-gene expression datasets for matching normal tissues.

PCAWG Scout

PCAWG Scout (http://pcawgscout.bsc.es/) provides a framework for -omics workflow and website templating to generate on-demand, in-depth analyses of the PCAWG data that are openly available to the whole research community. Views of protected data are available that still safeguard sensitive data. Through the PCAWG Scout web interface, users can access an array of reports and visualizations that leverage on-demand bioinformatic computing infrastructure to produce results in real time, allowing users to discover trends as well as form and test hypotheses.

Chromothripsis Explorer

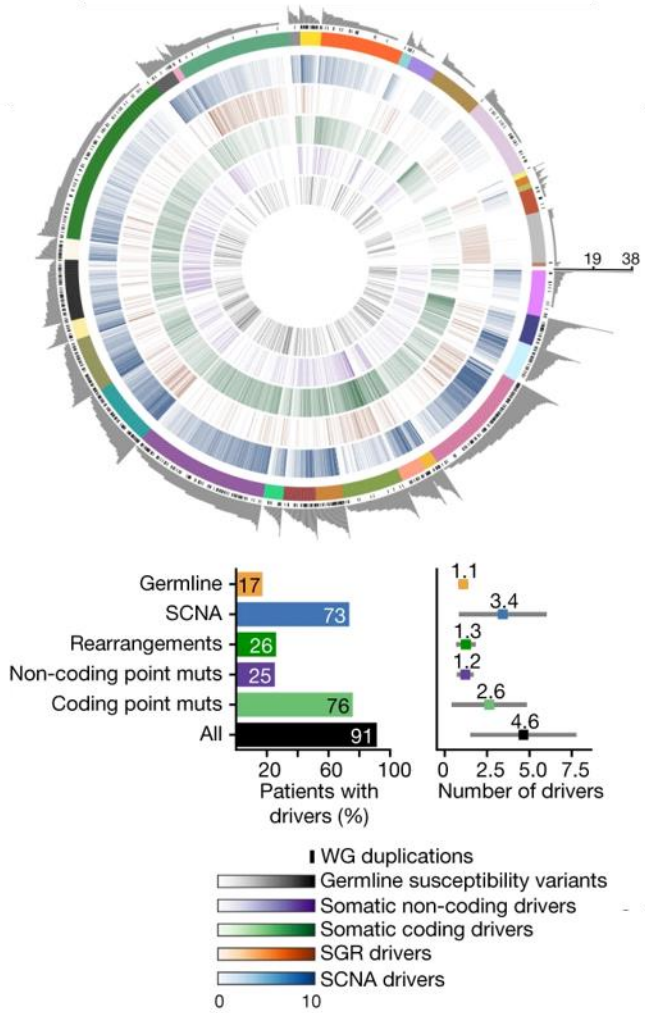
Chromothripsis Explorer (http://compbio.med.harvard.edu/chromothripsis/) is a portal that allows structural variation in the PCAWG dataset to be explored on an individual patient basis through the use of circos plots. Patterns of chromothripsis can also be explored in aggregated formats.

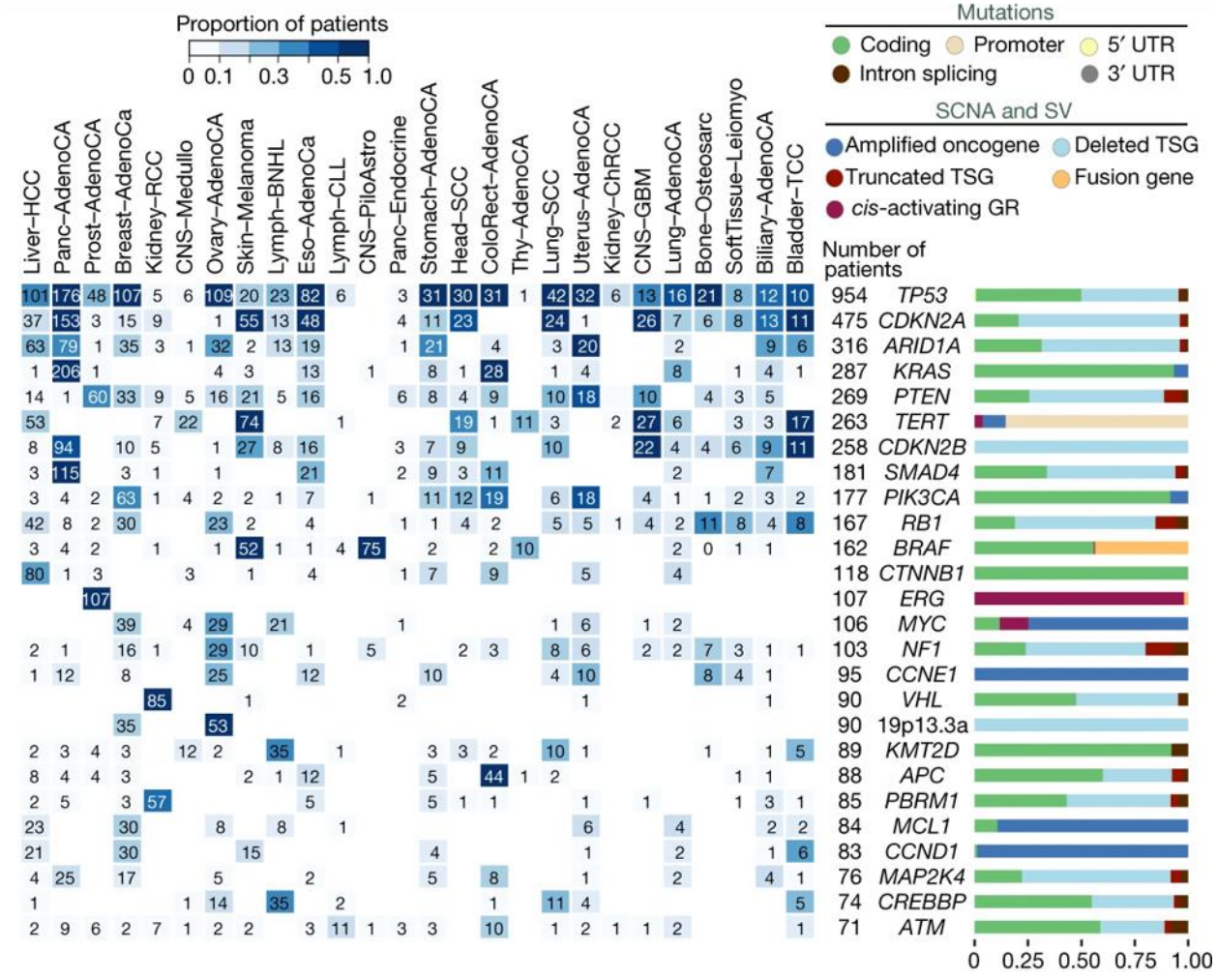
Analysis of PCAWG data

A comprehensive suite of companion papers that describe the analyses and discoveries across these thematic areas is copublished with this paper.

Scientific area	Key findings	Citation
Driver mutations		
Discovery of non-coding drivers	• Estimated ~10-fold more coding than non-coding driver point mutations. • Variation in point mutation density in non-coding regions influenced more by mutational processes than selection.	4
Drivers by pathways and networks	• Both coding and non-coding alterations contribute to cancer pathways. • Some pathways, such as RNA splicing, are primarily driven by non-coding mutations.	16
Evolution and heterogeneity		
Timing of cancer evolution	• Each tumour type has a distinct pattern of early and late-occurring driver events. • Earliest somatic mutations may occur decades prior to diagnosis, providing opportunities for early diagnosis. • Intra-tumour heterogeneity is widespread and tumour subclones contain drivers that are under positive selection.	7
Structural variants		
Patterns of structural variation	• Replication-based mechanisms of genome rearrangement frequent in many cancers, often causing driver structural variants. • 16 signatures of SV, including break-and-ligate patterns and copy-and-insert patterns, varying by size range, replication timing, tumour type and patient.	6
Functional consequence of structural variation	• 52 regions with recurrent structural breakpoints and 90 recurrently fused pairs of loci show evidence of positive selection. • Oncogenic fusions are shaped by juxtaposition of proto-oncogenes with tissue-specific regulatory elements.	4
Patterns of retrotransposition	• Many flavours of somatic retrotransposition in many cancers: LINE element mobilisation; transductions, pseudogenes, Alu elements. • Retrotranspositions can induce genomic instability, including large deletions and breakage-fusion-bridge cycles amplifying cancer genes.	10
Chromothripsis	• Chromothripsis pervasive across cancers, with frequency >50% in several tumour types. • Replicative processes and templated insertions contribute to rearrangement.	18
Mutational signatures		
Signatures of point mutations	• >70 distinct mutational signatures, encompassing SNVs, doublet subs and indels. • Multiple signatures from unknown processes of DNA damage, repair and replication.	5
Mutation distribution across genome	• Uneven distribution of somatic mutations and structural variants across the genome explained by epigenetic state of tissue, cell of origin and topological associated domains. • Can be used to identify a tumour's type and presumed tissue/cell of origin.	11,12,15
Transcriptional consequences of somatic mutation		
RNA effects of somatic mutation	• Genomic basis for RNA alterations across ~1200 tumours, including quantitative trait loci, allele specific expression and alternative splicing. • Link between mutational signatures and expression; classification of gene fusions; identification of genes recurrently altered at RNA level.	8,9
Others		
Tumour subtypes from genome sequencing	• Genomic distribution of somatic mutations, mutational signatures and driver mutations accurately distinguish major tumour types of primaries and metastases.	12
Mitochondrial DNA mutations	• Somatic mitochondrial truncating mutations frequent in certain cancer types, associated with activation of critical signaling pathways.	14
Telomere biology and sequences	• Activating *TERT* promoter mutations are the single most frequent non-coding driver. • In *ATRX*/*DAXX*-mutant tumours, aberrant telomere variant repeat distribution is common.	4,13

Panorama of driver mutations in cancer





The compendium of mutational driver genomic elements is available at Synapse IDs (syn11679360).

PCAWG tumours with no apparent drivers

- Although more than 90% of PCAWG cases had identified drivers, we found none in 181 tumours. Reasons for missing drivers have not yet been systematically evaluated in a pan-cancer cohort, and could arise from either technical or biological causes.

- **Technical causes**: poor-quality samples or failures in the bioinformatic algorithms

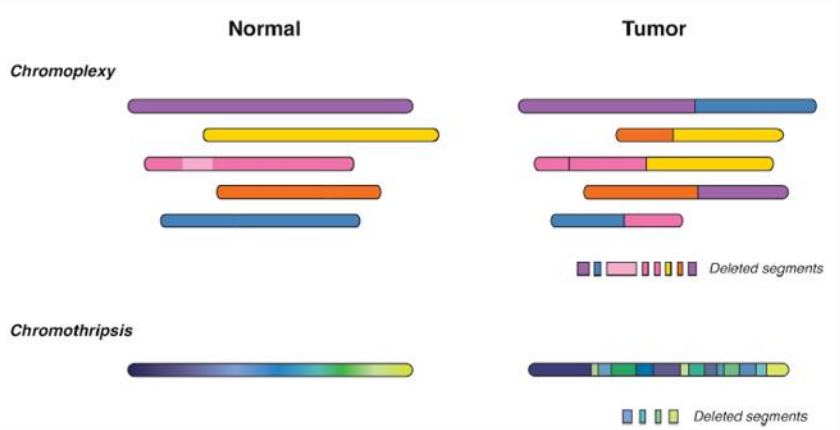
- **Biological causes**: tumours may be driven by mutations that are not yet discovered.

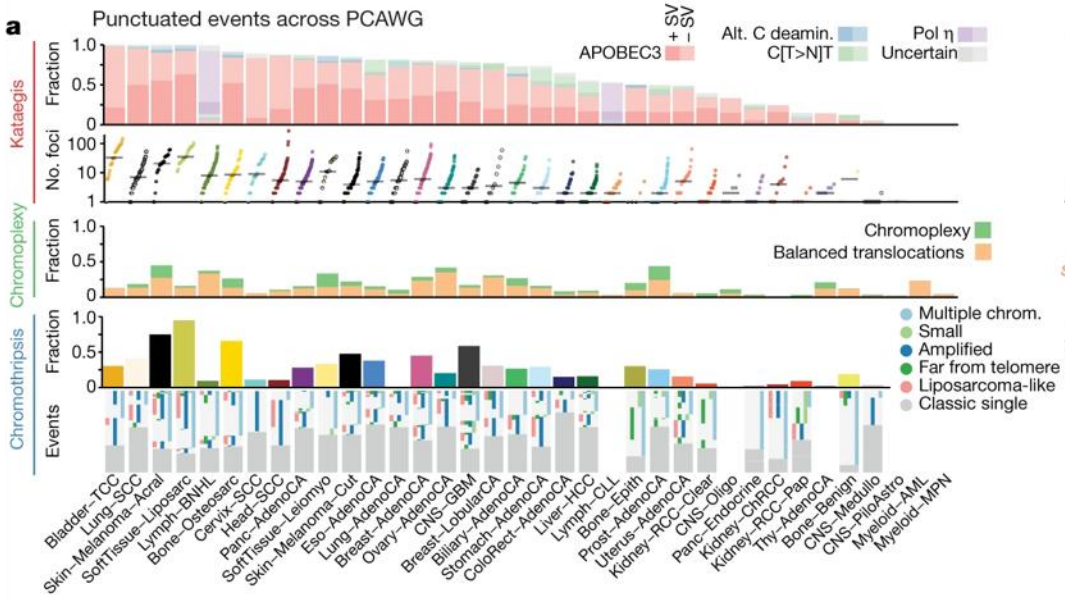
Patterns of clustered mutations and SVs

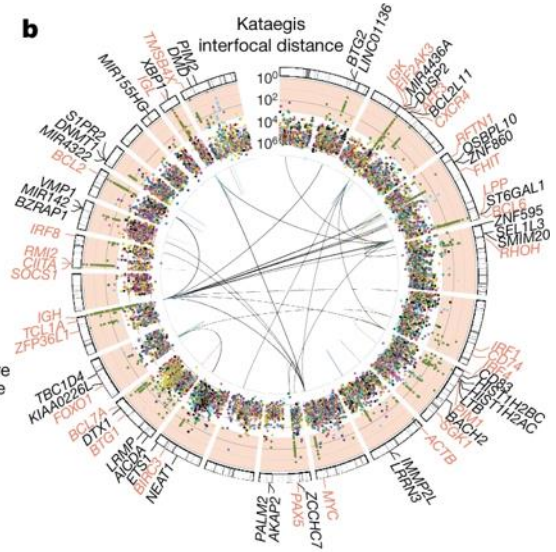
Kataegis is a focal hypermutations process that leads to locally clustered nucleotide substitutions.

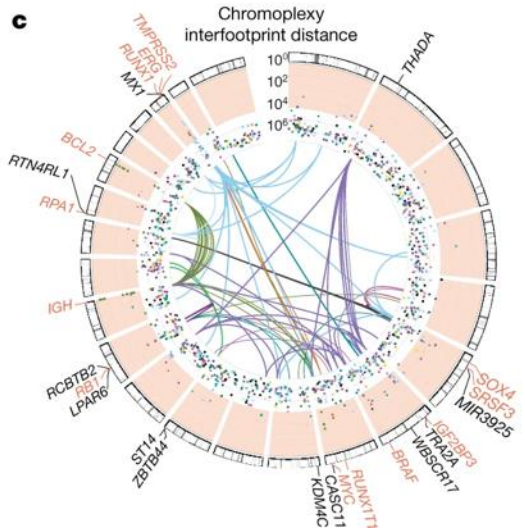
Chromoplexy refers to a class of complex DNA rearrangement observed in the genomes of cancer cells.

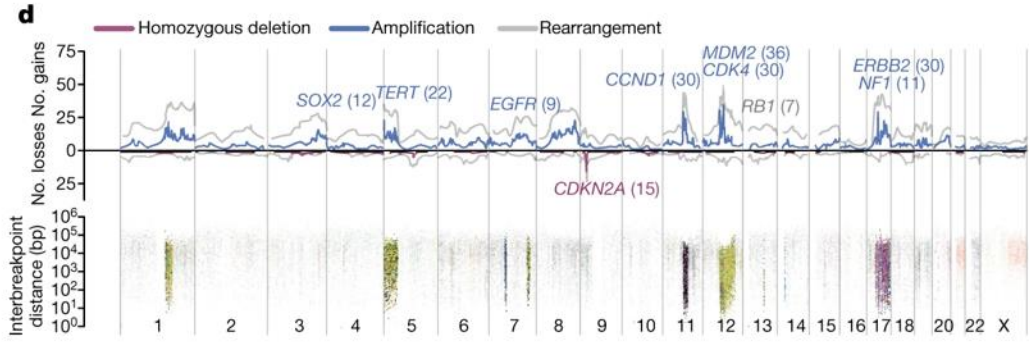
Chromothripsis is a mutational process by which up to thousands of clustered chromosomal rearrangements occur in a single event in localised and confined genomic regions in one or a few chromosomes.











Chromothripsis-induced gains or losses (grey) and amplifications (blue) or deletions (red)

Summary

- More than 90% of tumours had at least one identified driver mutation.
- Tumours with no apparent drivers could arise from either technical or biological causes.
- Three mutational processes, such as kataegis, chromoplexy, and chromothripsis, have been characterized in the PCAWG genomes.

Future directions

- The project produced large amount data with many types and integrative analysis of these data has discovered the biological insights of many aspects.

- Next step is that **research findings need to be translated into sustainable, meaningful clinical treatment**.

- Build comprehensive knowledge banks of clinical outcome and treatment data from patients with a wide variety of cancers.

