Weill Cornell Medicine

Coding and noncoding mutations across cancer types

Ekta Khurana, PhD

Assistant Professor

Meyer Cancer Center, Institute for Computational Biomedicine & Institute for Precision Medicine

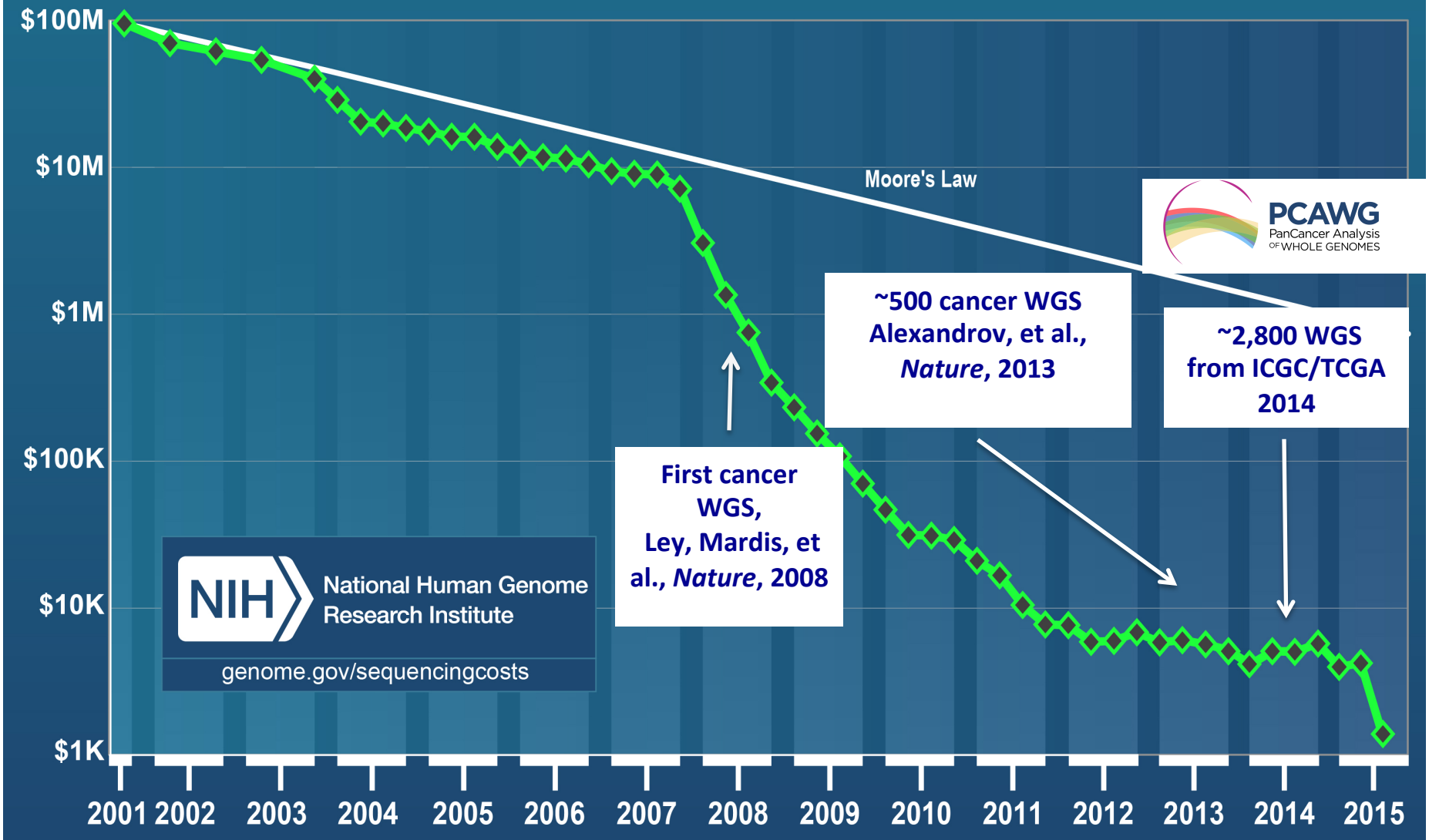
Department of Physiology and Biophysics

Weill Cornell Medicine, New York, NY

ekk2003@med.cornell.edu

@ekta_khurana

Number of cancer whole genomes sequenced



Genomic variants identified from sequencing

Human Ref.

ATGAACTGCAATTTCCAGAAGCATGCACCCTTGGAAG - - - TCTA

ATGAACTGCAAATTCCAGAAGCATG - - - - CTTGGAAGAGTTCTA

SNP Deletion Insertion

Small Indels < 50 bp

Large structural variants

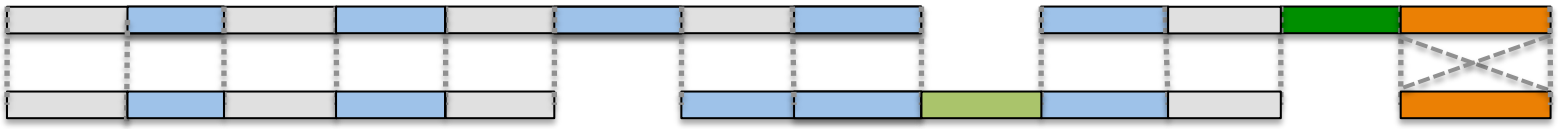
Duplications

Insertions

Deletions

Inversion

Human Ref.

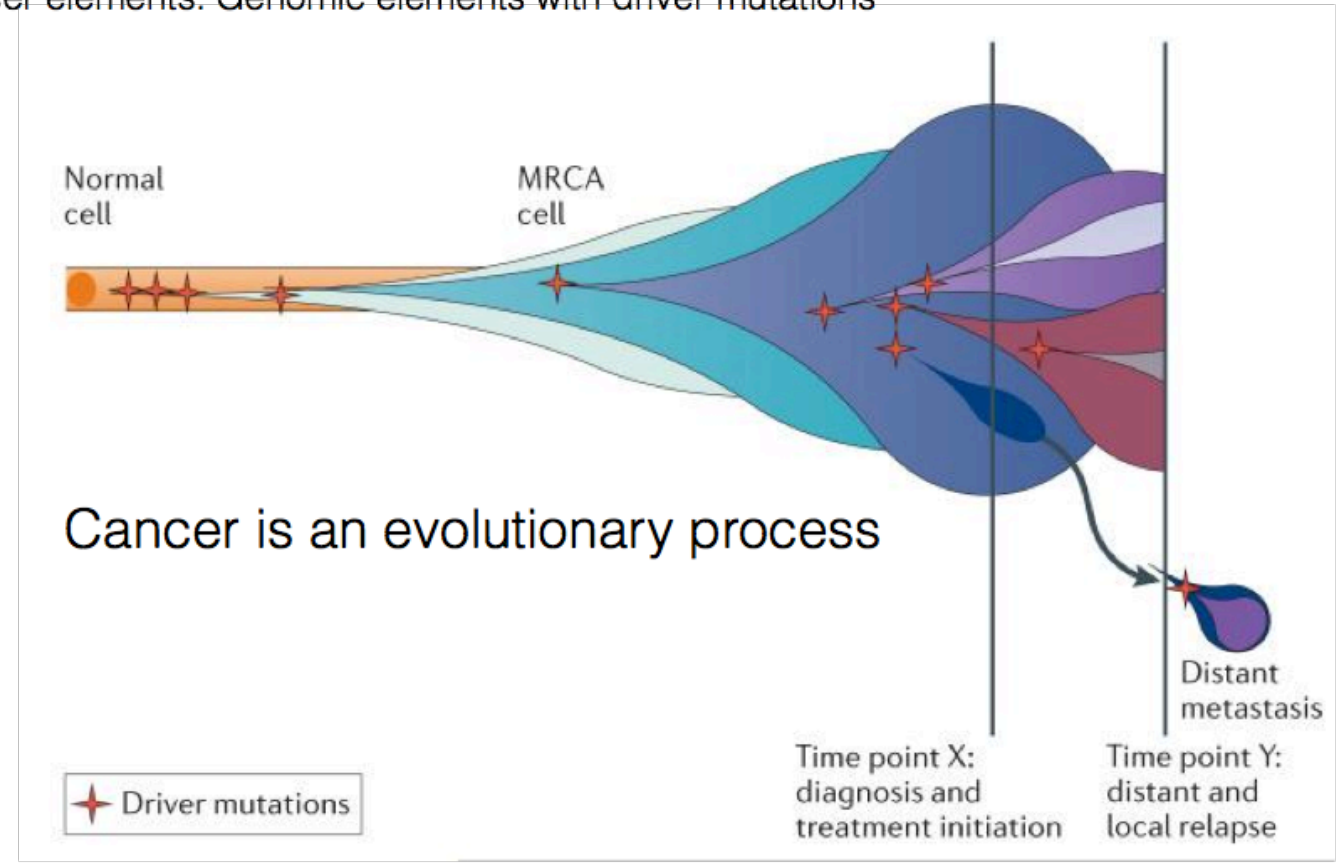


An average human genome contains ~4 million inherited variants and a tumor genome contains thousands of somatic variants.

Drivers versus Passengers

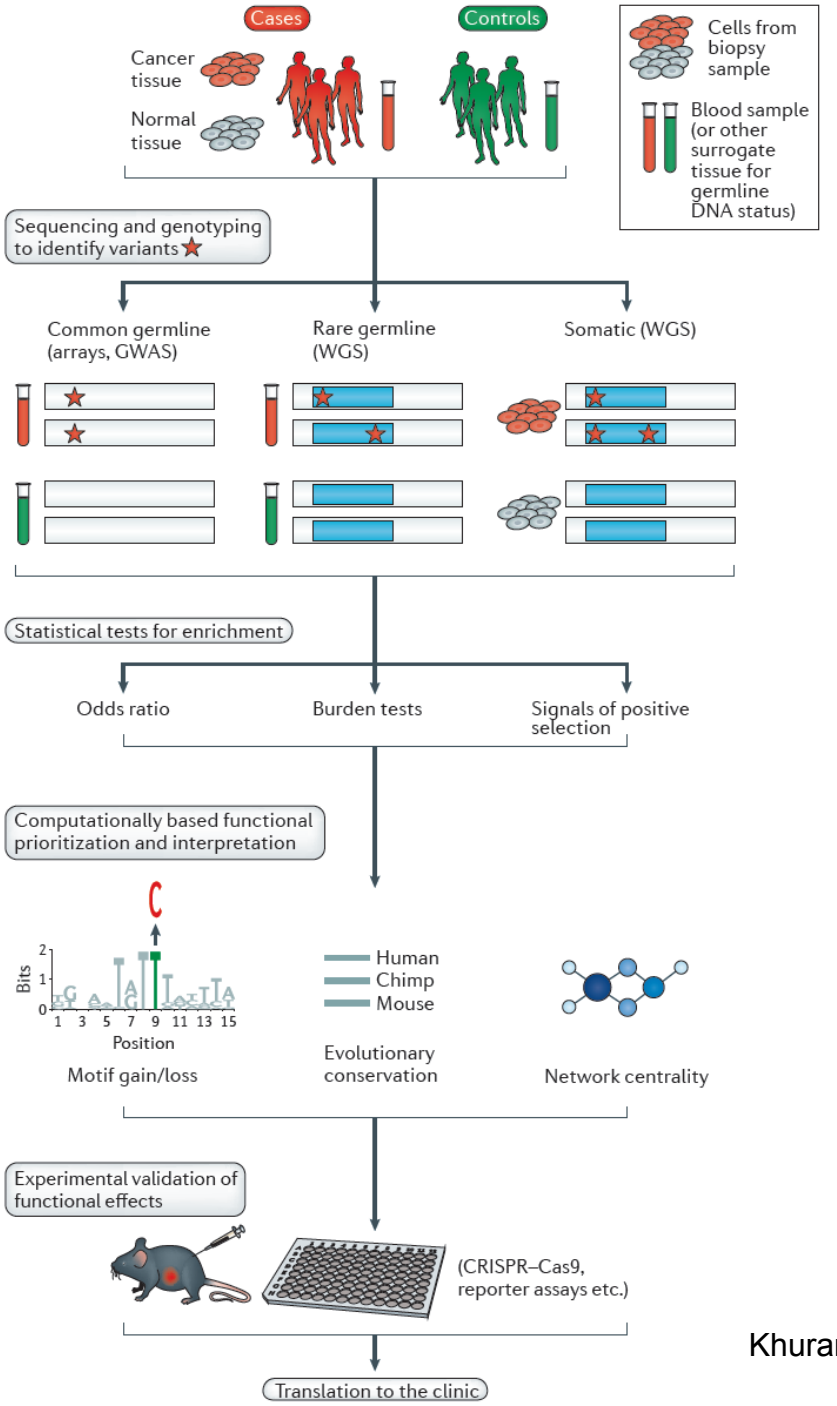
- Driver mutations: Confer selective advantage to tumour cells
- Passenger mutations: Do not confer selective advantage to tumour cells

- Cancer elements: Genomic elements with driver mutations



Yates and Campbell et al, Nat Rev Genet 2012

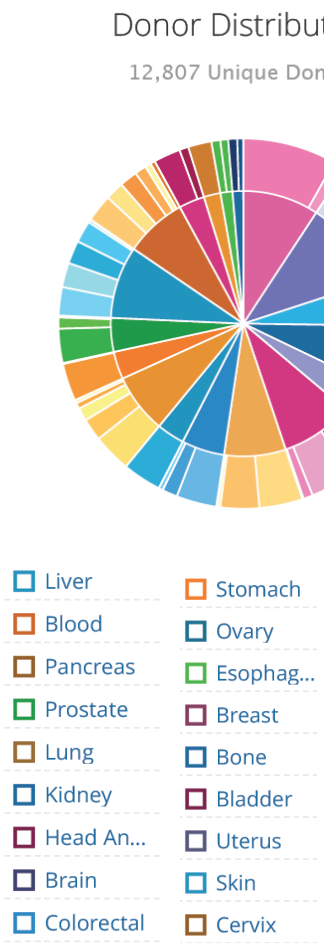
Identifying mutations associated with cancer