Journal Club

nature

Article | Open Access | Published: 05 February 2020

Analyses of non-coding somatic drivers in 2,658 cancer whole genomes

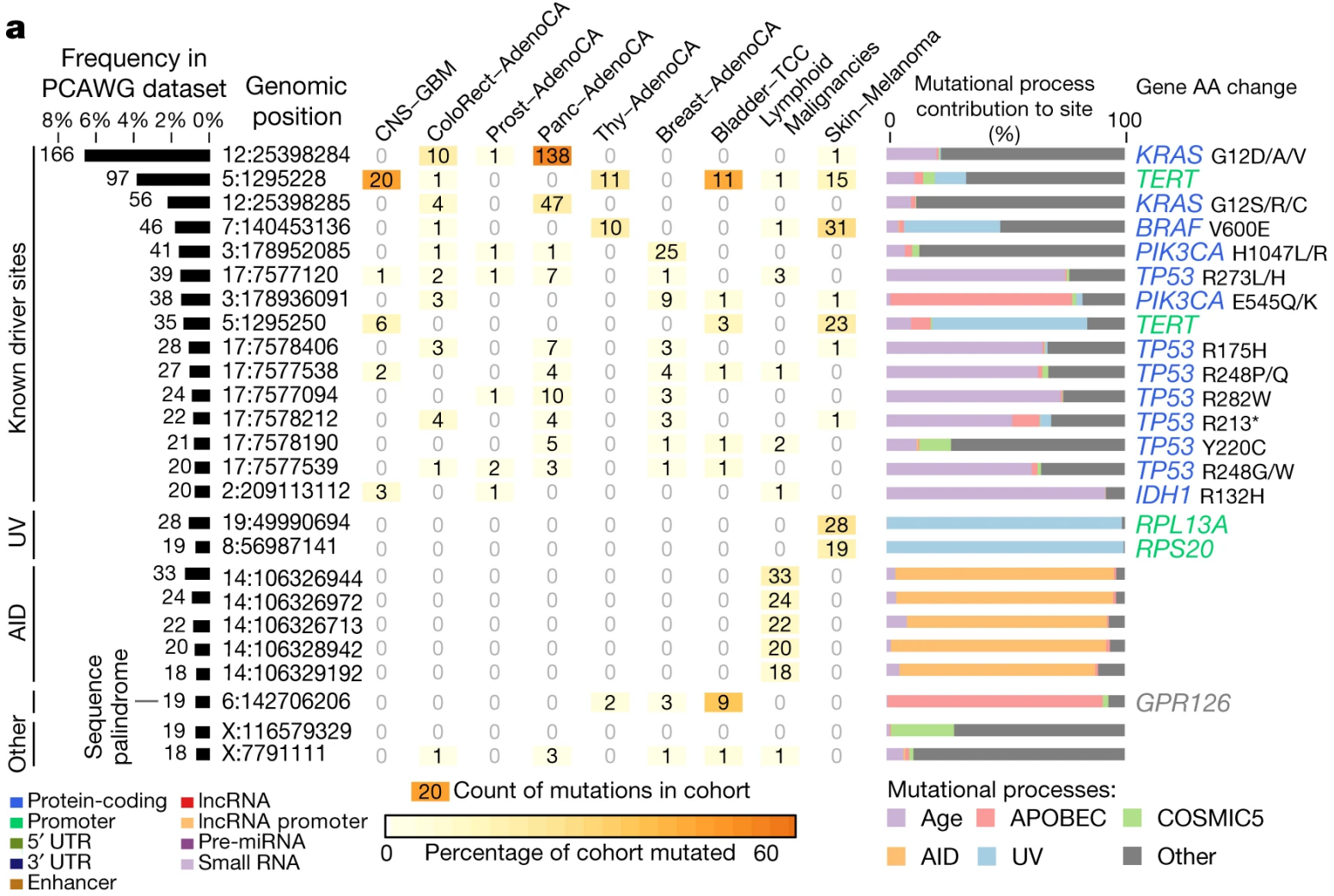
Esther Rheinbay, Morten Muhlig Nielsen, [...] PCAWG Consortium

Nature **578**, 102–111(2020) | Cite this article

40k Accesses | **3** Citations | **211** Altmetric | Metrics

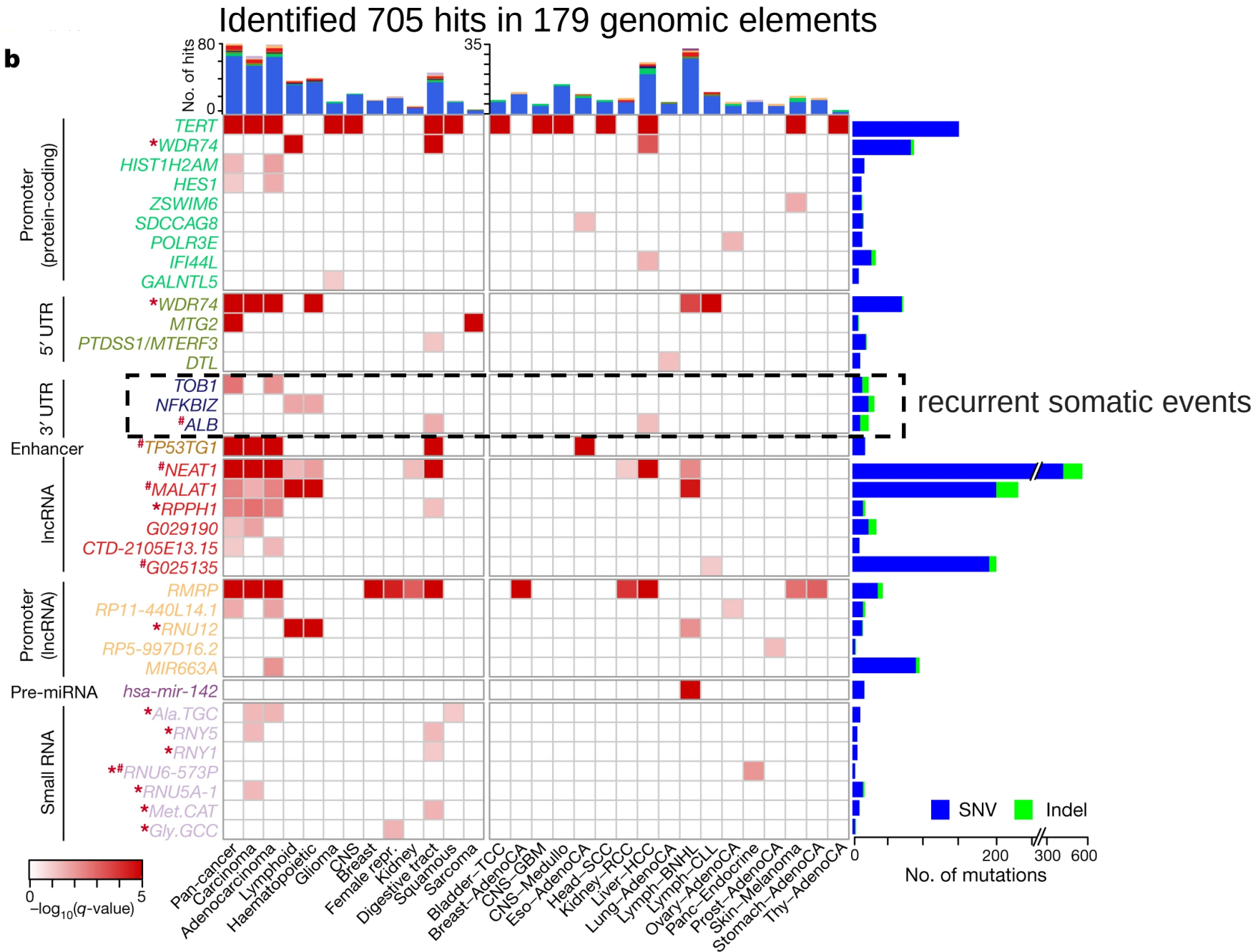
Xiaolong Cheng

Hotspot mutations across cancer types

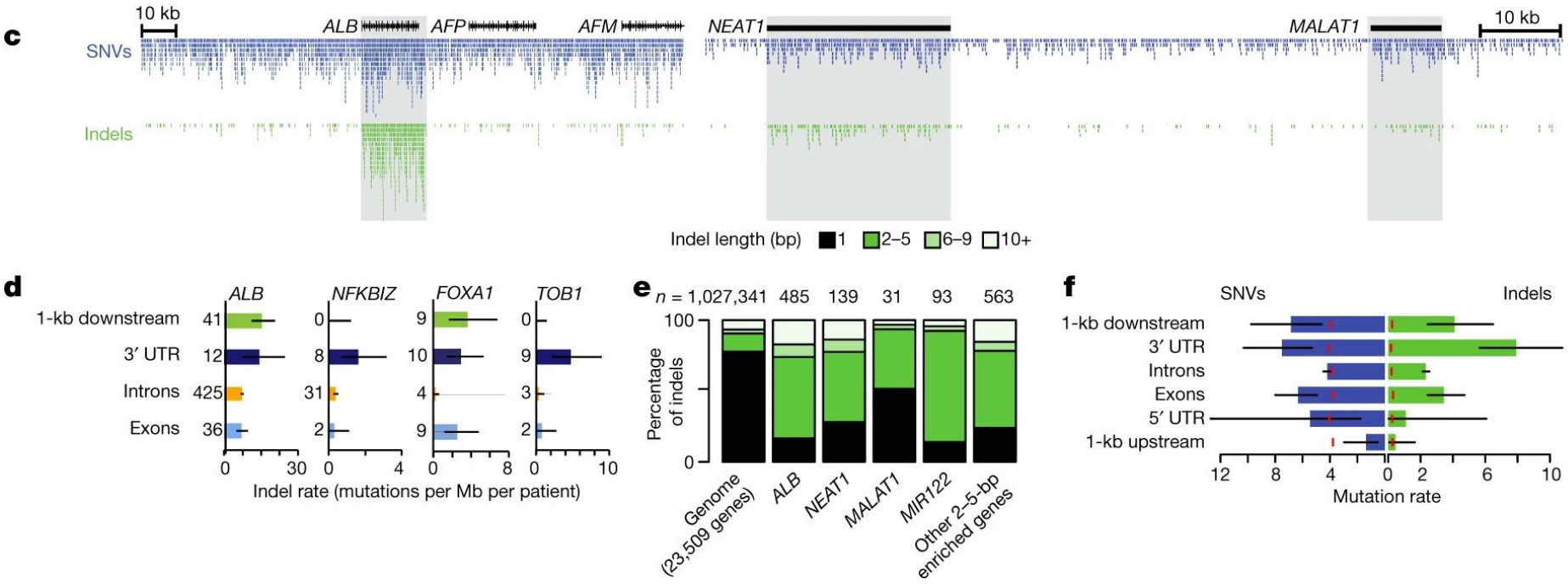


- Besides TERT promoter events, non-coding single-site hotspot drivers are infrequent or fall in regions with low sensitivity to detect mutations.

Candidate coding and non-coding drivers

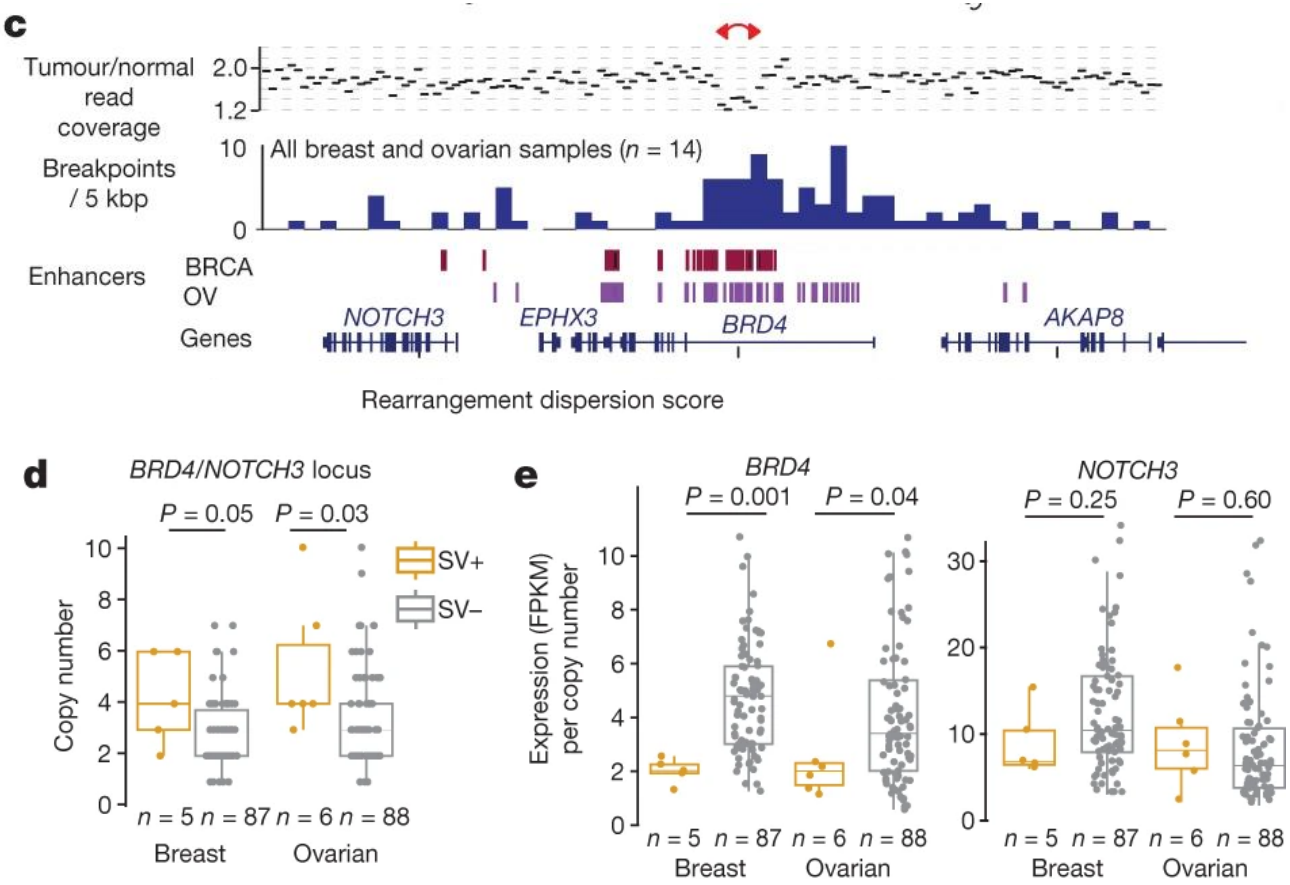


Transcription-associated indel signature



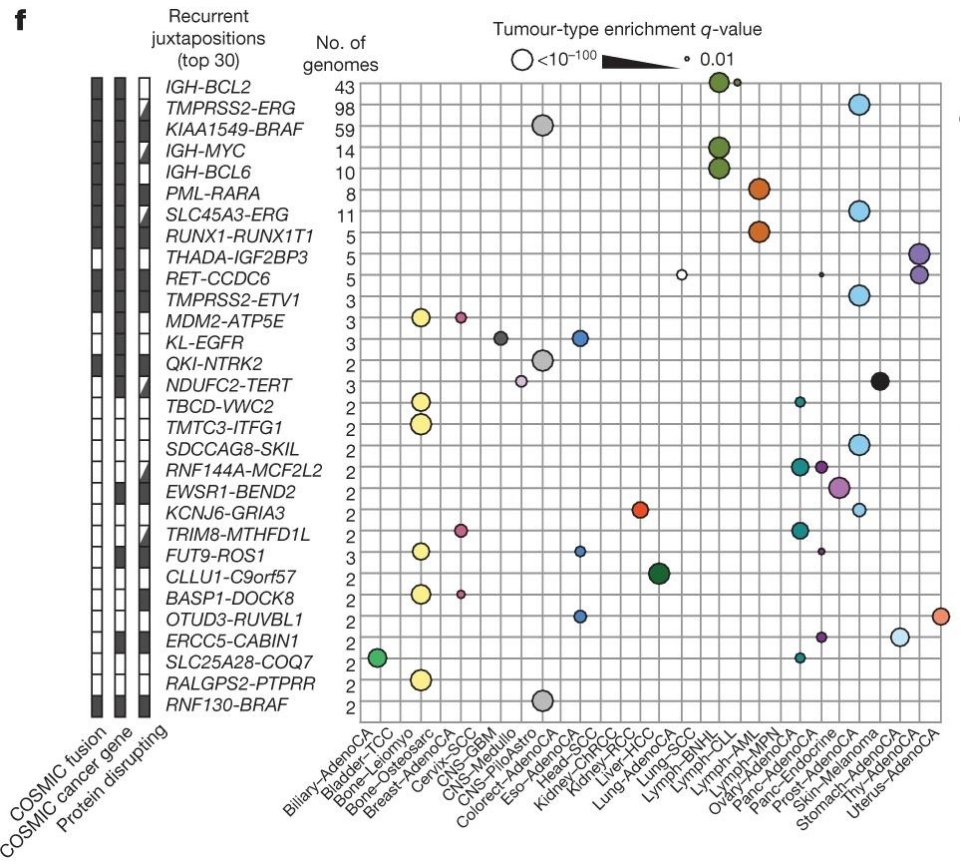
- The ALB indel rate is notably high throughout the UTRs, introns and exons, and even downstream (a pattern inconsistent with selection)
- FOXA1 has high indel rates throughout its locus, whereas the indels in NFKBIZ and TOB1 are in their 3′ UTRs, suggesting that these are driver events
- The indels in MALAT1, NEAT1, ALB and MIR122 are not driver events and are the result of a transcription-associated mutational process.

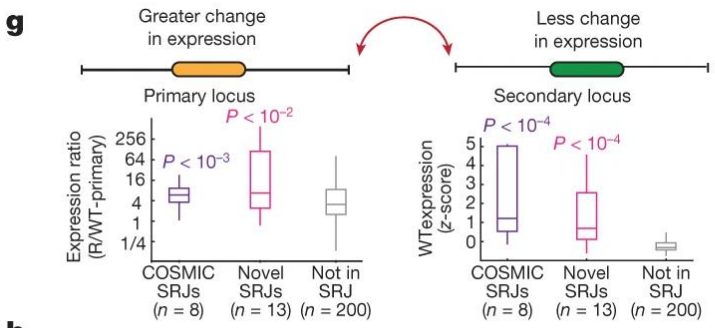
Novel structural-variant driver candidates

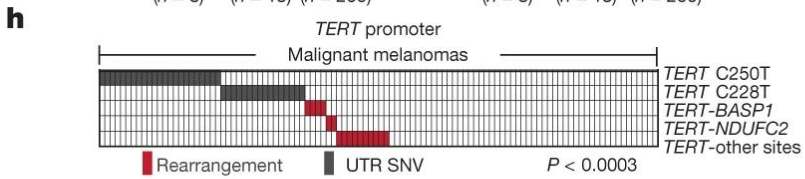


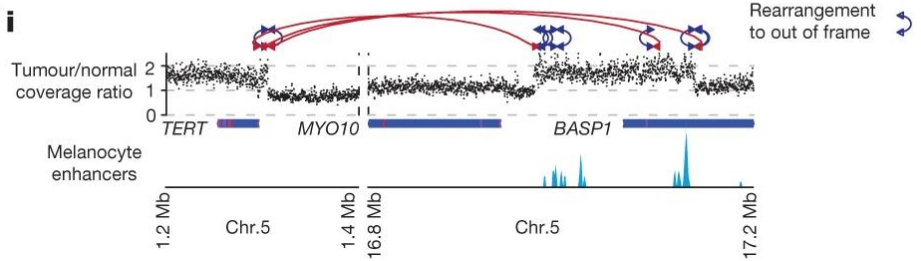
- These deletions were highly enriched in cancers that amplified a segment that includes BRD4 and NOTCH3
- However, the microdeletions are associated with a lower expression of BRD4 in breast and ovarian tumours, but not of the neighbouring gene NOTCH3
- To our knowledge, this is the first evidence of a recurrent microdeletion limiting expression of an amplified gene.