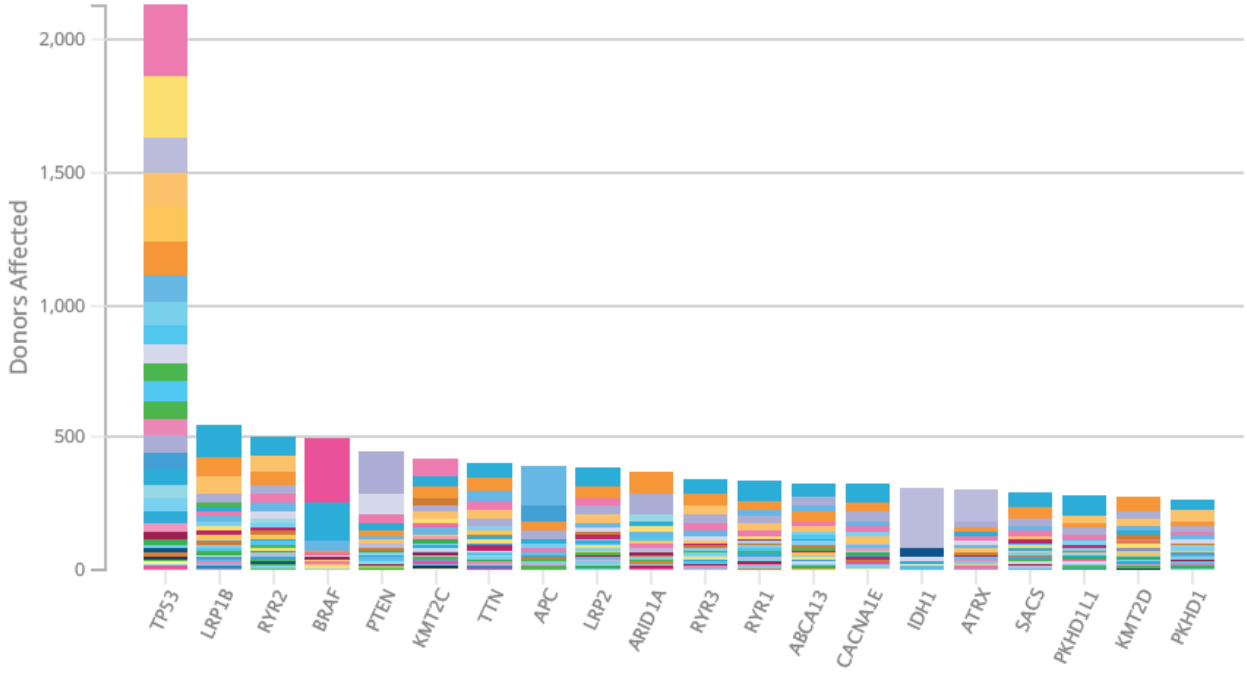


Khurana et al, *Nature Rev Genet*, 2016

Signals of positive selection

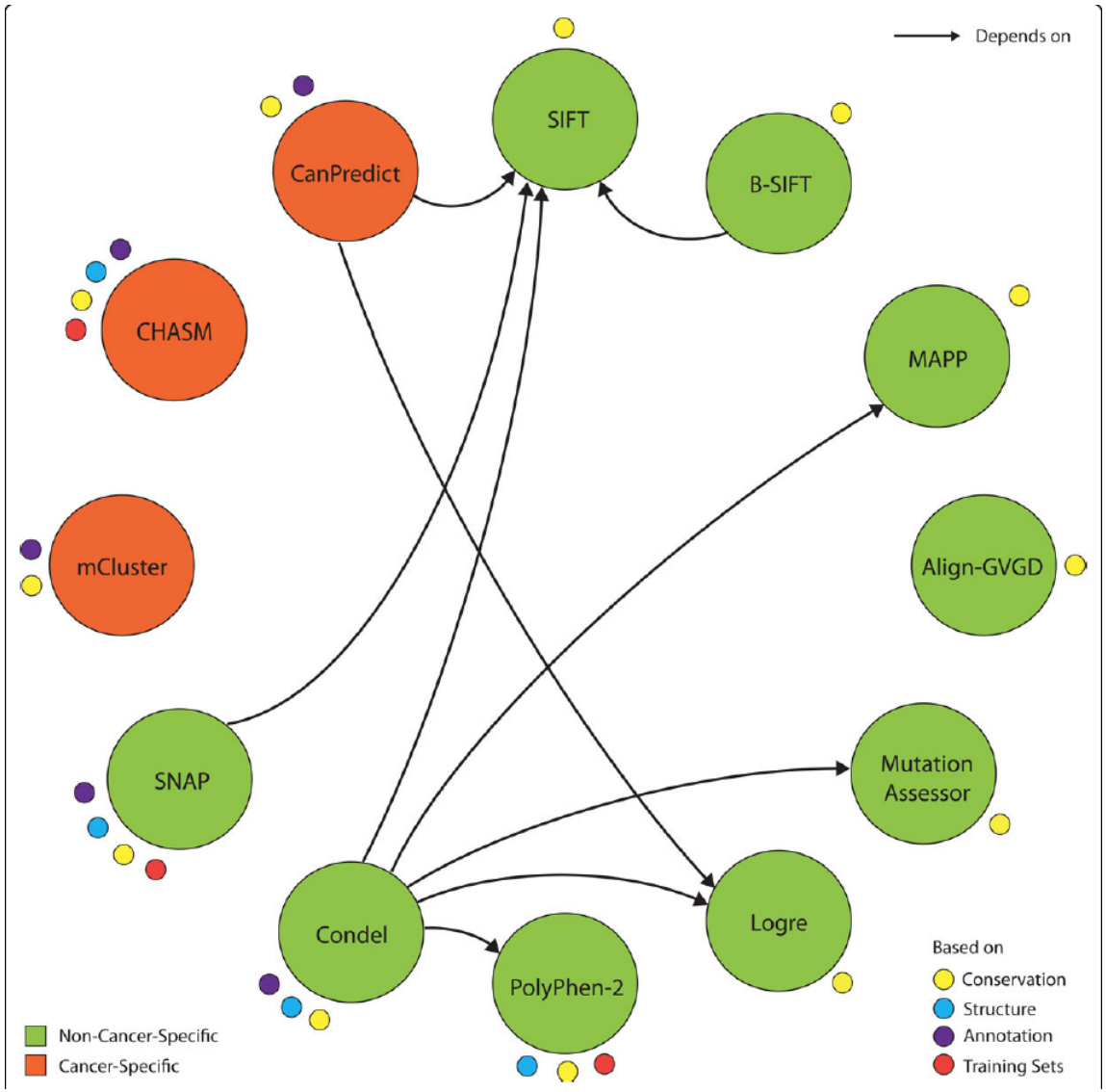
Mutation frequency of cancer genes





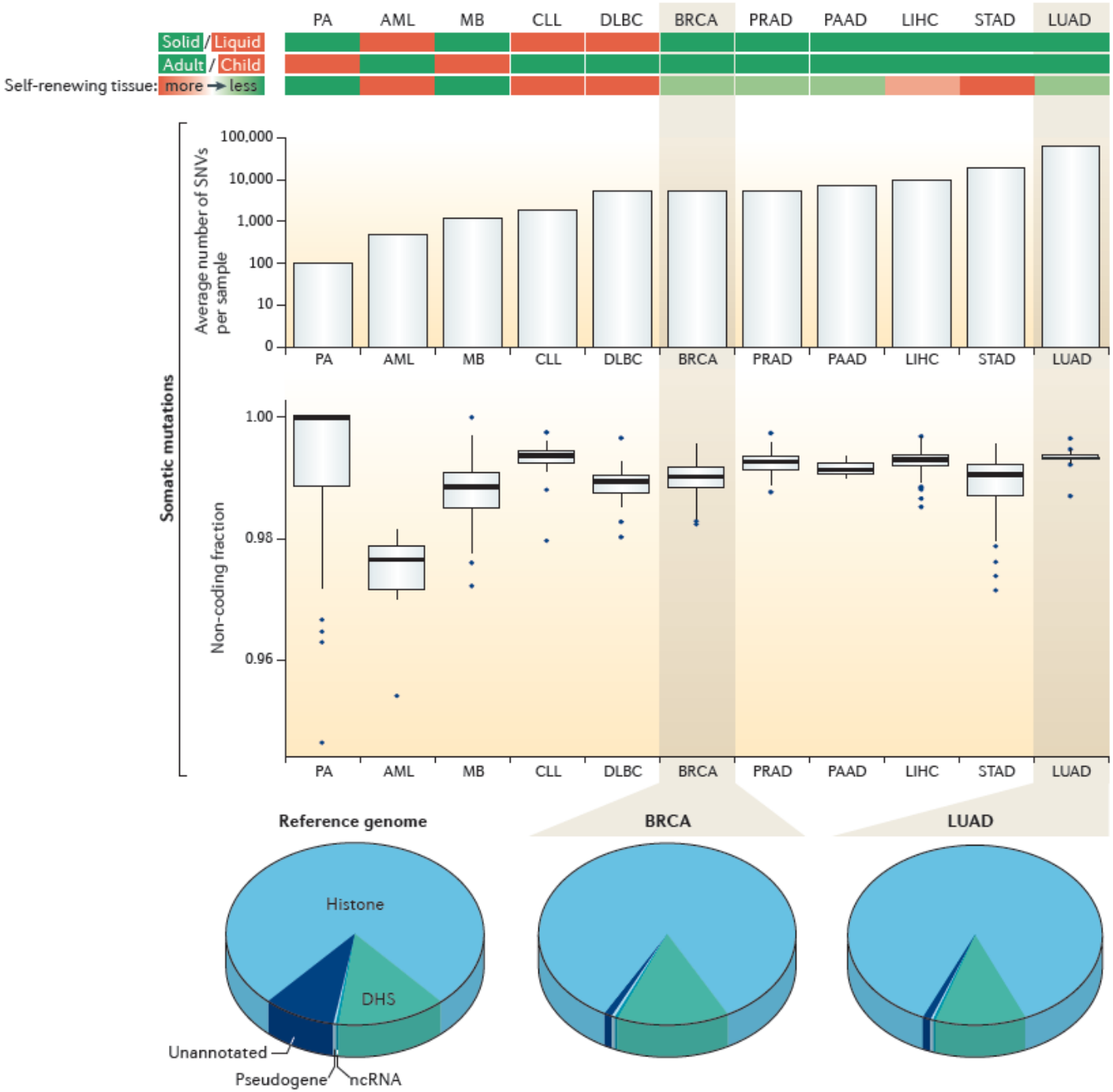
Source: International Cancer Genome Consortium (dcc.icgc.org)

Computational methods to predict functional impact of missense mutations



Gnad et al, BMC Genomics, 2013

Most variants are in noncoding regions

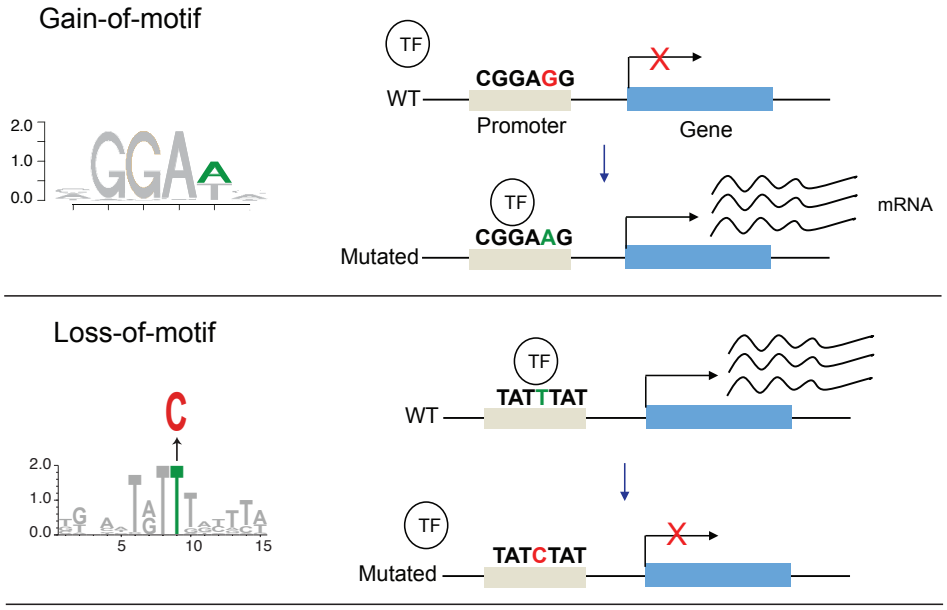


MB: medulloblastoma
DLBC: B cell lymphoma
STAD: gastric
BRCA: breast
PAAD: pancreatic
PRAD: prostate
LIHC: liver
PA: pilocytic Astrocytoma
LUAD: Lung adenocarcinoma

Khurana et al, *Nature Rev Genet*, 2016

Noncoding mutations can be significant drivers

Transcription factor (TF) binding disruption



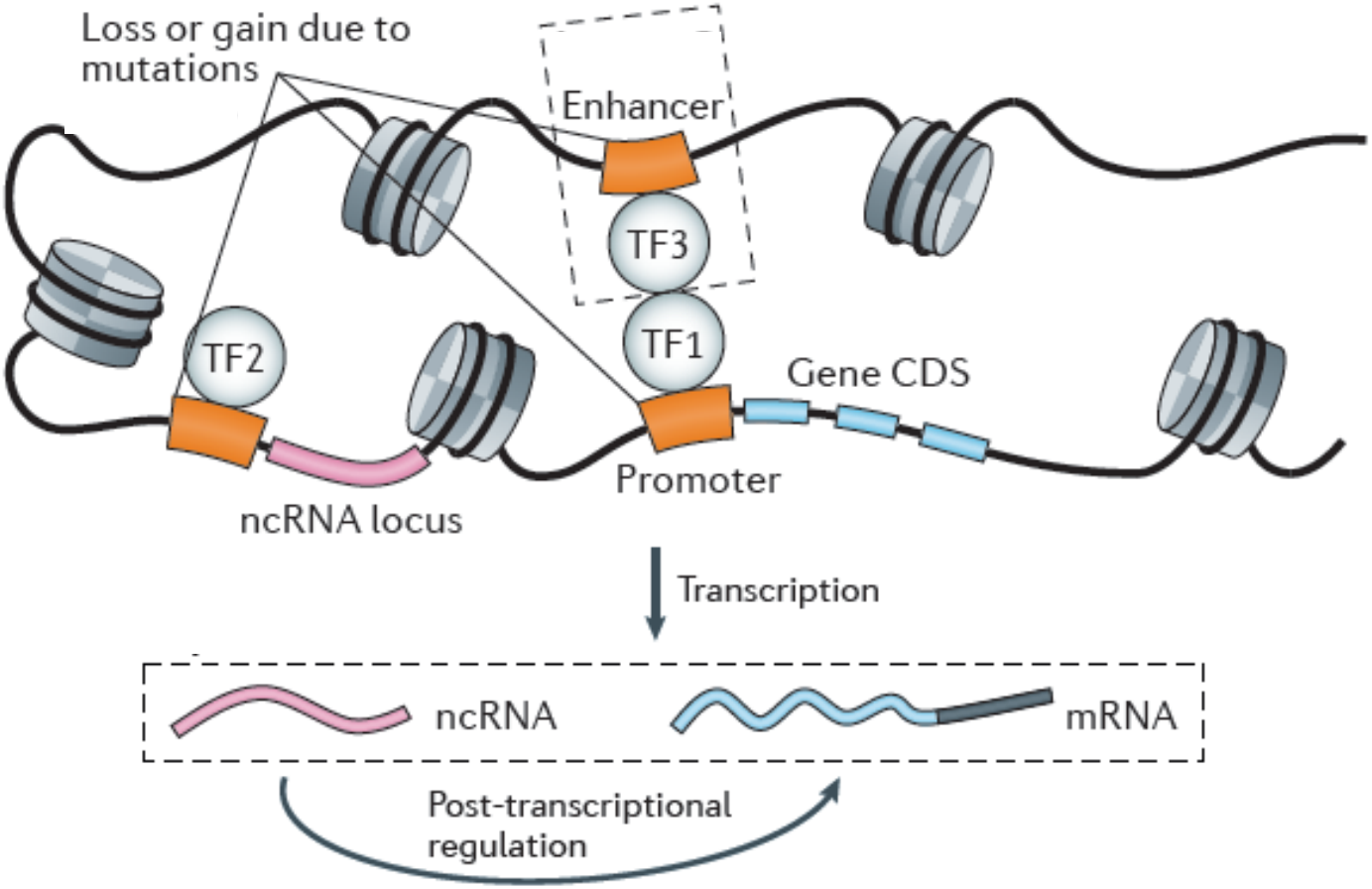
TERT promoter mutated in many different cancer types

Tumor type*	No. tumors	No. tumors mutated (%)
Chondrosarcoma	2	1 (50)
Dysembryoplastic neuroepithelial tumor	3	1 (33.3)
Endometrial cancer	19	2 (10.5)
Ependymoma	36	1 (2.7)
Fibrosarcoma	3	1 (33.3)
Glioma†	223	114 (51.1)
Hepatocellular carcinoma	61	27 (44.2)
Medulloblastoma	91	19 (20.8)
Myxofibrosarcoma	10	1 (10.0)
Myxoid liposarcoma	24	19 (79.1)
Neuroblastoma	22	2 (9)
Osteosarcoma	23	1 (4.3)
Ovarian, clear cell carcinoma	12	2 (16.6)
Ovarian, low grade serous	8	1 (12.5)
Solitary fibrous tumor (SFT)	10	2 (20.0)
Squamous cell carcinoma of head and neck	70	12 (17.1)
Squamous cell carcinoma of the cervix	22	1 (4.5)
Squamous cell carcinoma of the skin	5	1 (20)
Urothelial carcinoma of bladder	21	14 (66.6)
Urothelial carcinoma of upper urinary epithelium	19	9 (47.3)

- MYB motif created & drives *TAL1* overexpression in T-ALL (Mansour et al, *Science*, 2014)