Recurrent fusions target gene regulation

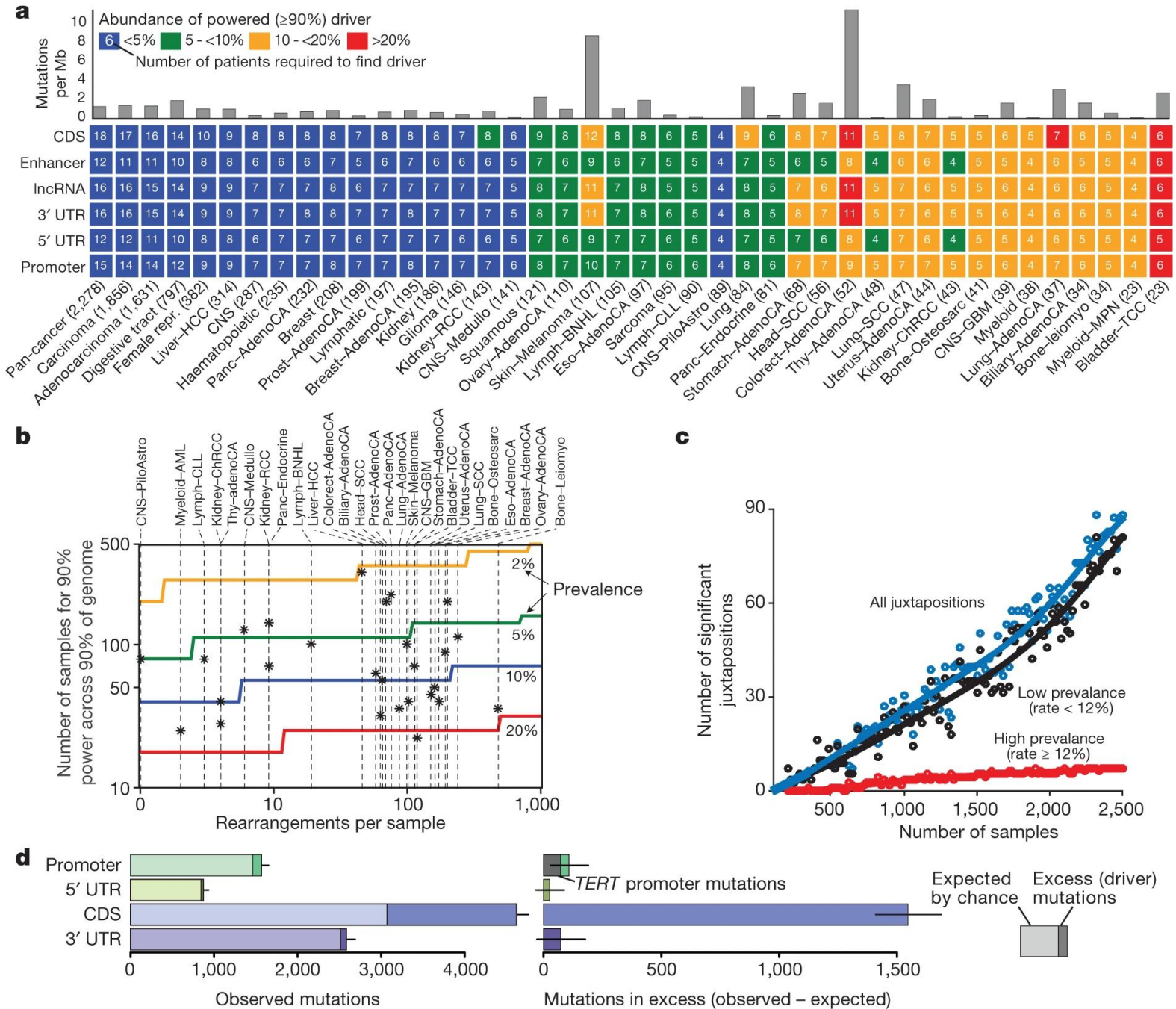








Paucity of non-coding drivers in cancer



Summary

- Confirmed previously reported drivers, raise doubts about others and identify novel candidates.
- They suggest that larger datasets and technological advances will continue to identify new non-coding drivers, albeit at considerably lower frequencies than protein-coding drivers.
- The work will provide a solid foundation for the incipient era of driver discovery from ever-larger numbers of cancer whole genomes.

Pathway and network analysis of more than 2500 whole cancer genomes

Matthew A. Reyna[1,2], David Haan[3], Marta Paczkowska[4], Lieven P.C. Verbeke[5,6], Miguel Vazquez[7,8], Abdullah Kahraman[9,10], Sergio Pulido-Tamayo[5,6], Jonathan Barenboim[4], Lina Wadi[4], Priyanka Dhingra[11], Raunak Shrestha[12], Gad Getz[13,14,15,16], Michael S. Lawrence[13,14], Jakob Skou Pedersen[17,18], Mark A. Rubin[11], David A. Wheeler[19], Søren Brunak[20,21], Jose M.G. Izarzugaza[20,21], Ekta Khurana[11], Kathleen Marchal[5,6], Christian von Mering[9], S. Cenk Sahinalp[12,22], Alfonso Valencia[7,23], PCAWG Drivers and Functional Interpretation Working Group, Jüri Reimand[4,24*], Joshua M. Stuart[3*], Benjamin J. Raphael[1*] & PCAWG Consortium

Nat Com, 2020

Mutations in cancer genome

- **Diver mutation**

A mutation is causally implicated in oncogenesis. It has growth advantage on the cancer cell and has been positively selected in the microenvironment.

- **Passenger mutation**

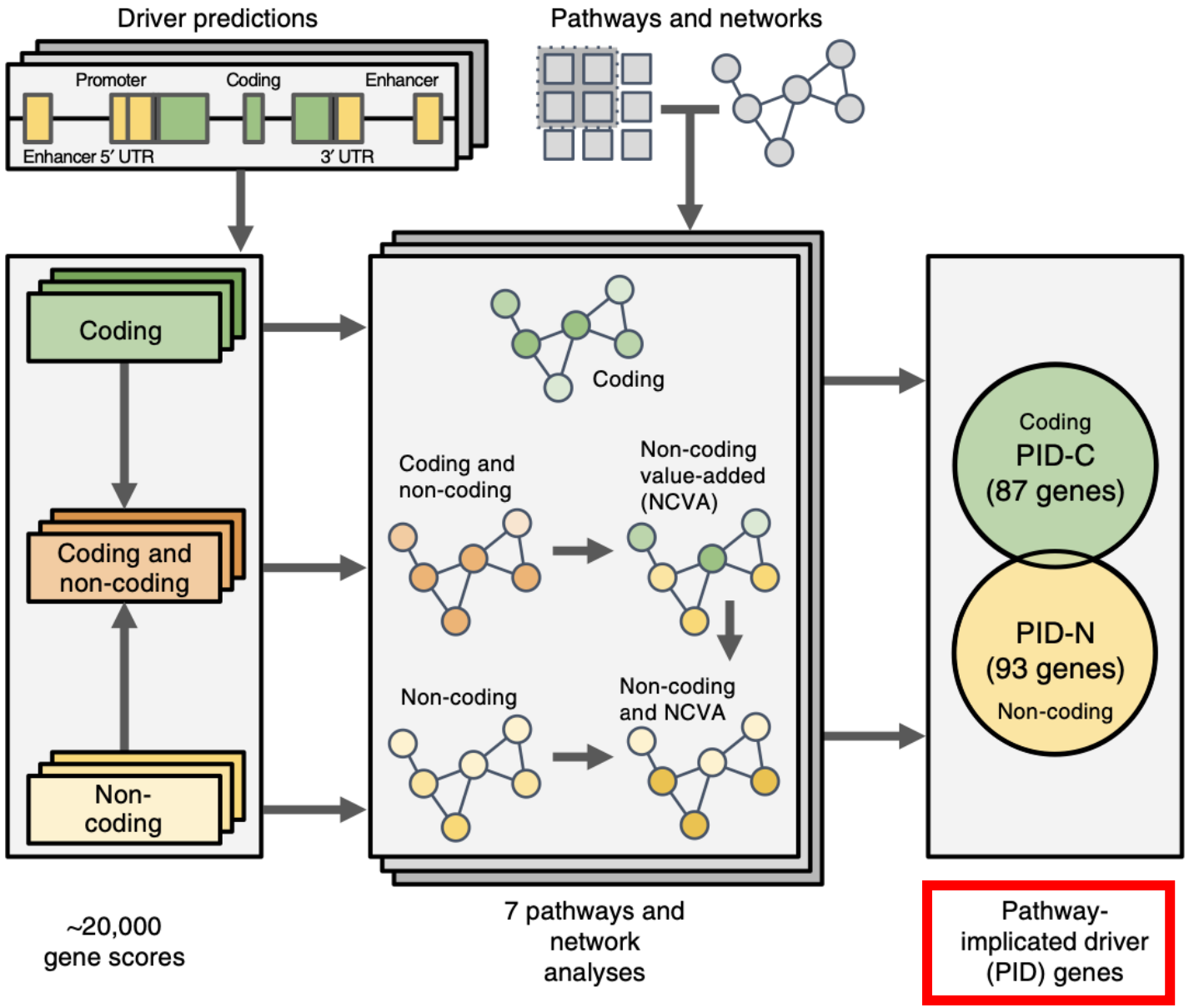
A mutation has not been selected, has not conferred growth advantage and has therefore not contributed to cancer development.

Goal

Cancer driver mutations in protein-coding genes has greatly expanded in the past decade. However, non-coding cancer driver mutations are less well-characterized

The author perform multi-faceted pathway and network analyses of non-coding mutations across 2583 whole cancer genomes from 27 tumor types

Overview of the pathway and network analysis approach

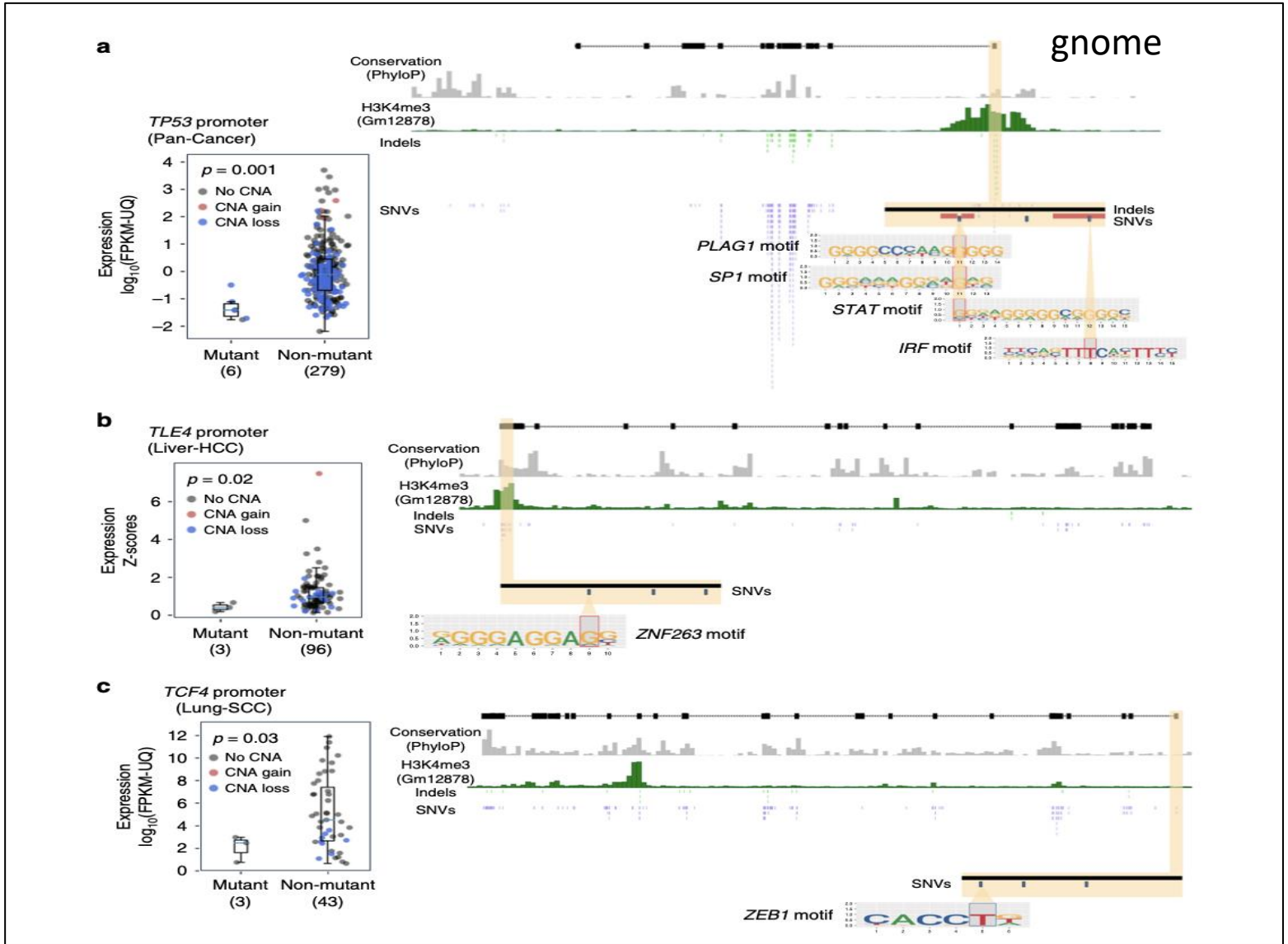


5 genes in common:
CTNNB1, DDX3X, SF3B1, TGFBR2, and TP53

Suggests that coding and non-coding driver mutations occur in largely distinct sets of cancer genes

Gene expression changes are correlated with mutations in PID-N genes

Validated the Impact of non-coding mutations on gene expression

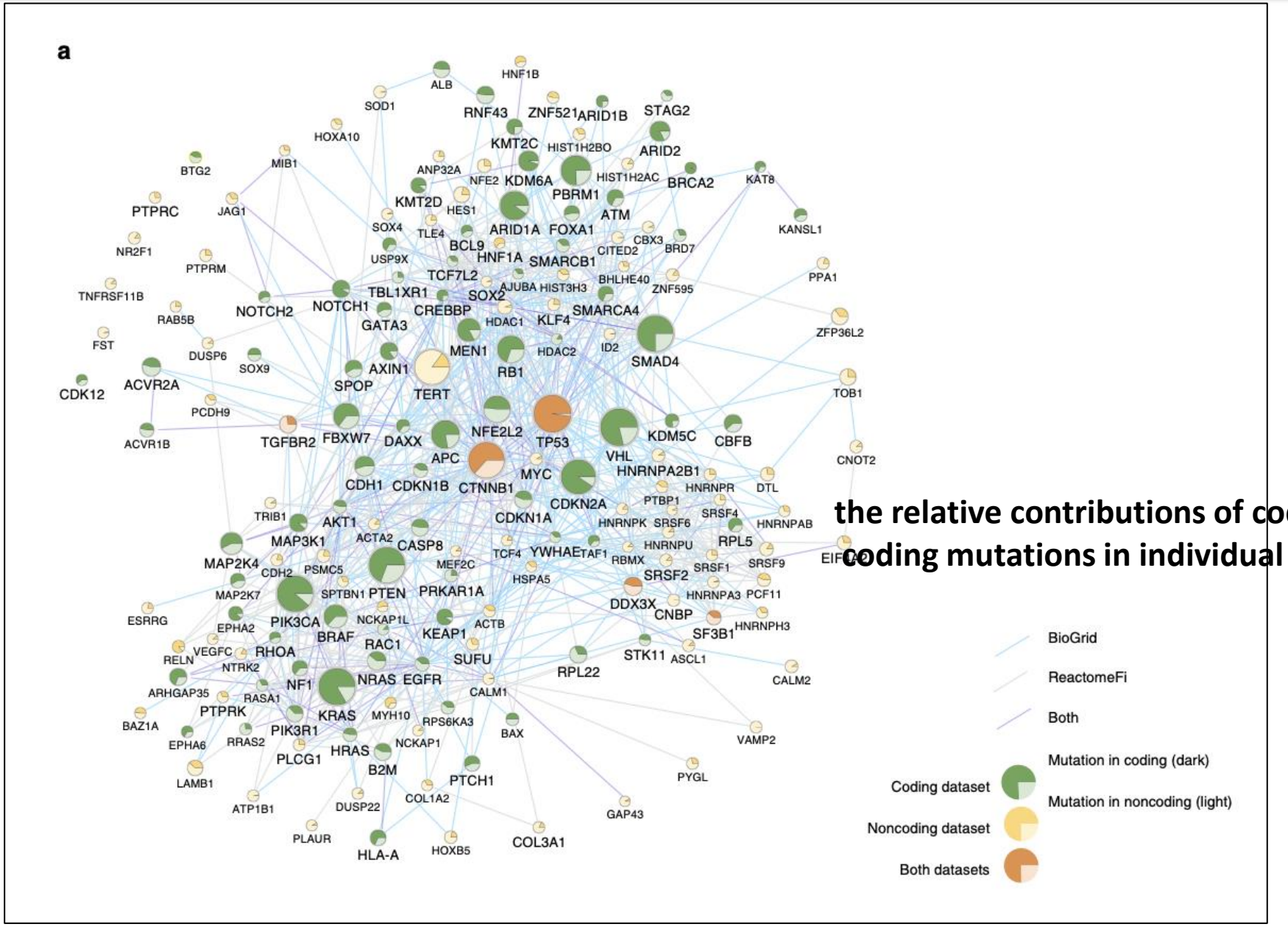


TP53 is well-known role as a tumor suppressor gene

TLE4 is a transcriptional co-repressor that binds to several transcription factors. Function as tumor suppressor in AML

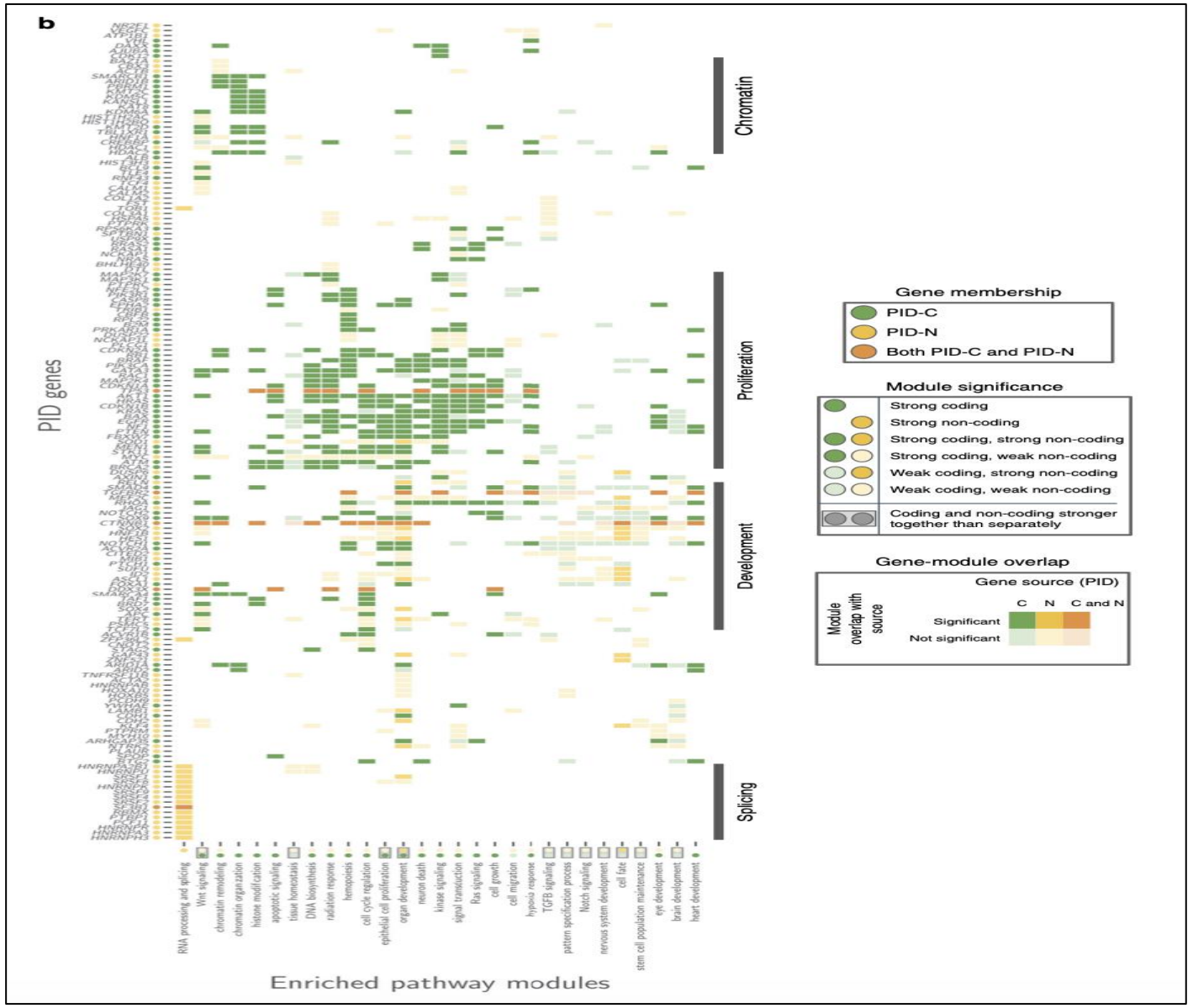
TCF4 is part of the TCF4/β-catenin complex and encodes a transcription factor that is downstream of the Wnt signaling pathway

Pathway and network modules containing PID-C and PID-N genes



the relative contributions of coding and non-coding mutations in individual subnetworks

Pathway and network modules containing PID-C and PID-N genes



- **4 biological processes:**

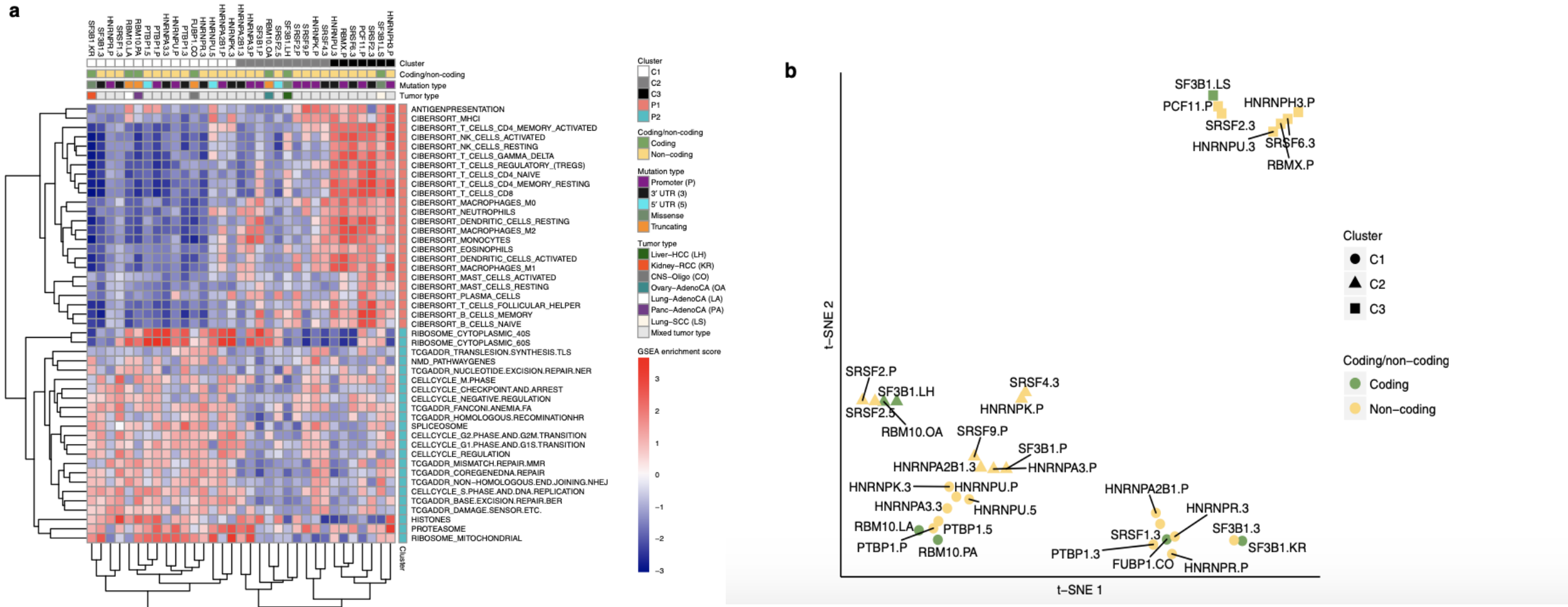
chromatin organization, cell proliferation, development, and RNA splicing

- **The Wnt signaling pathway was targeted by coding mutations and non-coding mutations**

- **RNA splicing pathways were affected primarily by non-coding mutations**

Evaluate the association between RNA-splicing related PID-N genes and expression of other genes

RNA splicing factors are targeted primarily by non-coding mutations and alter expression of similar pathways



samples containing non-coding mutations in well-known RNA splicing factors exhibit similar gene expression signatures as samples with coding mutations in these genes

Summary

- Coding and non-coding driver mutations largely target different genes and make varying contributions to pathways and networks perturbed in cancer.

- RNA splicing is primarily altered by non-coding mutations in this cohort, and samples containing non-coding mutations in RNA splicing factors exhibit similar gene expression signatures as samples with coding mutations in these genes.

- Investigation of the coding and non-coding mutations that perturb these pathways and networks will enable more accurate patient-stratification strategies.

nature genetics

ARTICLES

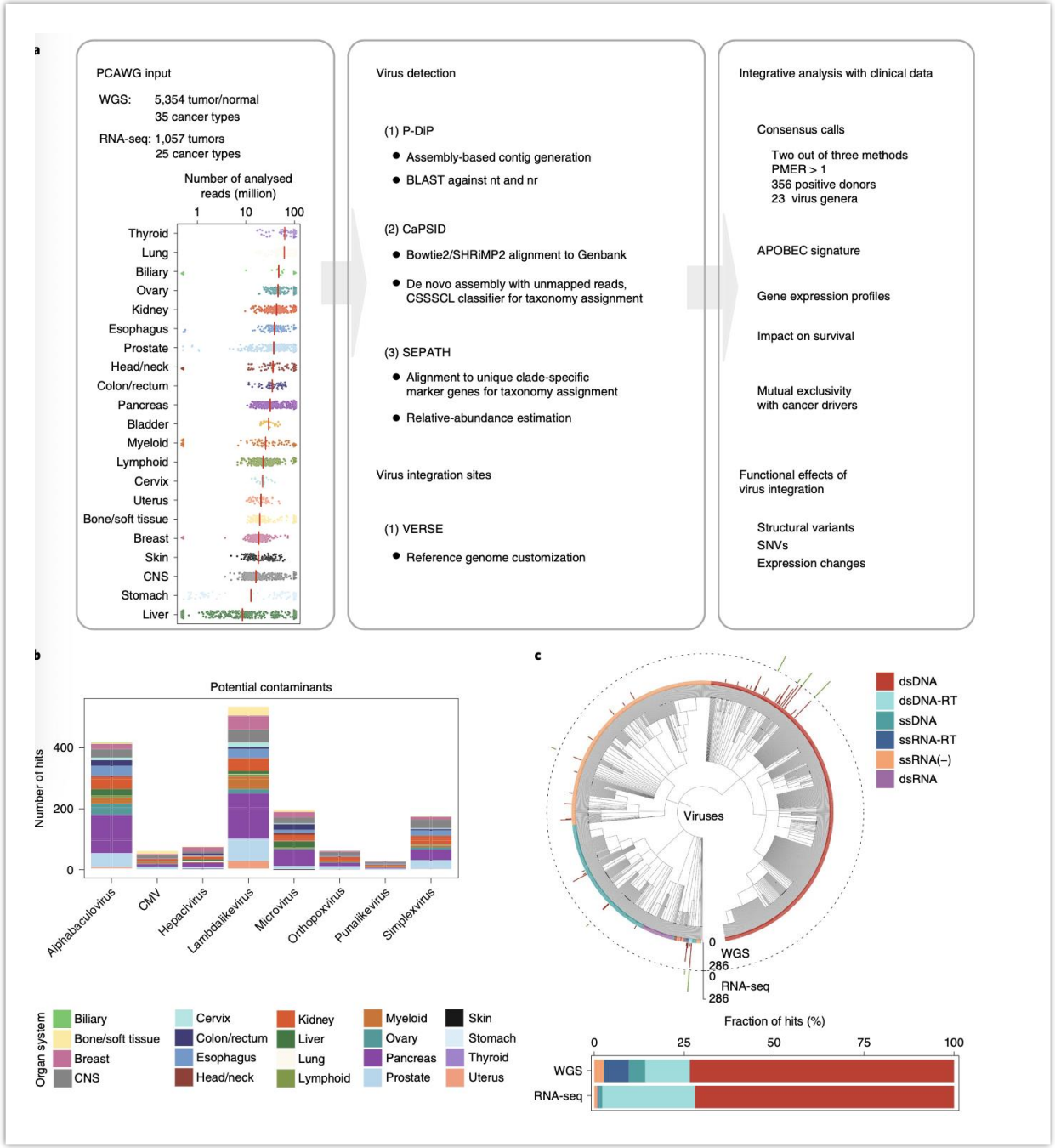
https://doi.org/10.1038/s41588-019-0558-9

OPEN

The landscape of viral associations in human cancers

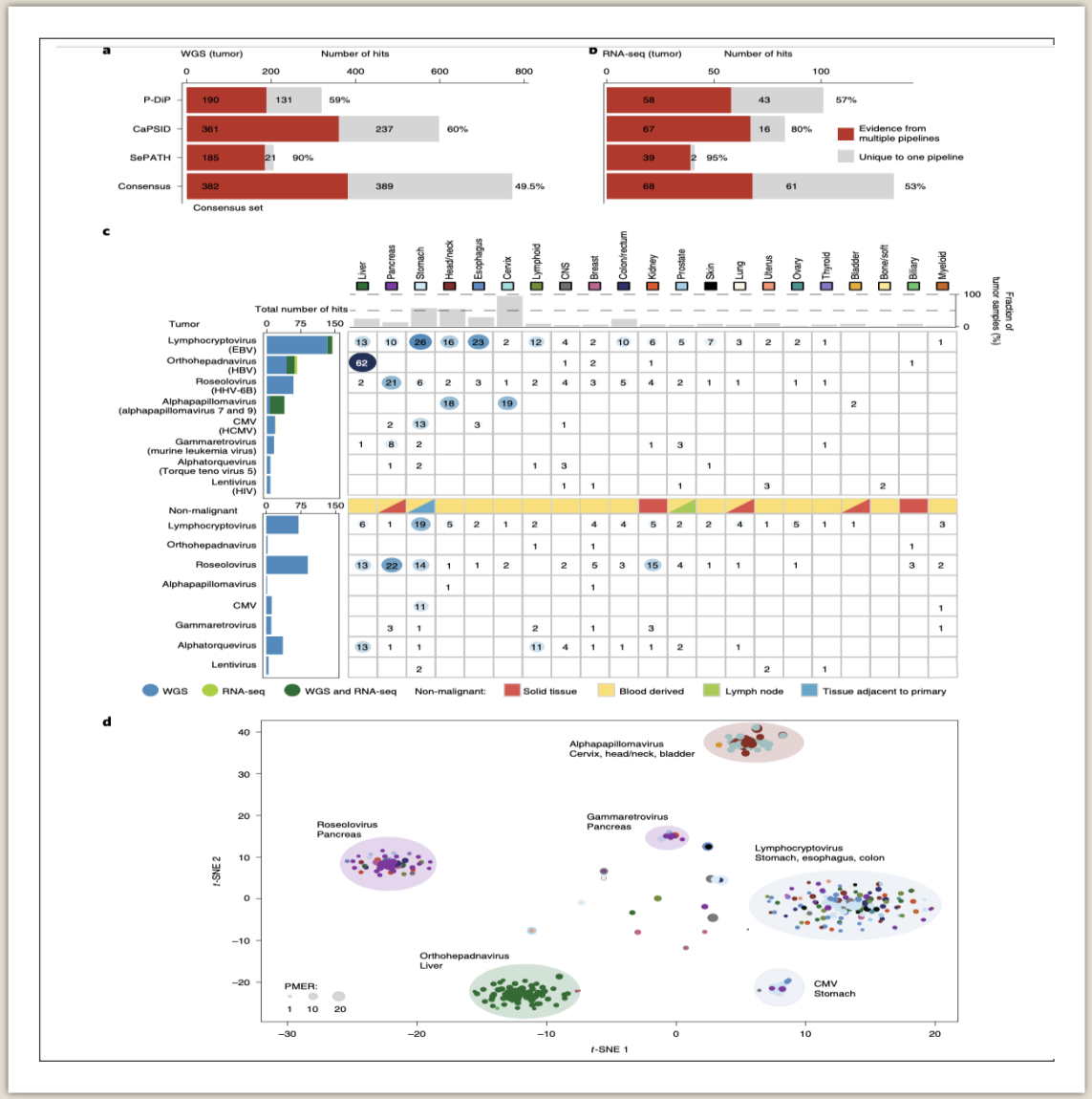
Marc Zapatka[1,20,21], Ivan Borozan[2,21], Daniel S. Brewer[3,4,21], Murat Iskar[1,21], Adam Grundhoff[5], Malik Alawi[5,6], Nikita Desai[7,8], Holger Sültmann[9,10], Holger Moch[11], PCAWG Pathogens[12], Colin S. Cooper[4,13], Roland Eils[14,15,16], Vincent Ferretti[17,18], Peter Lichter[1,10,20*] and PCAWG Consortium[19]