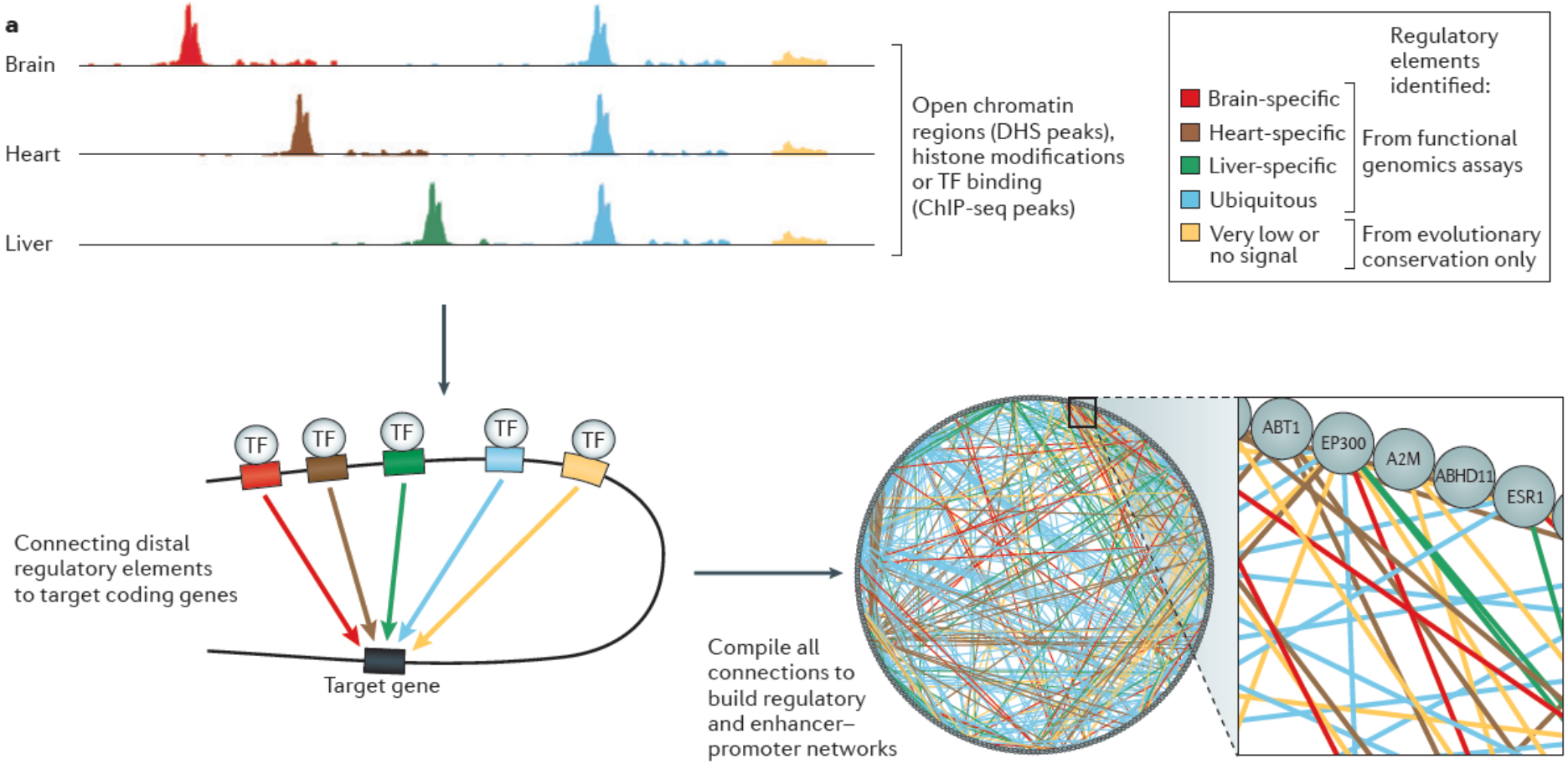
Killela et al, *PNAS*, 2013
Horn et al, *Science*, 2013
Huang et al, *Science*, 2013

Noncoding elements in the genome

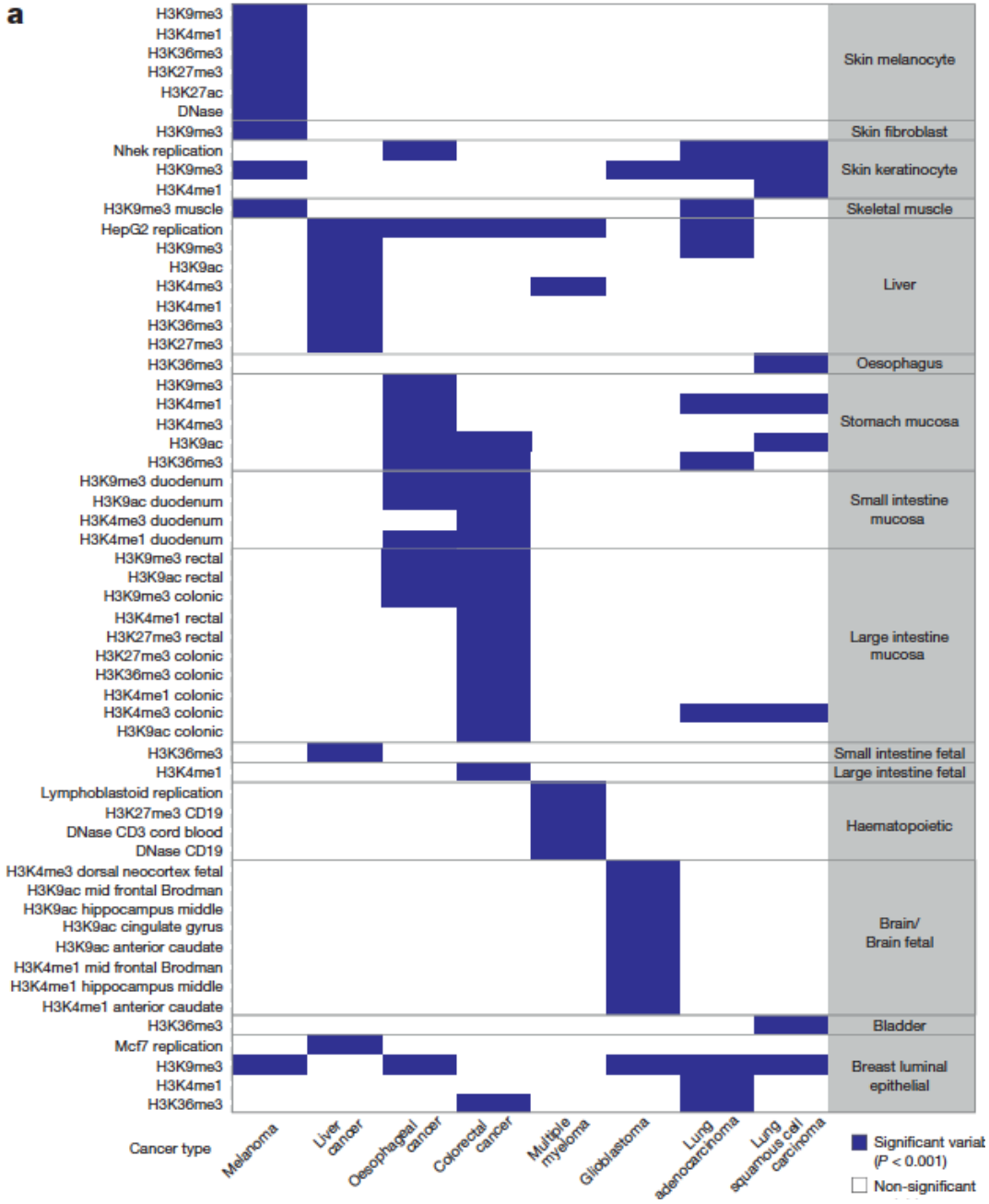


Khurana et al, *Nature Rev Genet*, 2016

Noncoding variants act via tissue-specific regulatory networks



Khurana et al, *Nature Rev Genet*, 2016

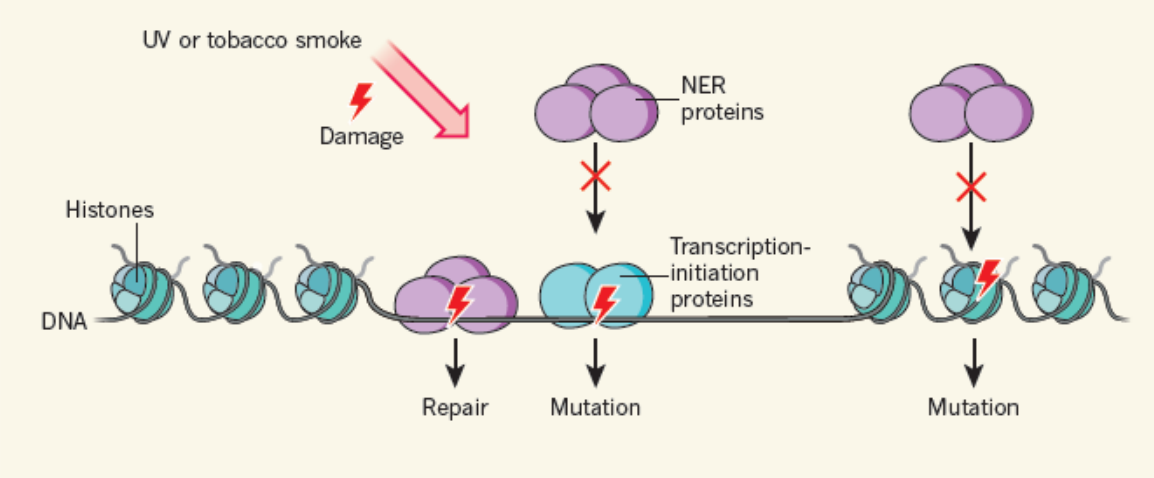


Need to account for heterogeneity of mutation rate in cancer cells when identifying drivers