Sitong Chen



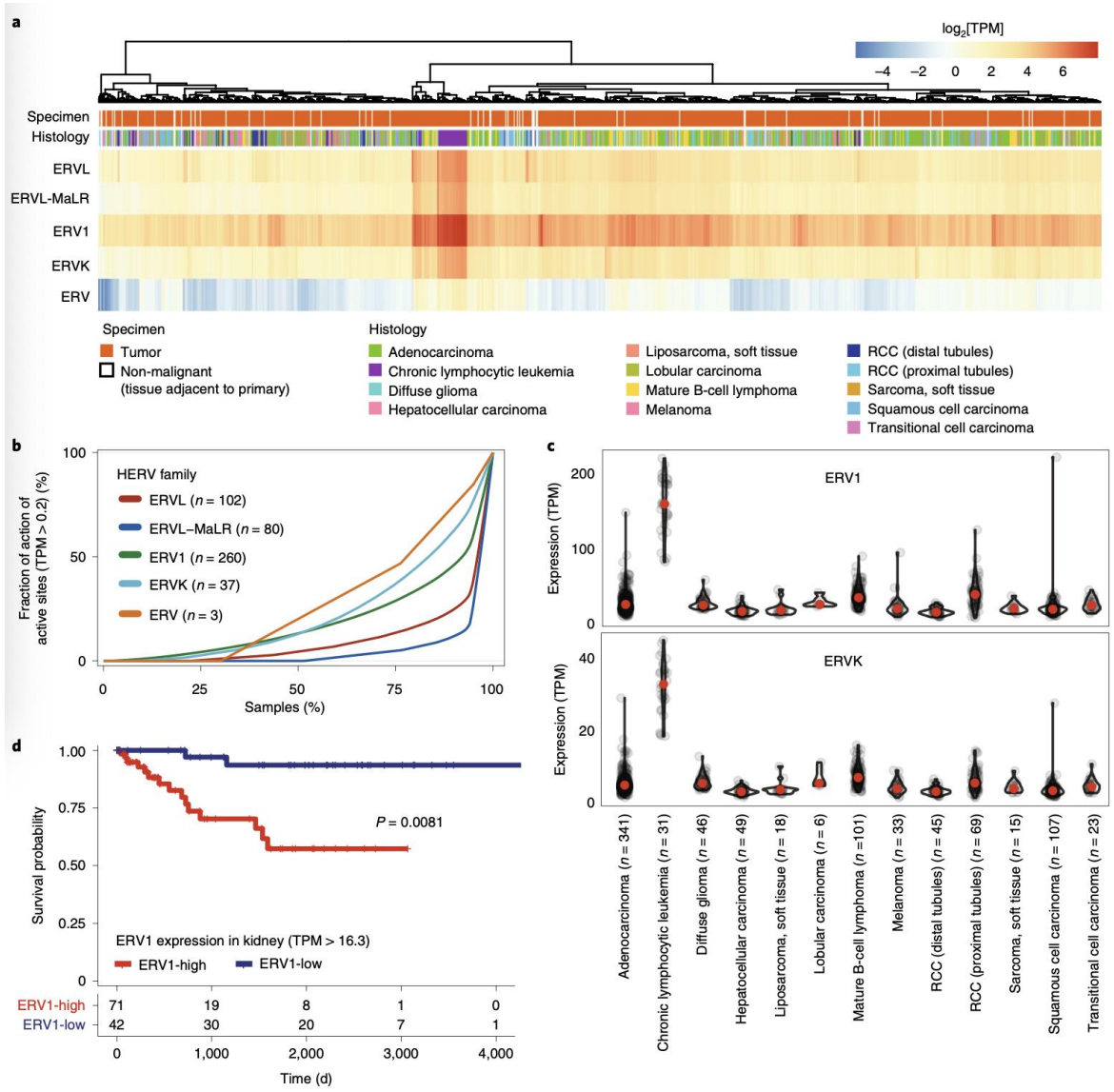
Overview

- A: Workflow.
- C: Summary of virus group



Consensus for detected viruses in WGS and RNA-seq data

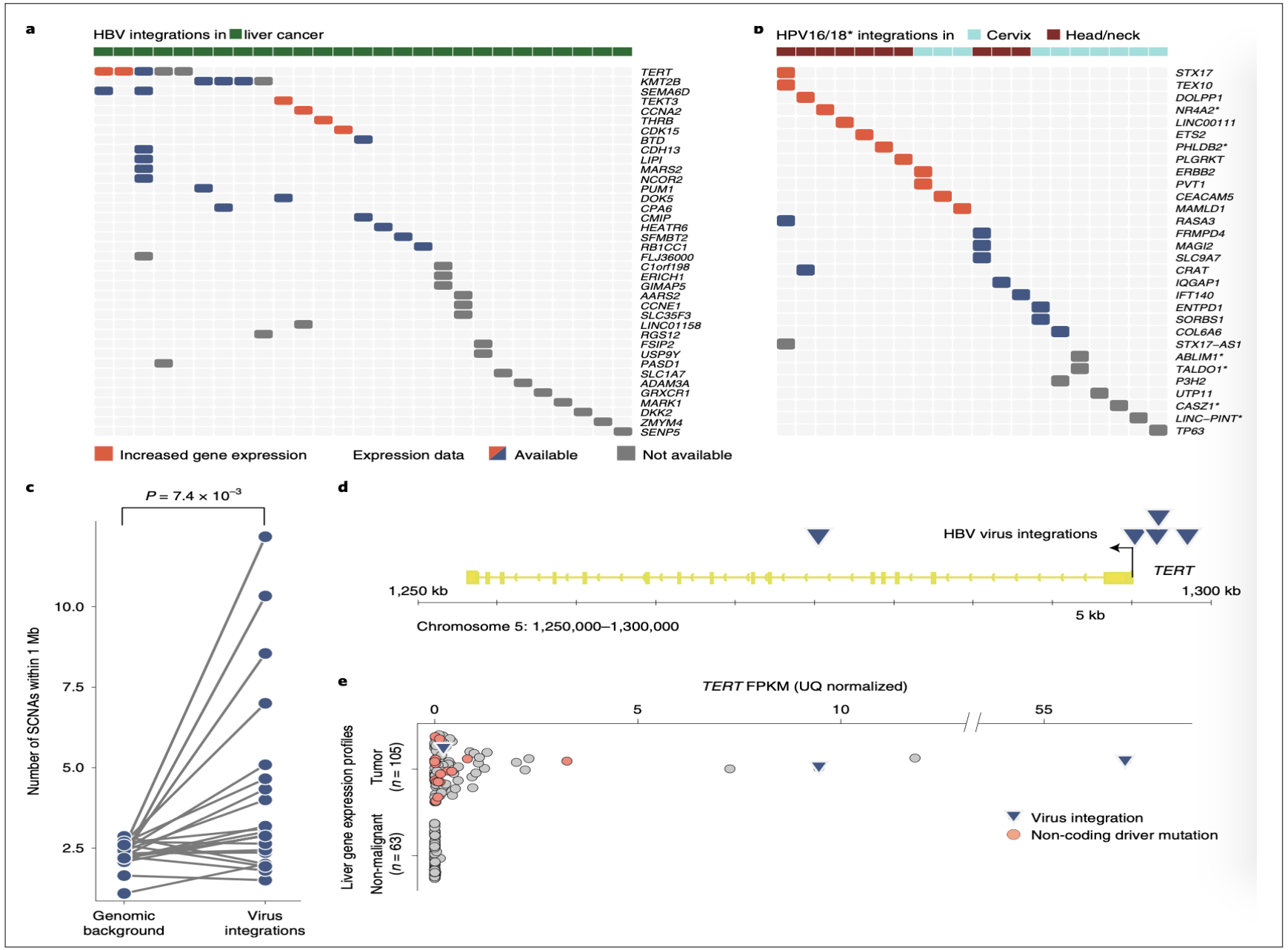
- consensus hit : Virus detection in a sample by at least two pipelines.
- CaPSID is the most sensitive pipeline.



Expression of ERVs

- Human endogenous retroviruses (HERV)

The effect of virus integration



Summary

1. Expression of HERVs is associated with prognosis in clear cell renal cell carcinoma
2. Viral integration will affect gene expression.

nature

Article | Open Access | Published: 05 February 2020

Genomic basis for RNA alterations in cancer

Qing Chen

Research Technician

Wei Li's Lab

About the article

Children's National

Comprehensive study

Database

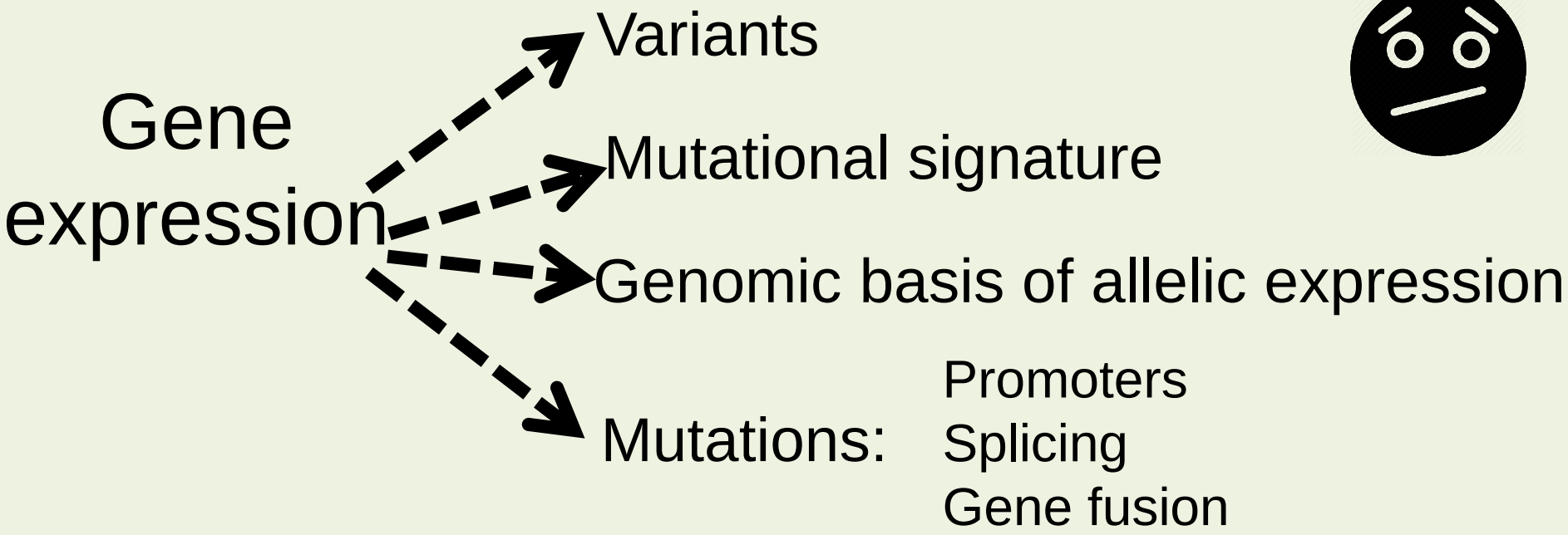
PCAWG
ICGC
TCGA

Cancer associated gene alterations

RNA phenotypes ← Genetic changes

1188 donors
27 tumor types
Transcriptomes
WGS

About the article



Landscape of RNA alterations

Co-occurrence of RNA and DNA alterations

Recurrent RNA alteration in driver genes

Some terminologies used in the article

Children's National

Pan-Cancer Analysis of Whole Genomes (PCAWG)
International Cancer Genome Consortium (ICGC)
The Cancer Genome Atlas (TCGA)

Quantitative trait locus (QTL) analysis
Genotype-Tissue Expression (GTEx) Project

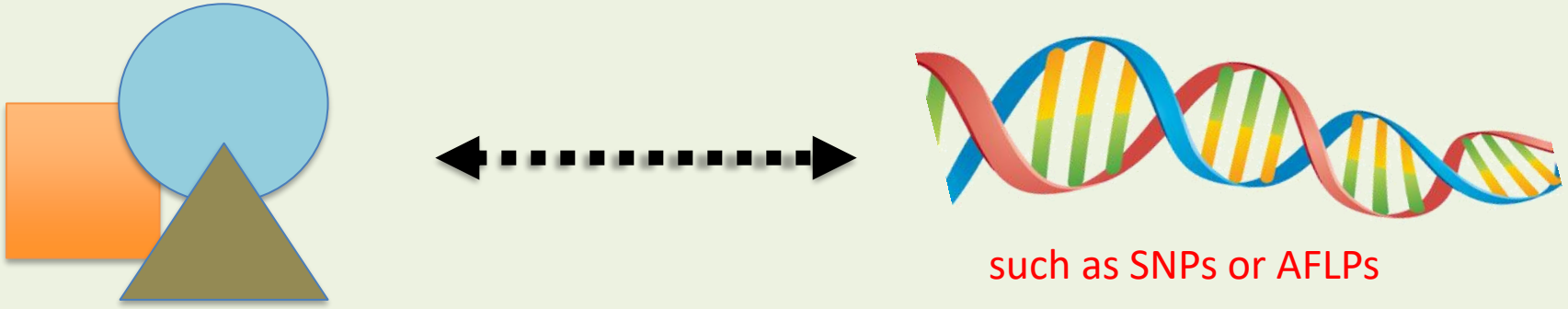
single-nucleotide variants (SNVs)
somatic copy-number alterations (SCNAs)

allele-specific expression (ASE) analysis

QTL

Children's National

Quantitative trait locus (QTL) analysis is a statistical method that links two types of information—phenotypic data (trait measurements) and genotypic data (usually molecular markers)—in an attempt to explain the genetic basis of variation in complex traits (Falconer & Mackay, 1996; Kearsey, 1998; Lynch & Walsh, 1998).

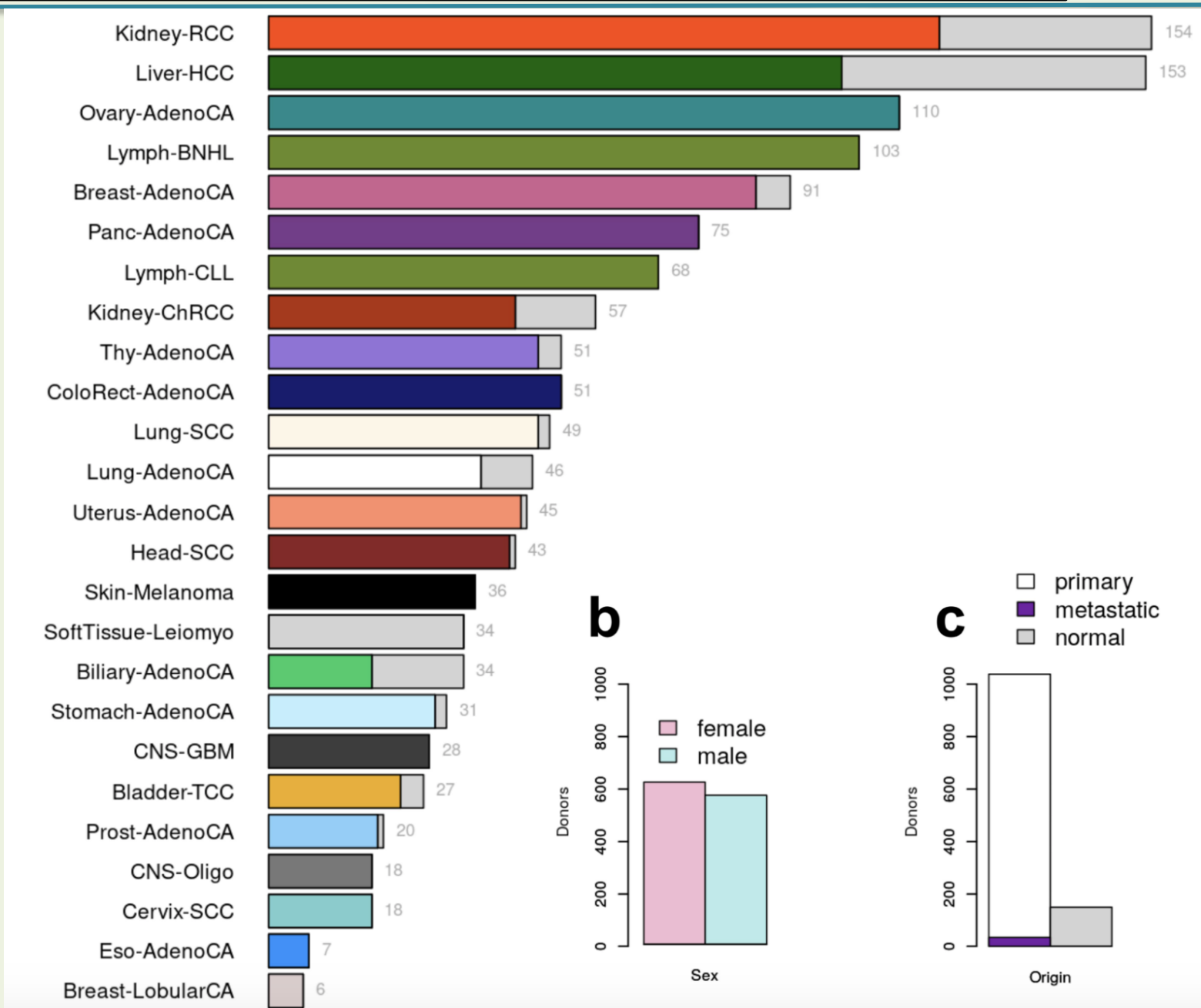


such as SNPs or AFLPs

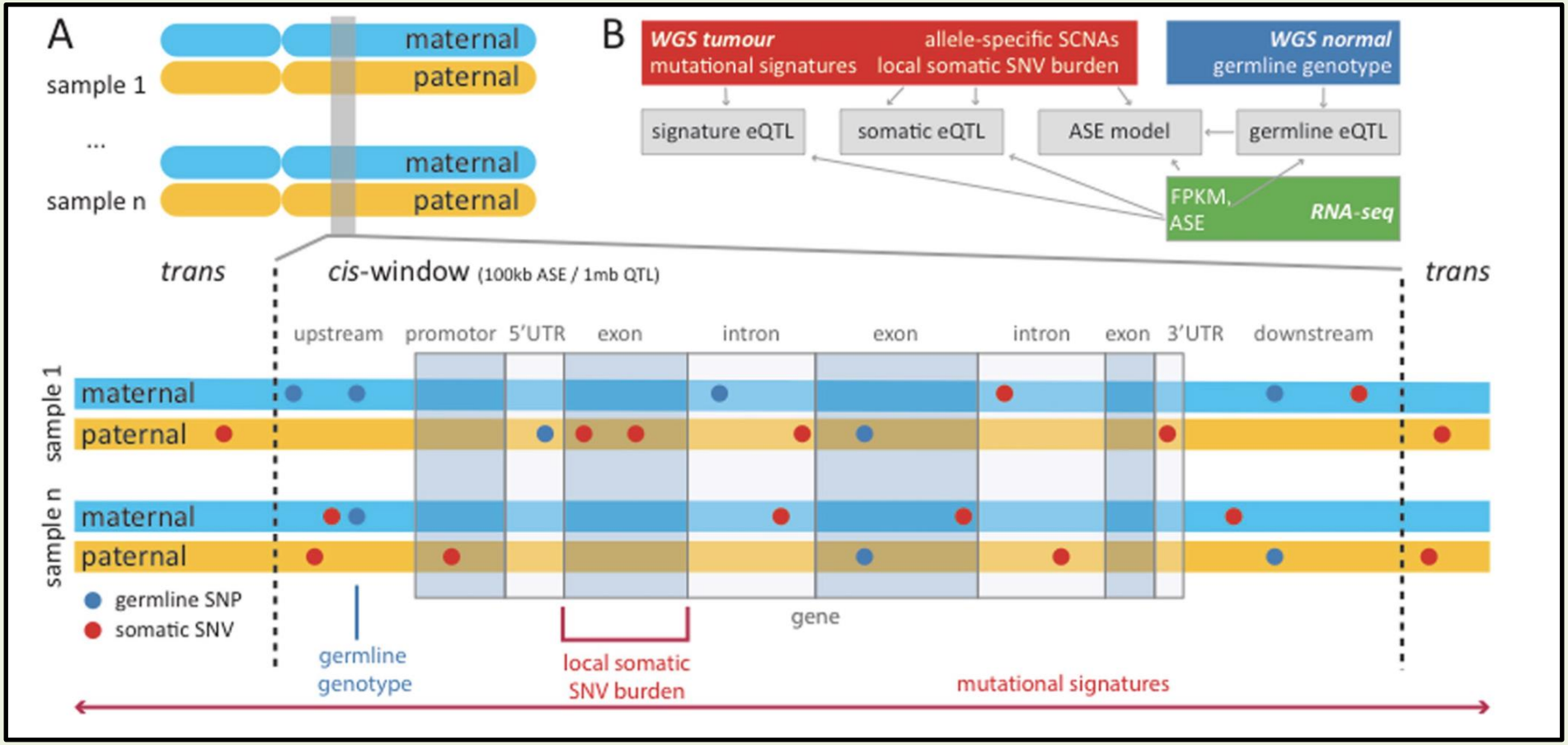
Expression quantitative trait loci (eQTLs) are genomic loci that explain all or a fraction of variation in expression levels of mRNAs

Splicing quantitative trait loci (abbreviated sQTLs or splicing QTLs) are quantitative trait loci that regulate alternative splicing of pre-mRNA.