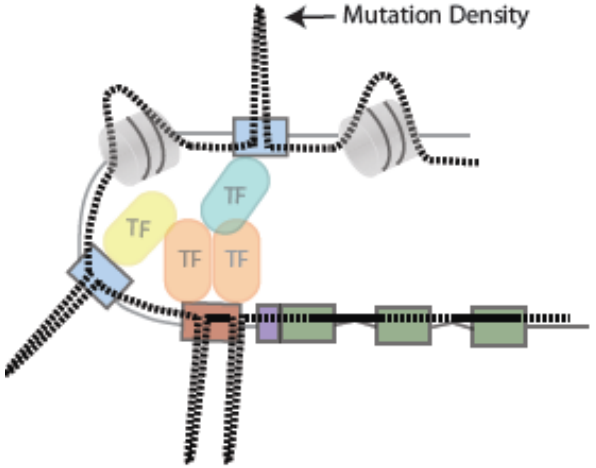
- Histone modification marks
- DNase I hypersensitive sites
- Replication timing

Polak *et al. Nature* **518**, 360-364 (2015)

Co-variates of mutation rates: Increased mutation density at TF binding sites in melanoma and lung cancer

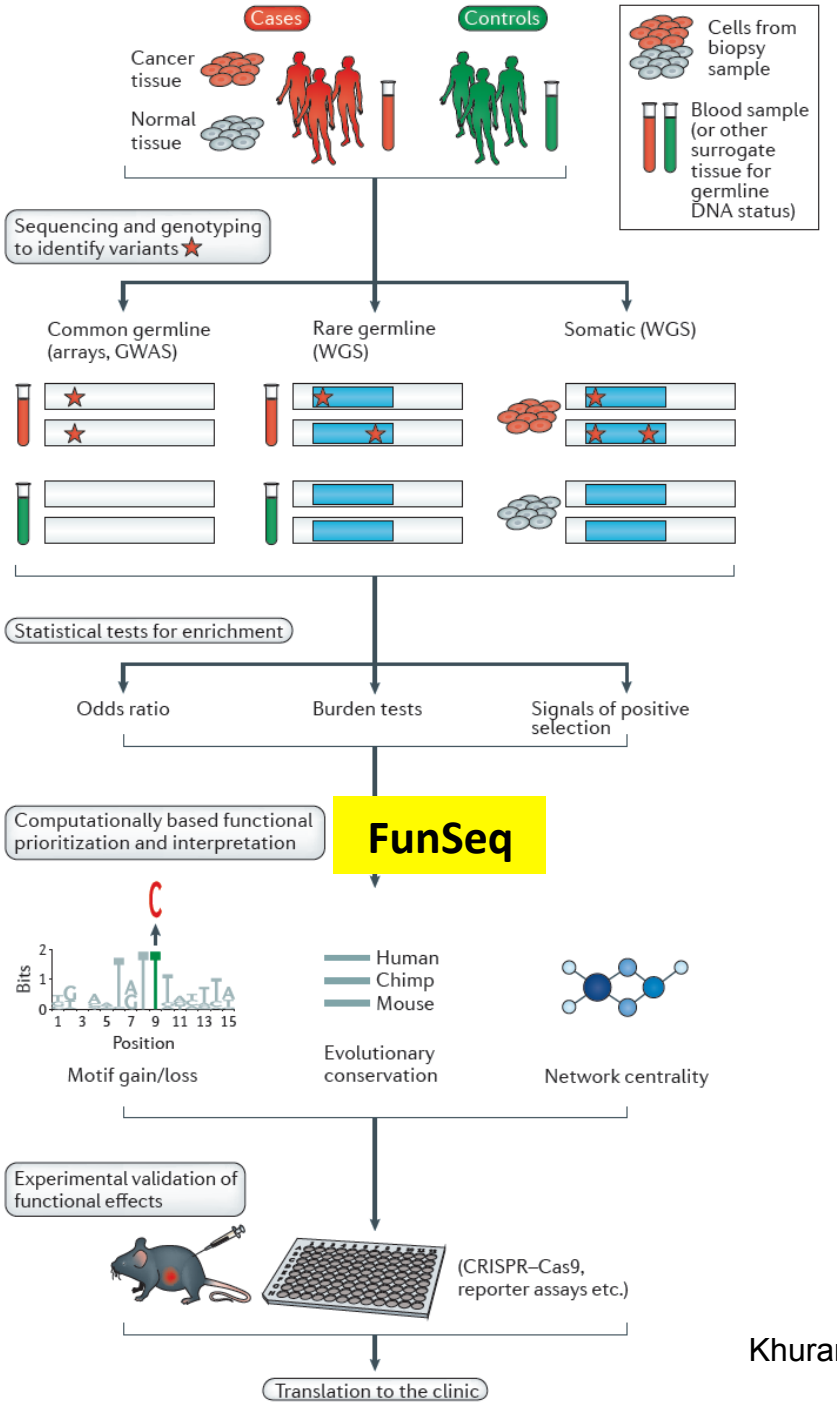


Perera et al, *Nature*, 2016
Sabarinathan et al, *Nature*, 2016
Khurana, *Nature News & Views*, 2016



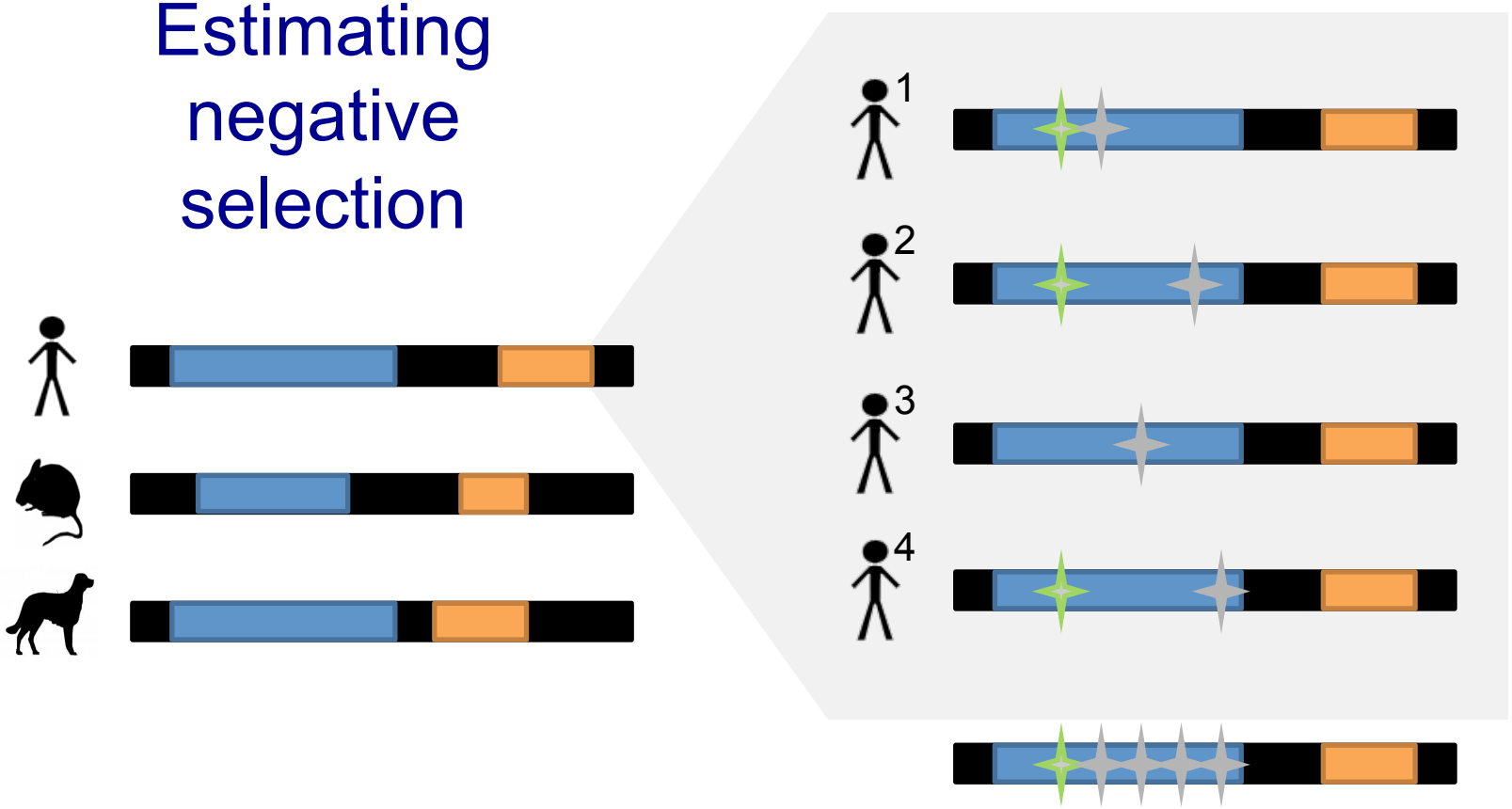
Cuykendall et al, *COISB*, 2017

Identifying mutations associated with cancer



Khurana et al, *Nature Rev Genet*, 2016

Estimating negative selection



Evolutionary conservation

- Typically defined by comparison across species

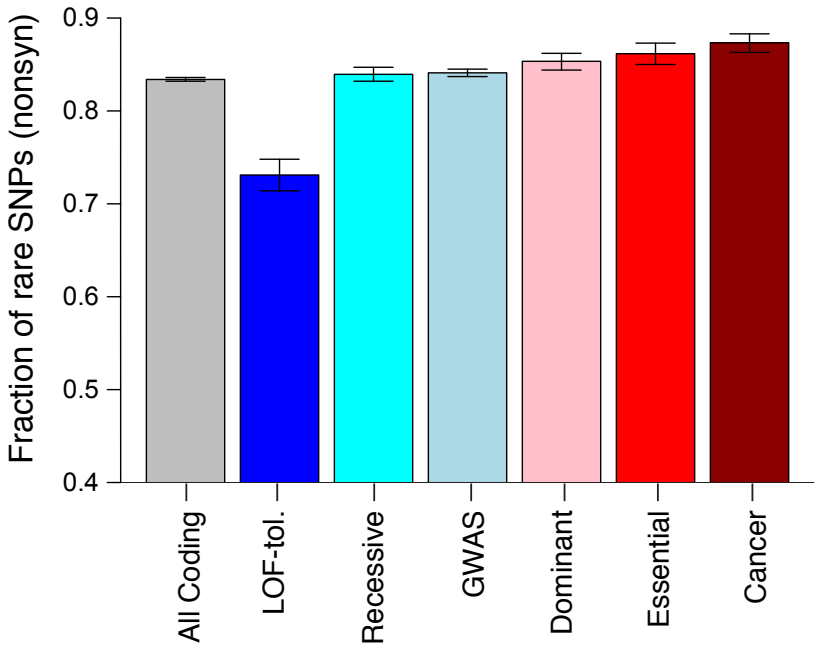
Conservation among humans

- Depletion of common variants/**<u>Enrichment</u>** of rare variants

Common variant Rare variants

Fraction of rare variants = (Num of rare variants/ Total num of variants)

Enrichment of rare SNPs as a metric for negative selection



(rare=derived allele freq < 0.5%)

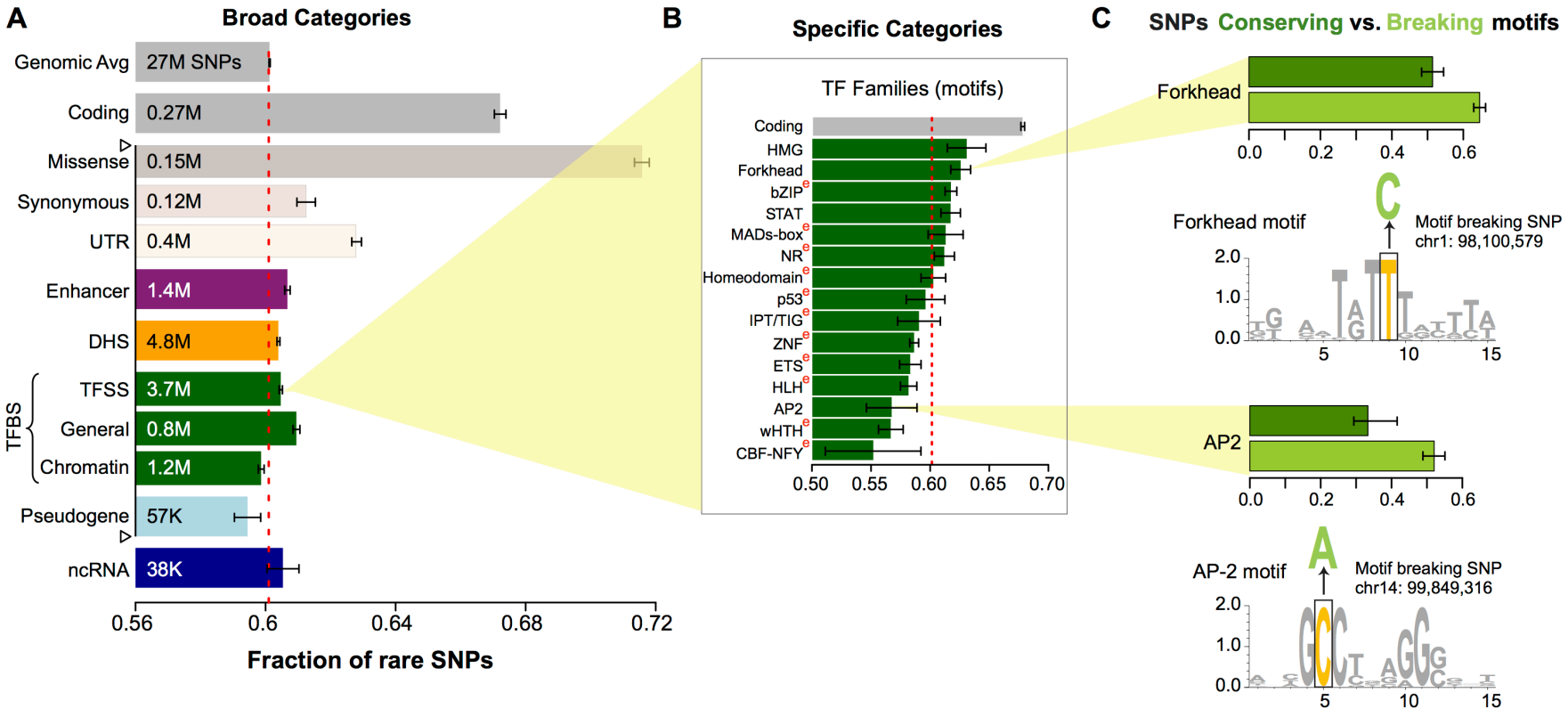
- Depletion of common polymorphisms in regions under selection

  Negative selection restricts the allele frequency of deleterious mutations.
- Results for coding genes consistent with known phenotypic impacts
- Other metrics for selection
  - Evolutionary conservation (e.g. GERP)
  - SNP density (confounded by mutation rate)

LOF-tol (Loss-of-function tolerant): least negative selection
Cancer: most selection