Pan-cancer expression profiling of 1,188 PCAWG donors

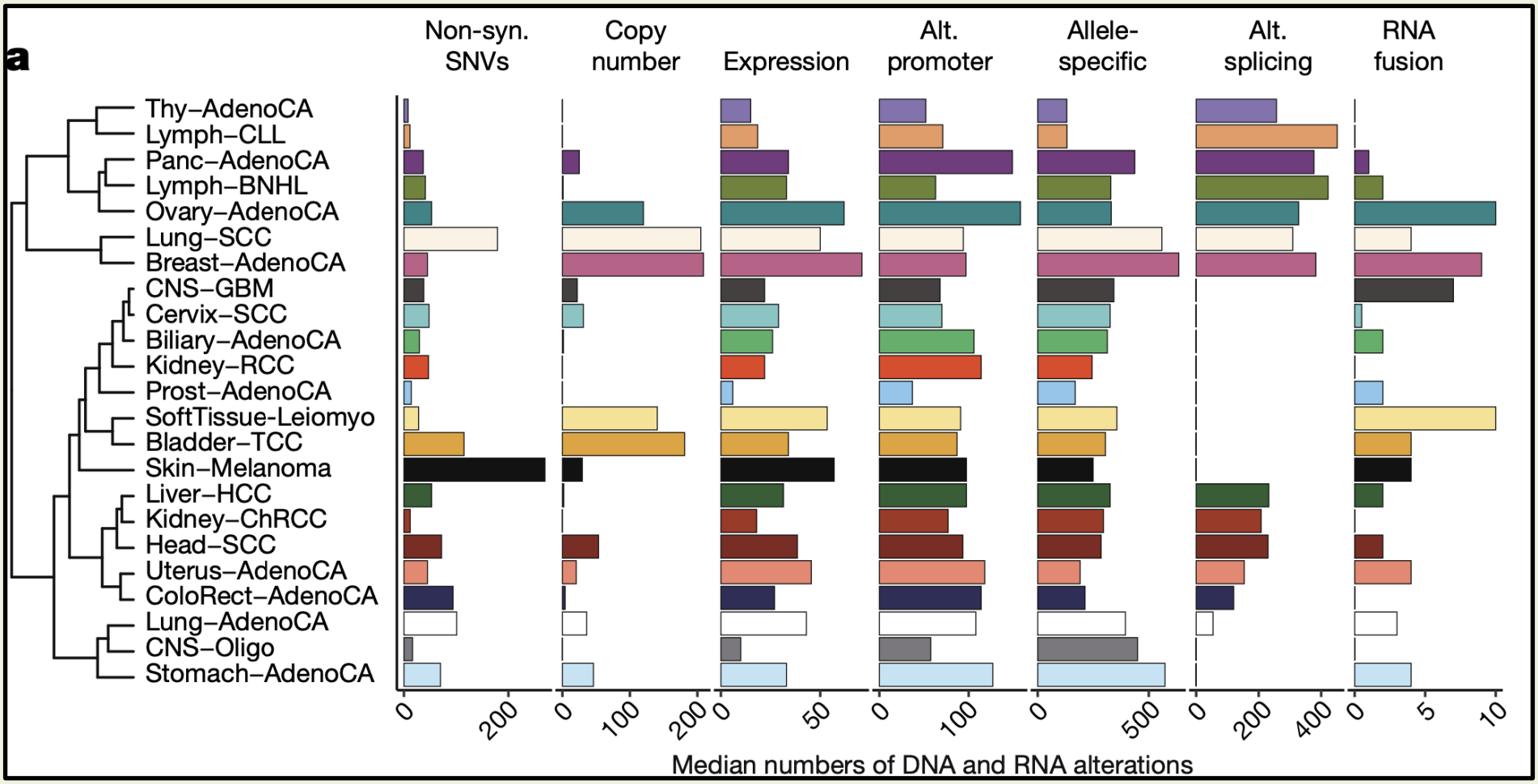


Overview of the different sources of genetic variation considered in the analysis



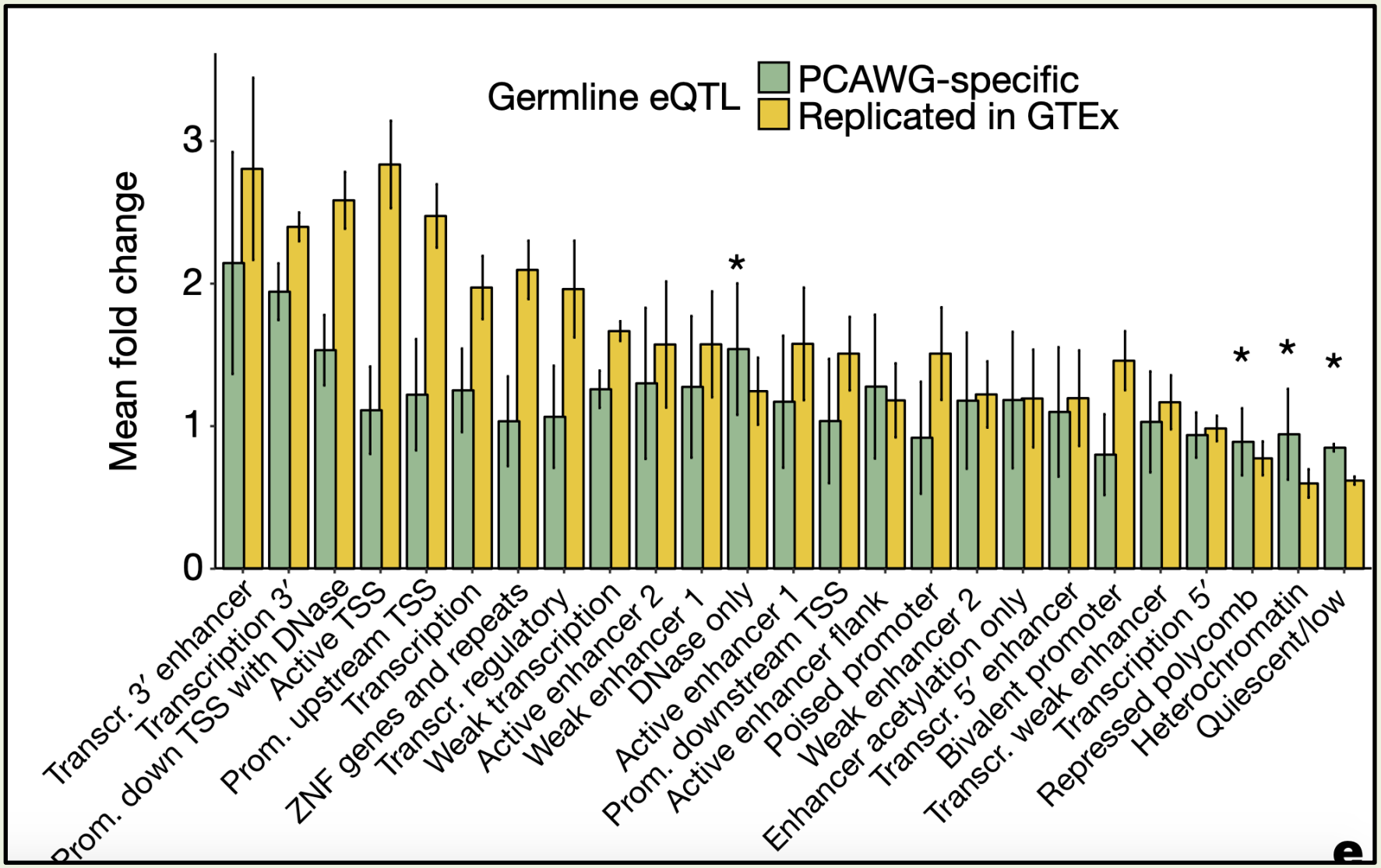
Global view of DNA and RNA alterations that affect tumors

Children's National



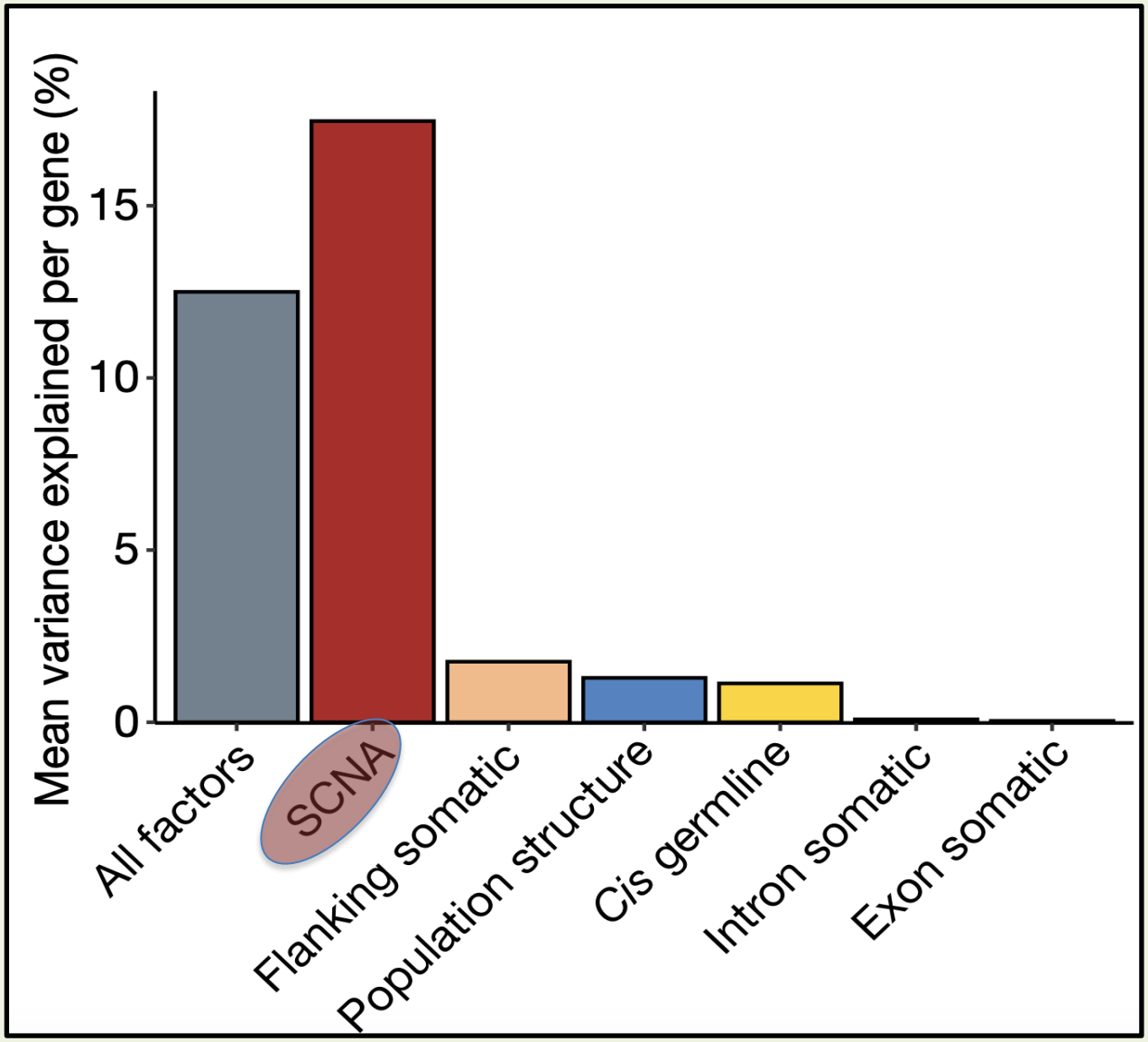
different types of cancer contain distinct combinations of DNA- and RNA-level alterations

Cancer-specific germline *cis*-eQTLs



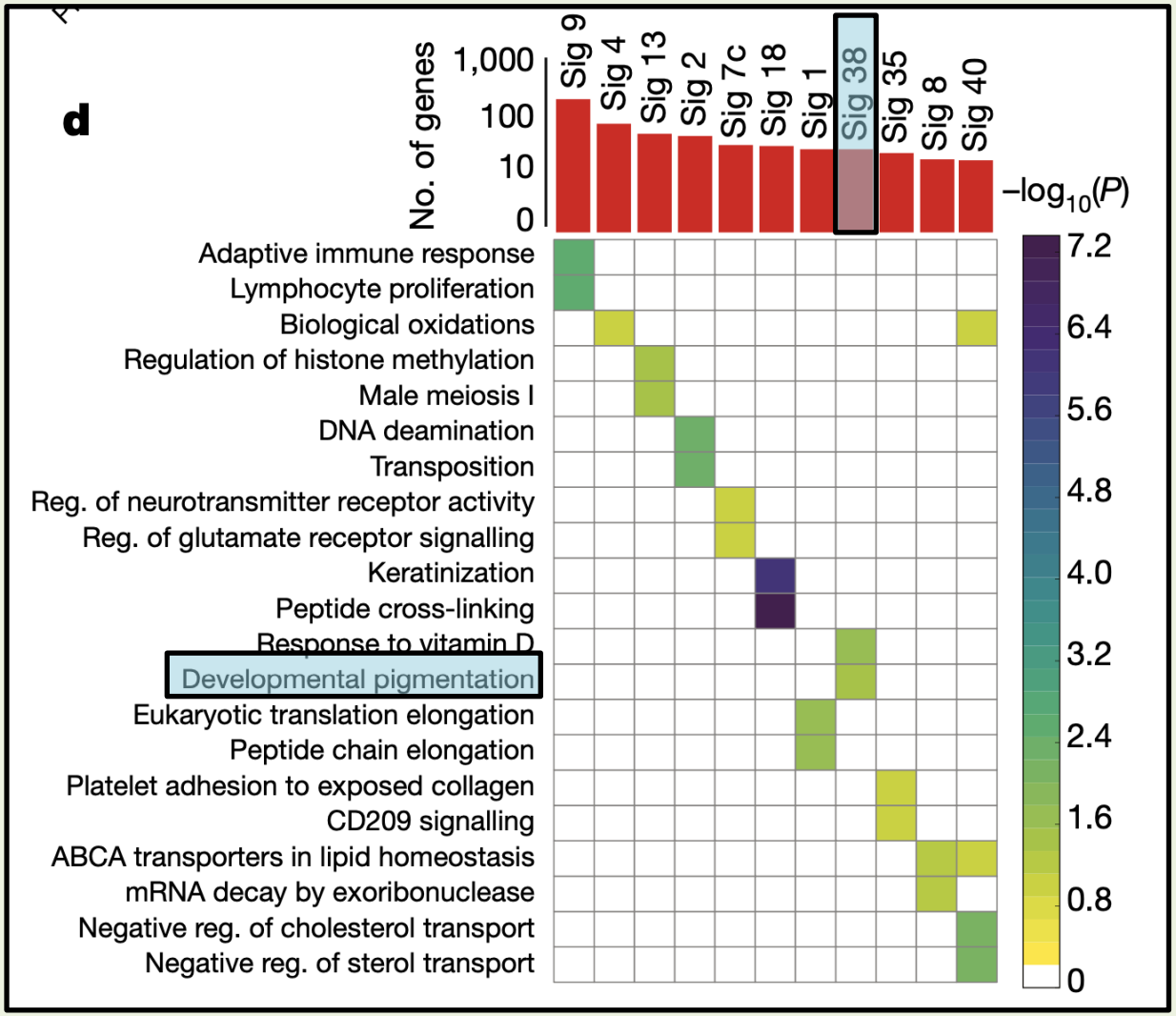
germline framework of gene expression regulation is largely conserved in cancer tissues.

Variance component analysis for gene expression levels



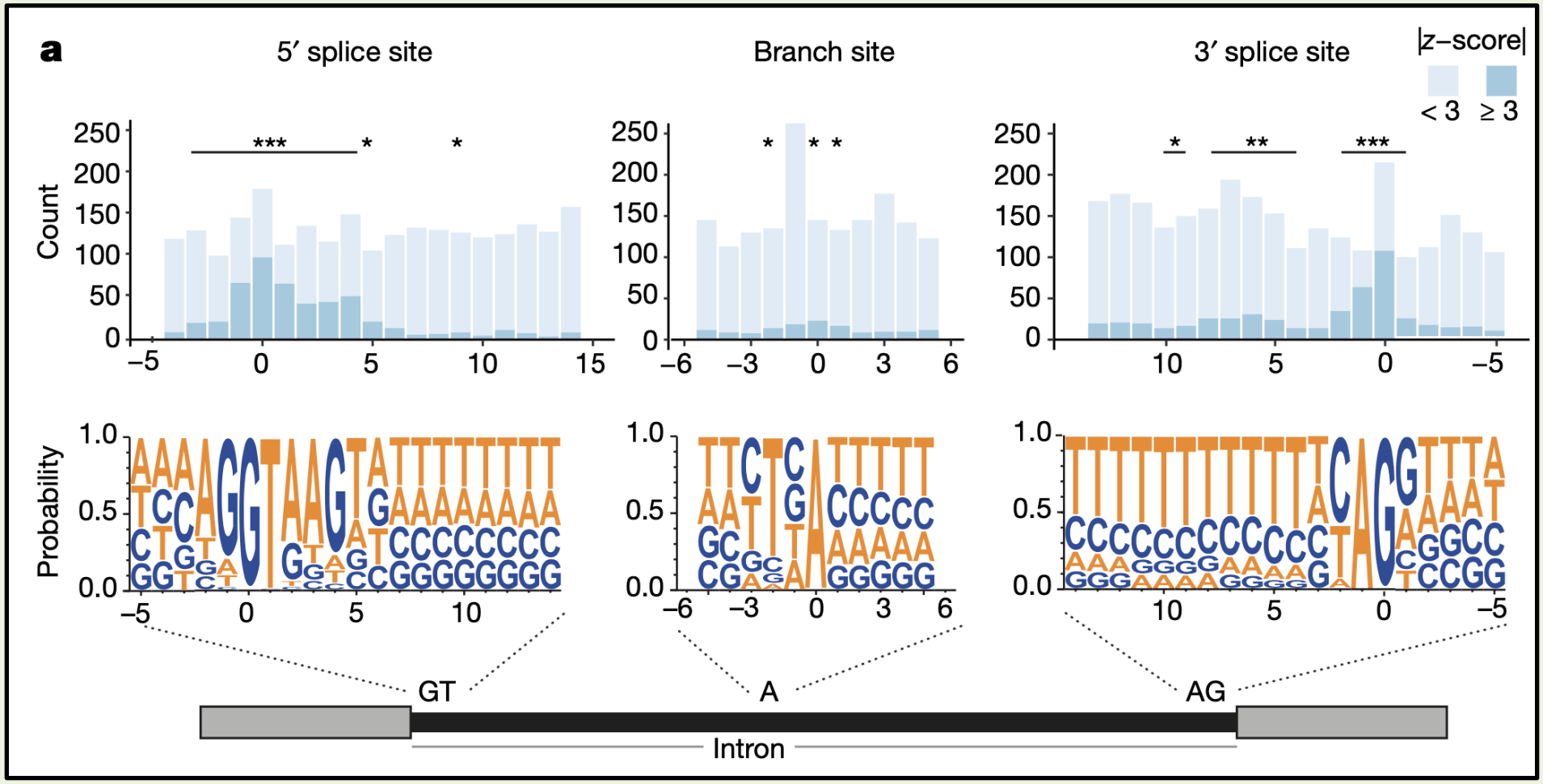
SCNAs as the major driver of expression variation

pan-cancer association analysis between genome-wide mutational signatures and gene expression levels



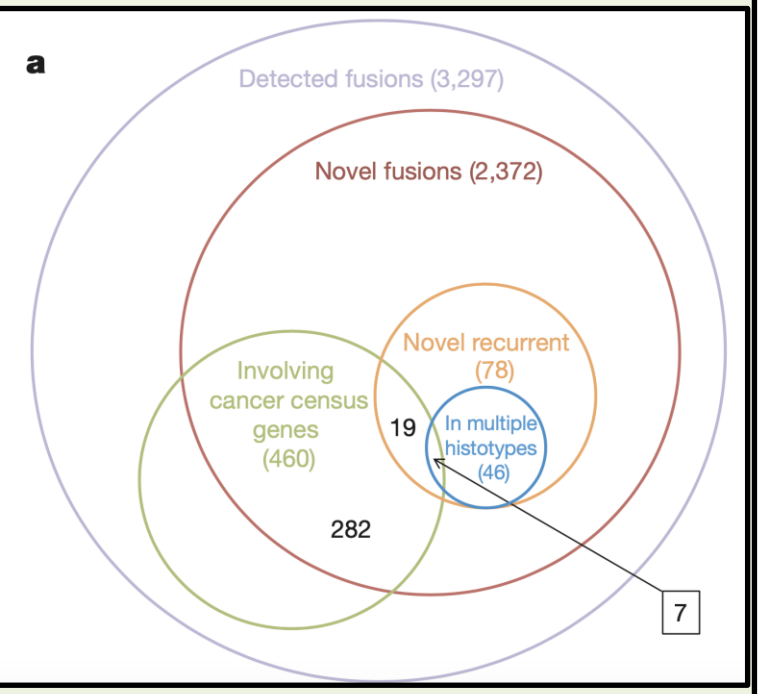
significant associations between mutational signatures (Sig) and gene expression.

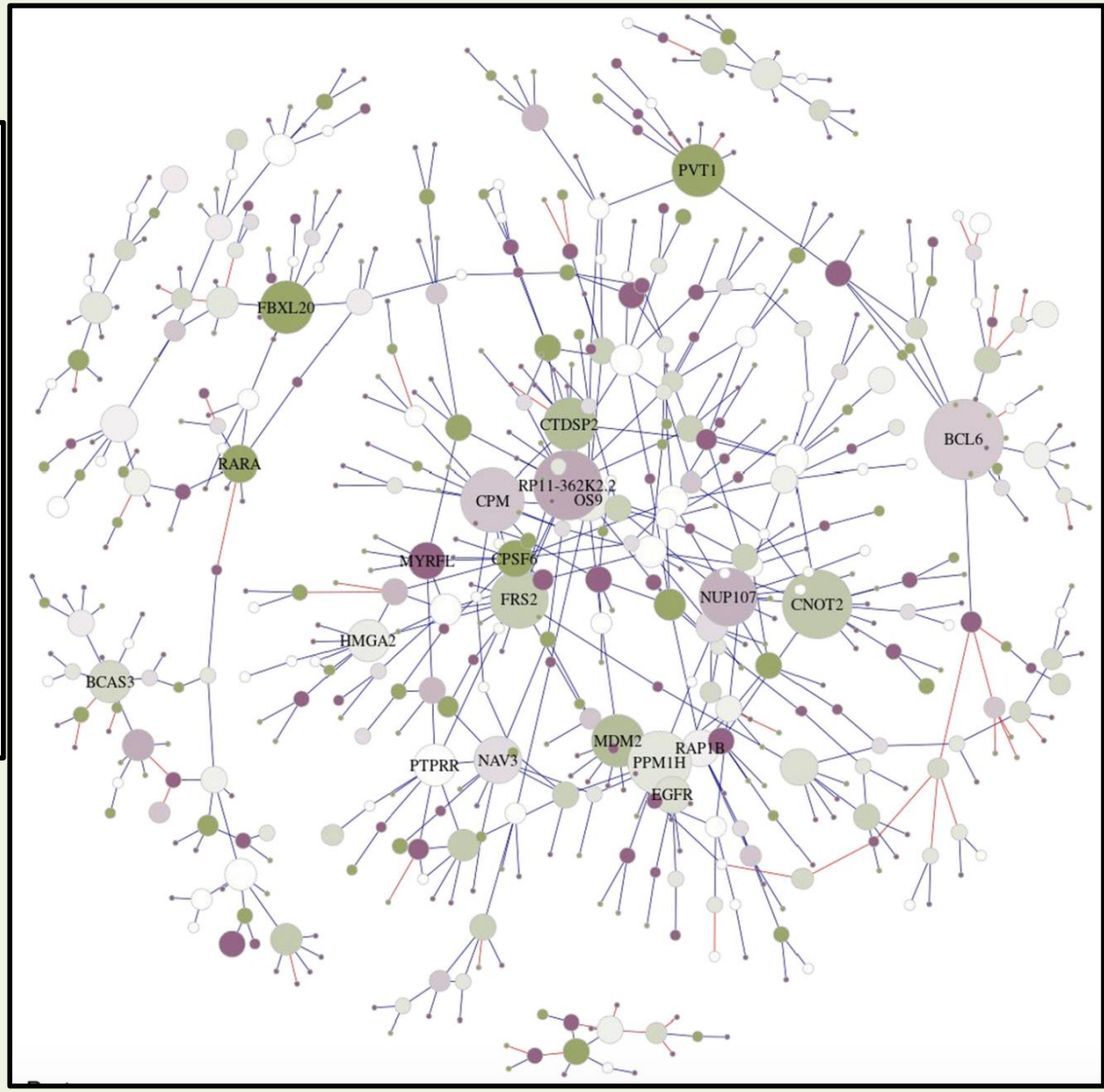
Position-specific effect of somatic mutations on alternative splicing



somatic mutations in these sites can affect splicing.

Patterns of gene fusions across cancer





Network with connected clusters of at least 10 genes.

Comments on this paper

Children's National

Outline was not so clearly

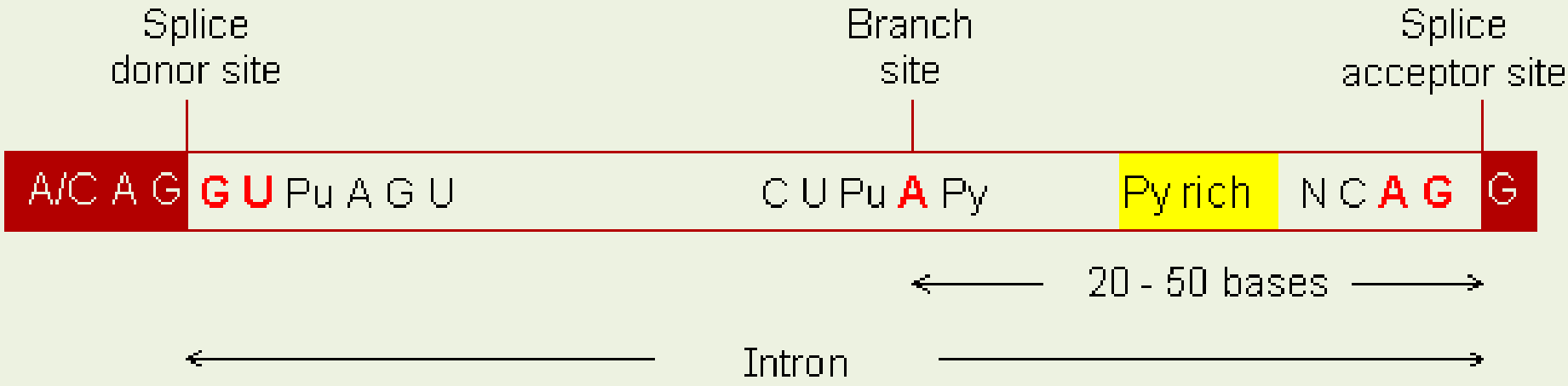
Novelty

Main figure ? extended figure

New analysis methods

allele-specific expression (ASE)

Thank you for your attention !


Splice
donor site
Branch
site
Splice
acceptor site
A/C A G G U Pu A G U
C U Pu A Py
Py rich
N C A G
G
20 - 50 bases
Intron

ARTICLE

https://doi.org/10.1038/s41467-019-13885-w OPEN

High-coverage whole-genome analysis of 1220 cancers reveals hundreds of genes deregulated by ~~rearrangement-mediated *cis*-regulatory alterations~~

Somatic Structural Variants (SVs)

Yiqun Zhang[1], Fengju Chen[1], Nuno A. Fonseca[2], Yao He[3,4], Masashi Fujita[5], Hidewaki Nakagawa[5], Zemin Zhang[3,4], Alvis Brazma[2], PCAWG Transcriptome Working Group, PCAWG Structural Variation Working Group, Chad J. Creighton[1,6,7,8*] & PCAWG Consortium

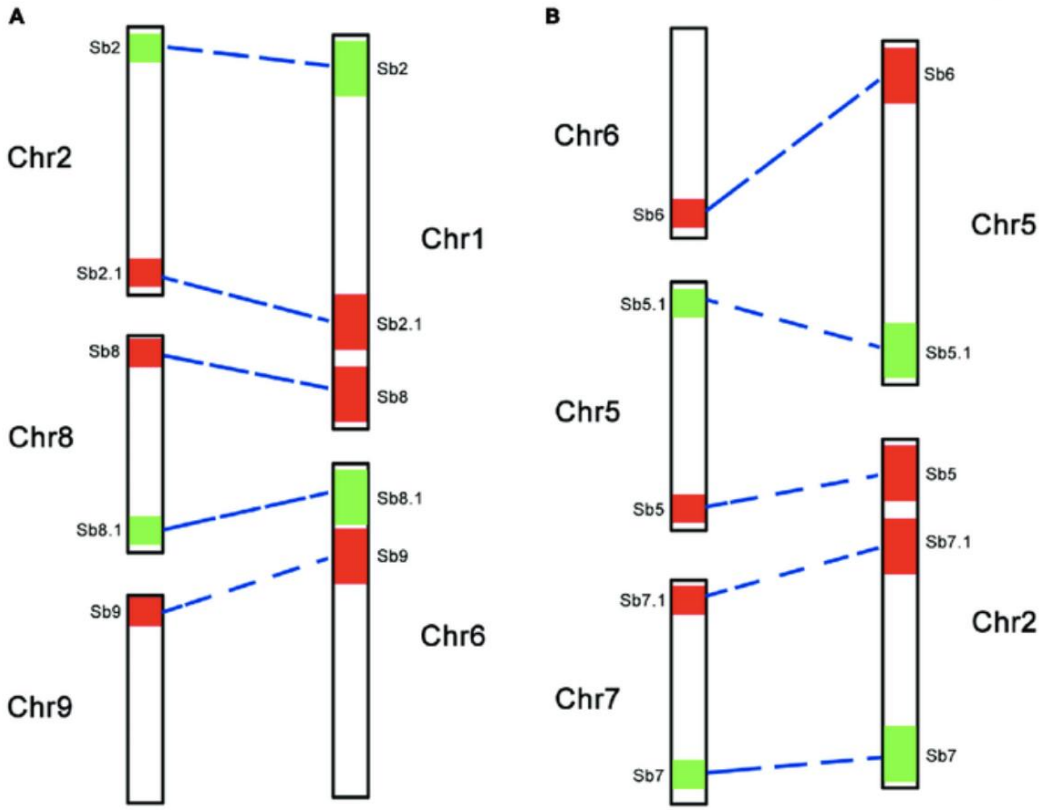
Lin Yang

Introduction

Breakpoints in the genome are locations on a chromosome where DNA might get deleted, inverted, or swapped around.

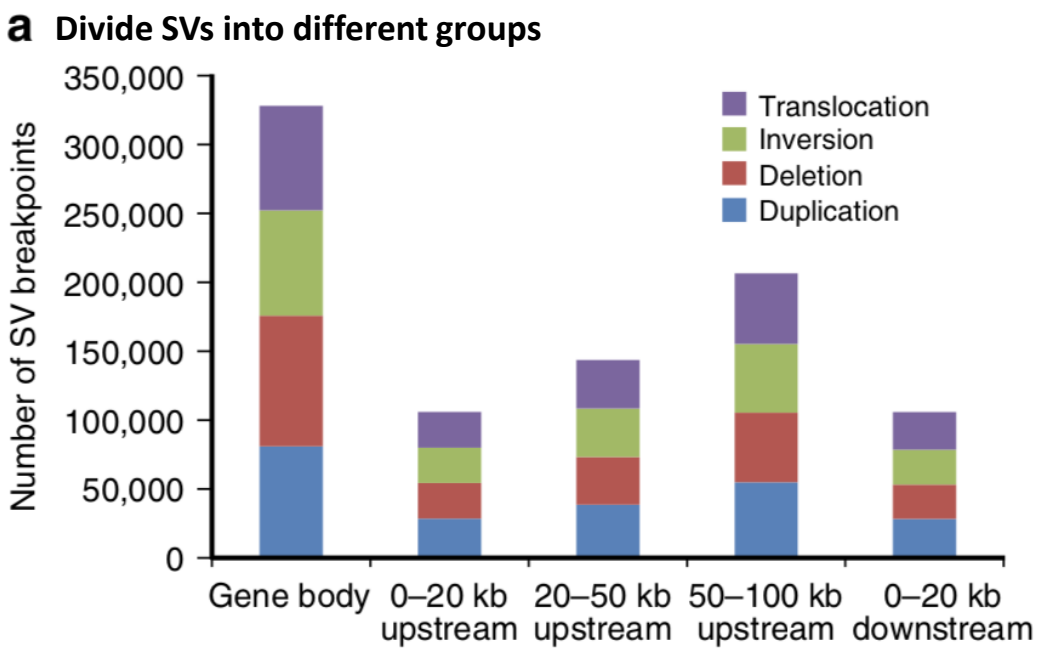
Breakpoints associated with Genomic rearrangement can potentially alter the regulation of nearby genes, e.g., by disrupting specific regulatory elements.

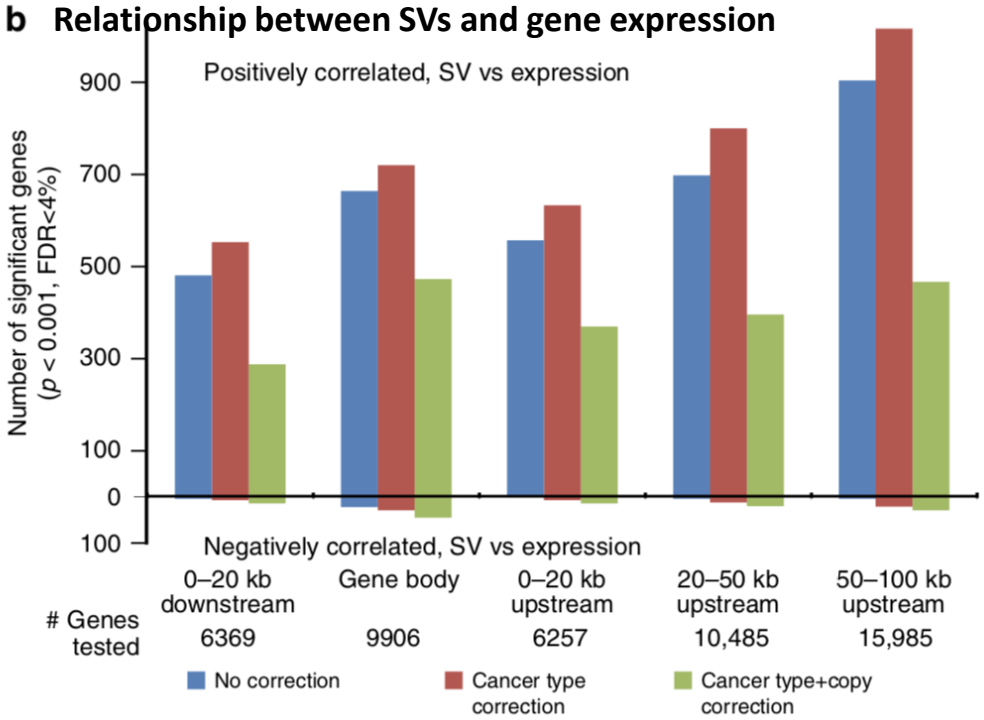
Somatic Structural Variants (SVs), representing genomic rearrangement events, each event involving two breakpoints from different genomic coordinates becoming fused together

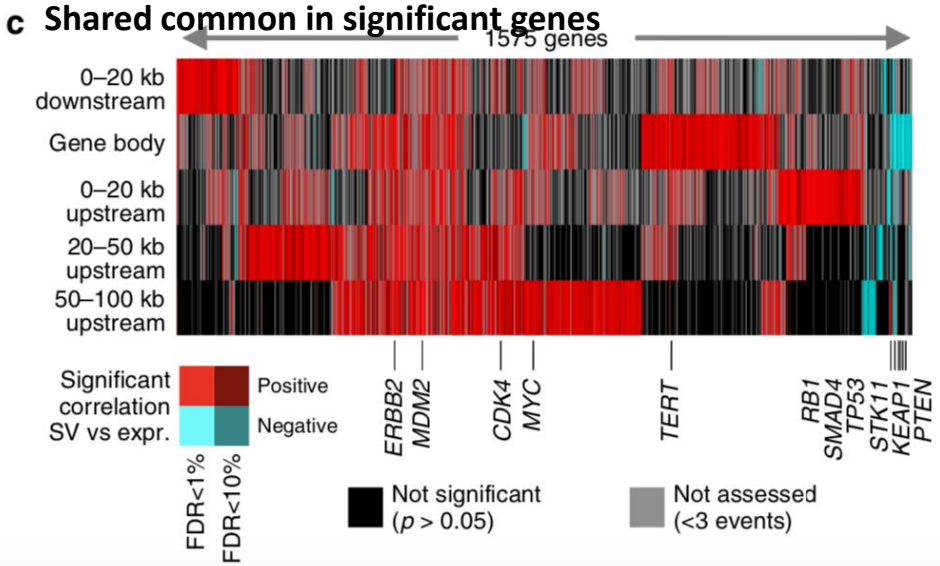


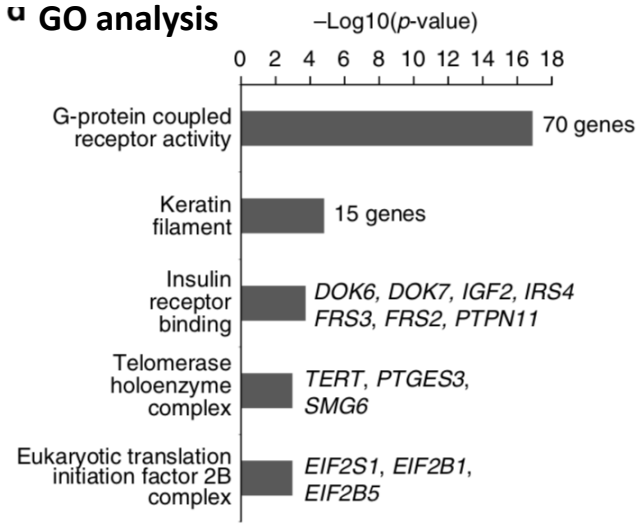
Children's National

Most SVs are positively correlated with gene expression



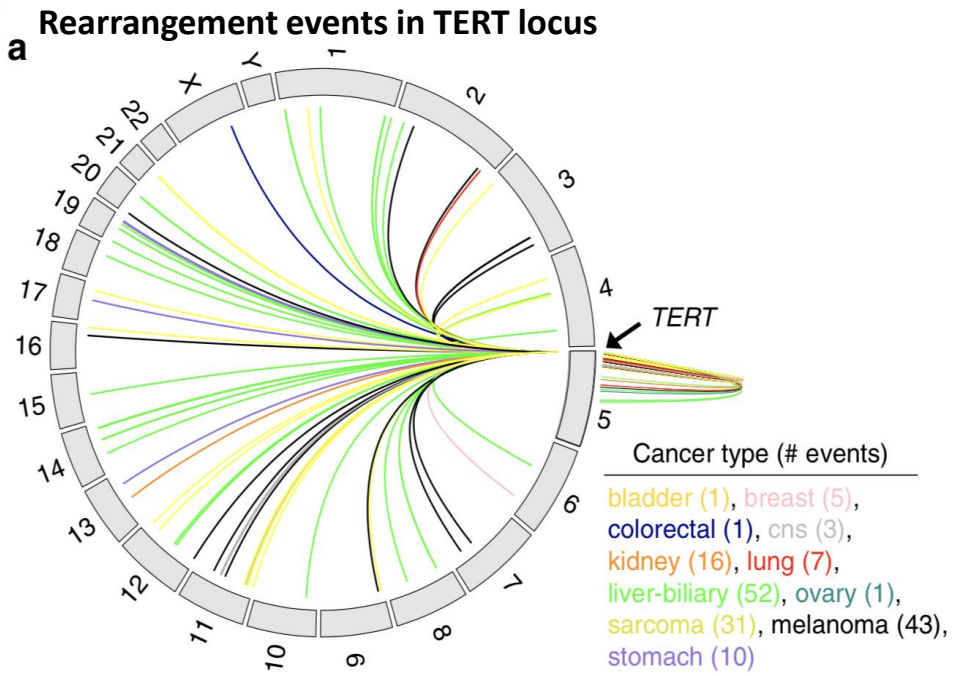


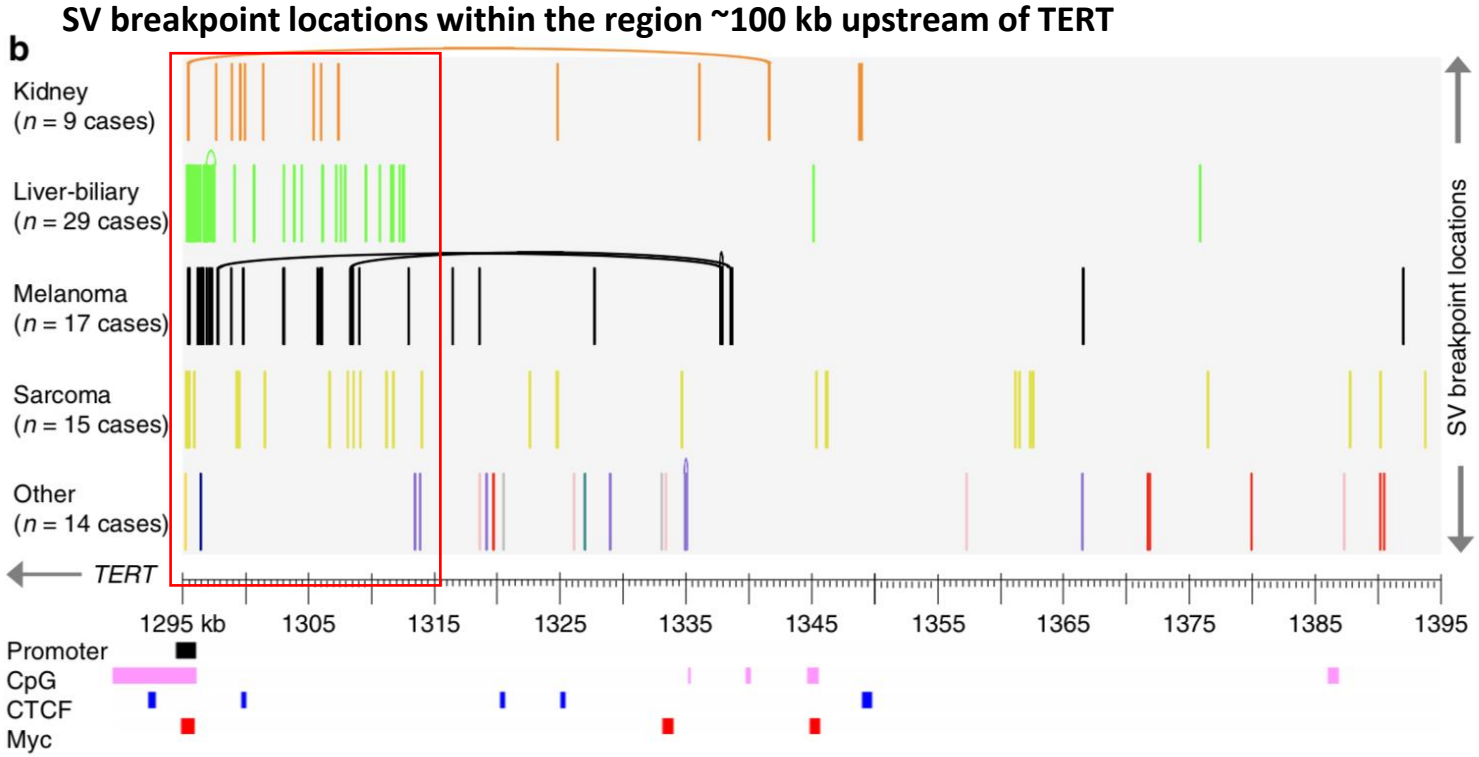


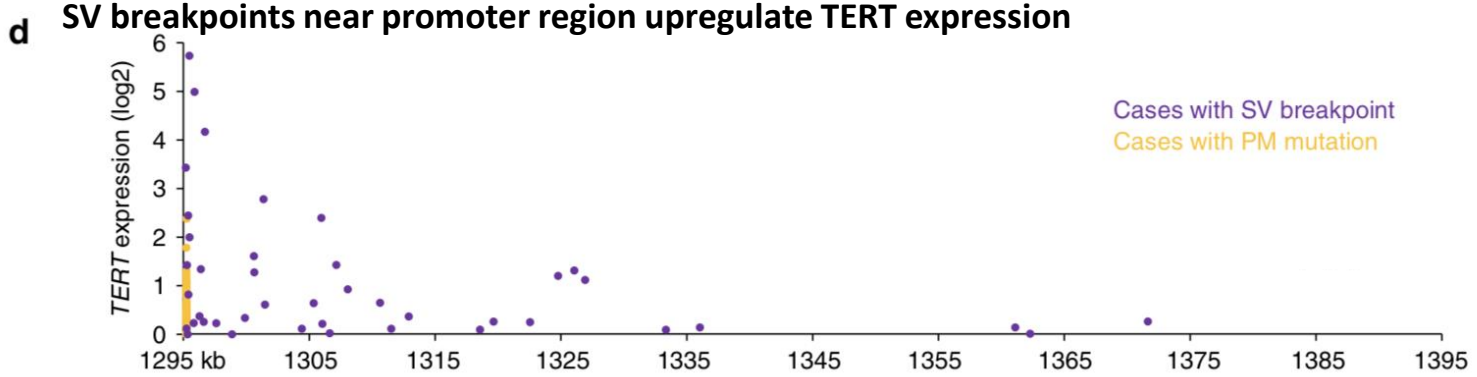


Children's National

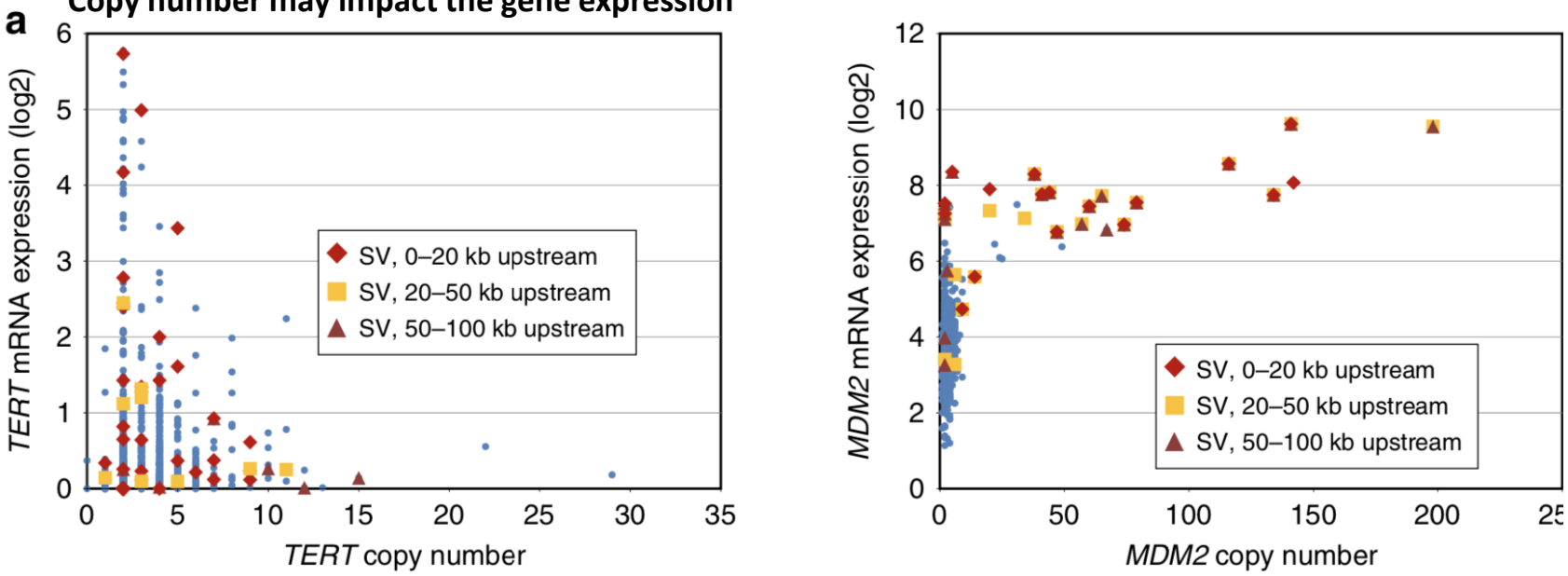
Key driver genes in cancer impacted by nearby SV breakpoints



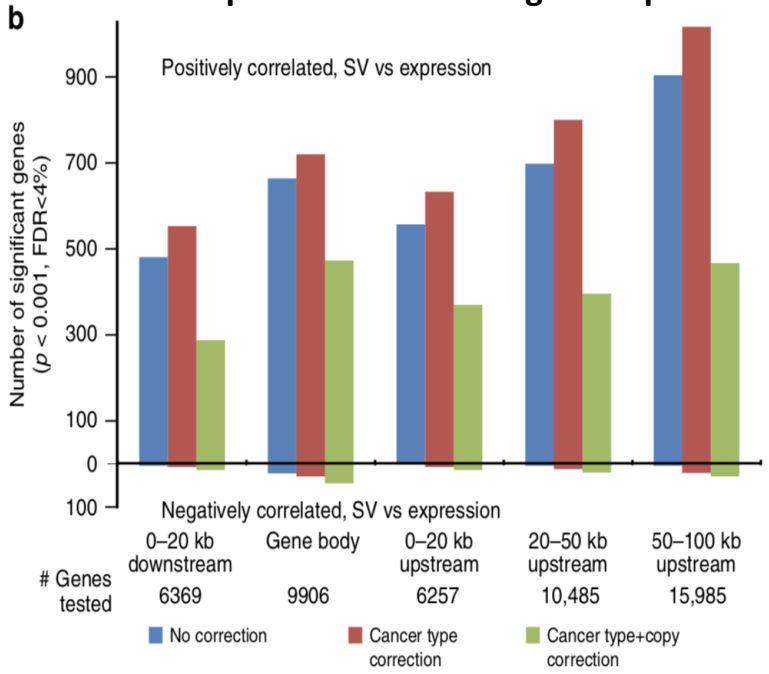




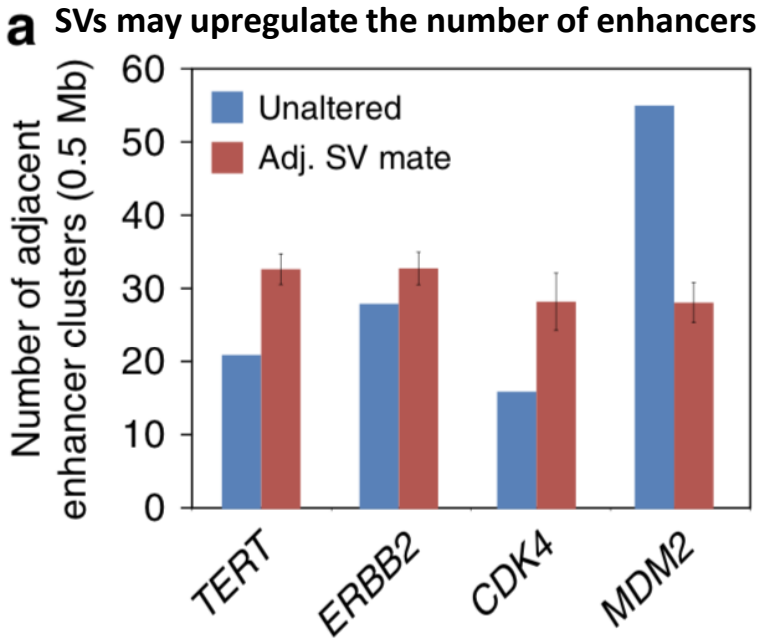
Copy number may impact the gene expression

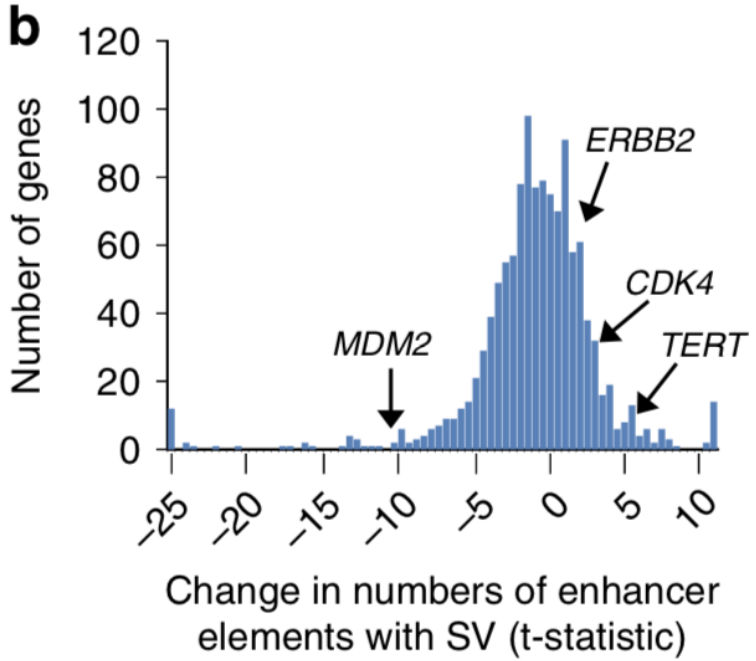


Relationship between SVs and gene expression



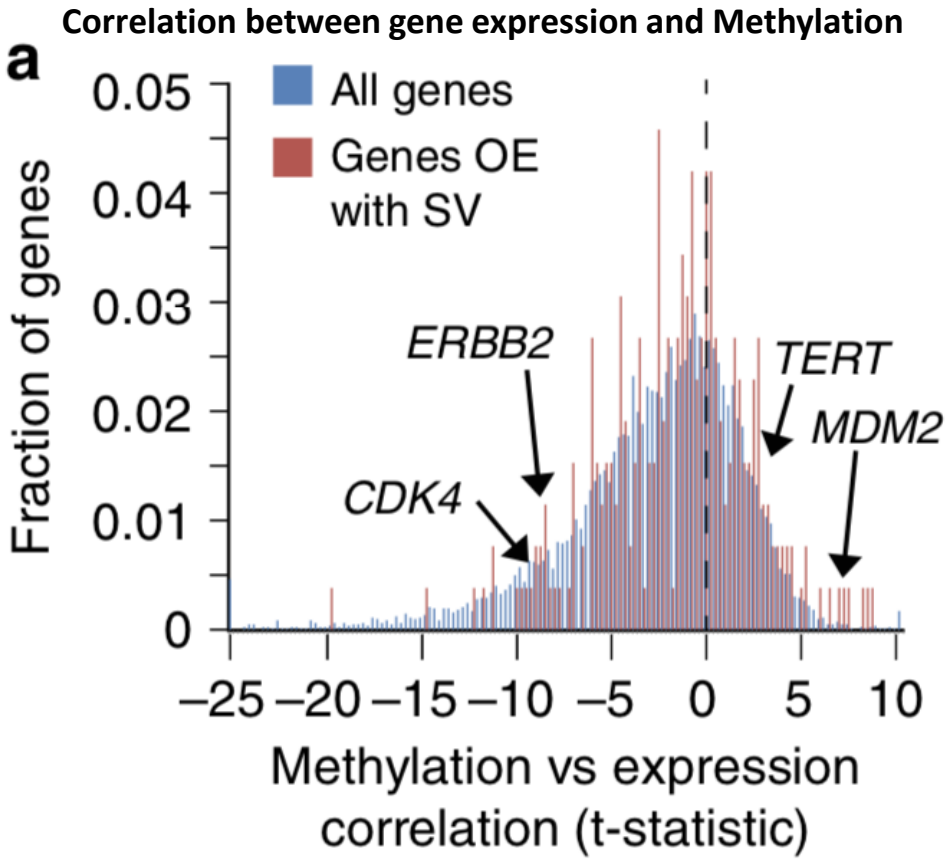
SVs may upregulate the number of enhancers

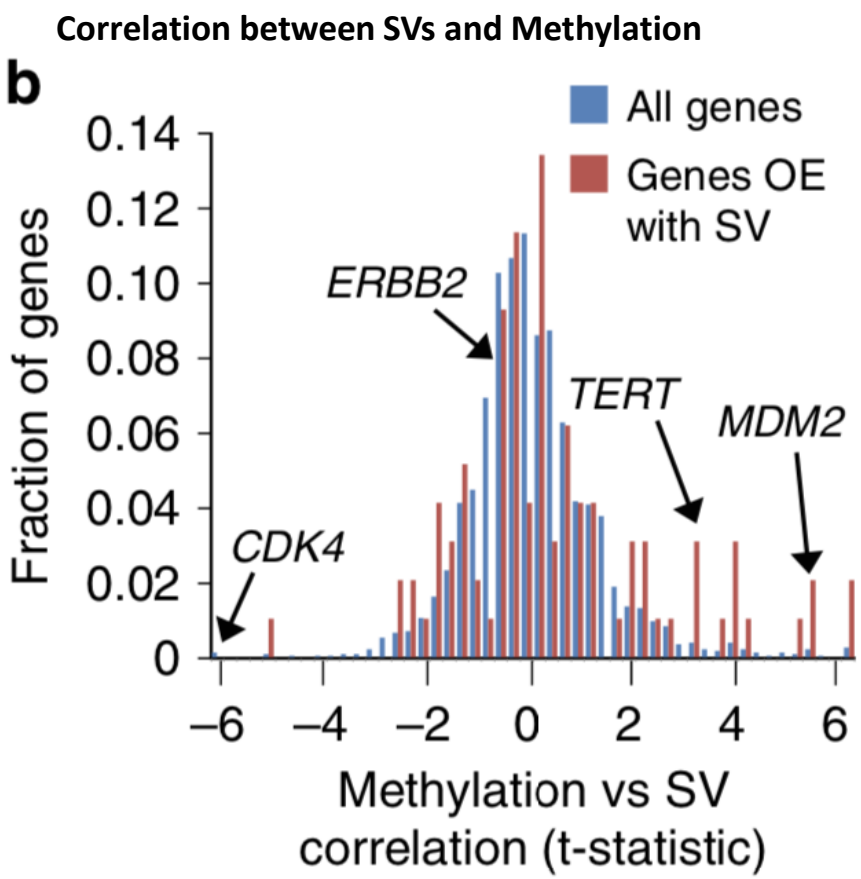




Children's National

SVs may involve inactivation of repressor elements





Children's National

Summary

- For the majority of these genes, expression increases rather than decrease with corresponding breakpoint events.
- For many genes, SVs are significantly associated with increased numbers or greater proximity of enhancer regulatory elements near the gene
- SVs may involve inactivation of repressor elements

Children's National