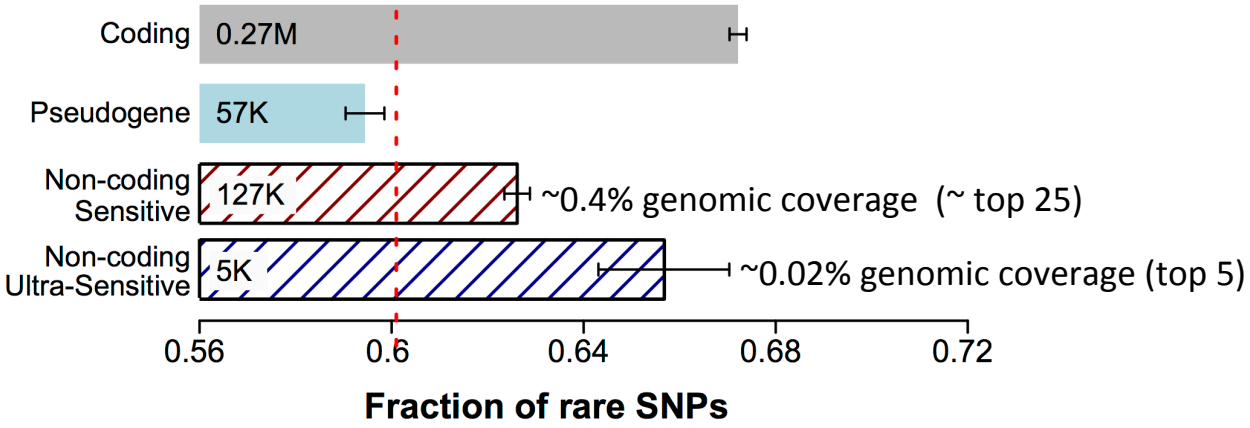
Khurana et al., *Science*, 2013

Organism-level negative selection in noncoding elements

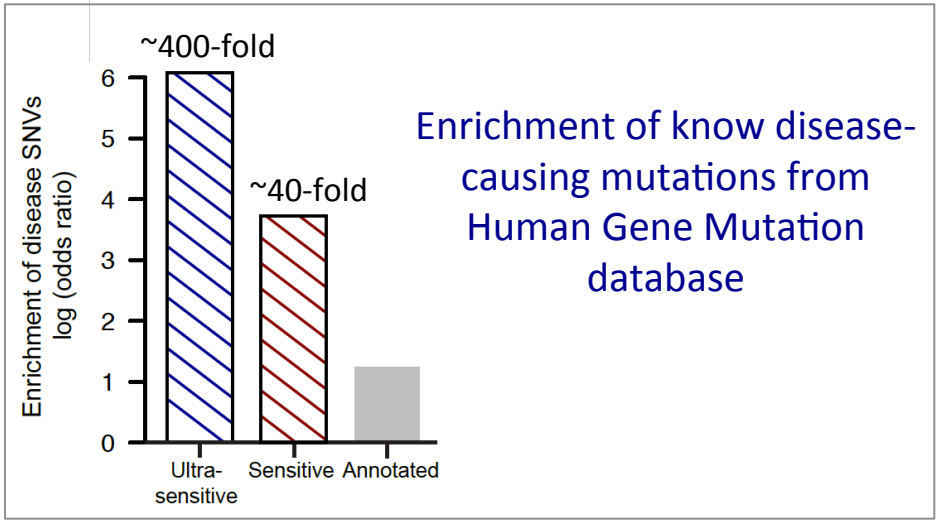


Khurana et al., *Science*, 2013

Which noncoding categories are under very strong “coding-like” selection ?

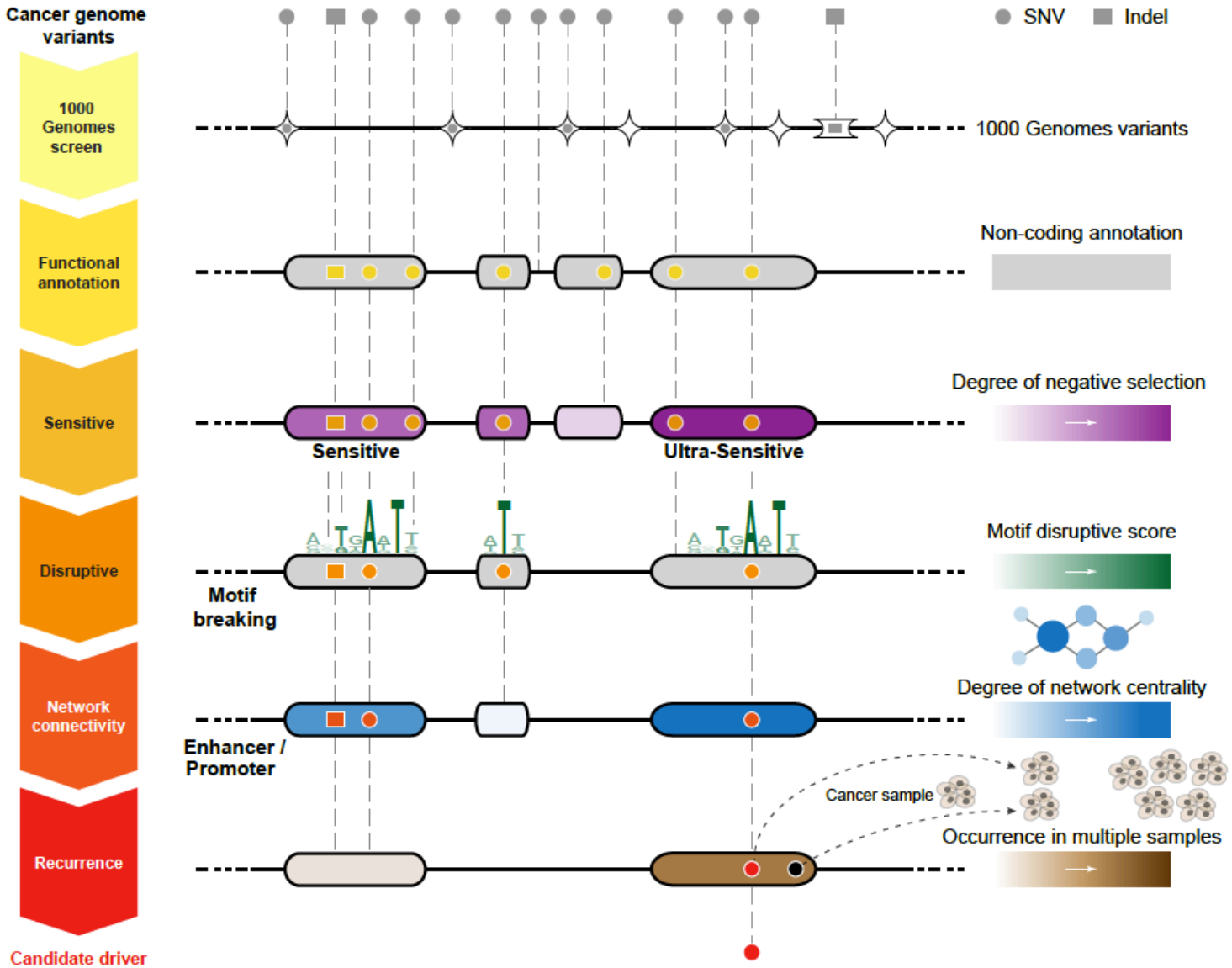


- Top categories among ranked 102 categories
- Binding peaks of some general TFs (eg *FAM48A*)
- Core motifs of some TF families (eg *JUN*, *GATA*)
- DHS sites in spinal cord and connective tissue



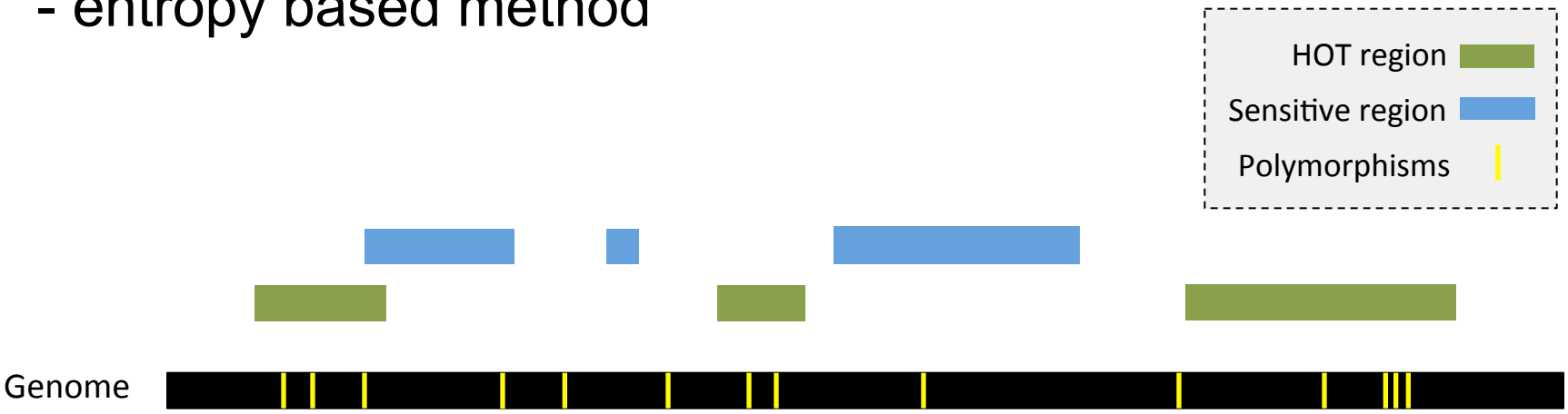
Khurana et al., *Science*, 2013

Identification of noncoding mutations with high impact: FunSeq



FunSeq2: Feature weight

- Weighted with mutation patterns in natural polymorphisms (features frequently observed weighed less)
- entropy based method



Feature weight: $w_d = 1 + p_d log_2 p_d + (1 - p_d) log_2 (1 - p_d)$

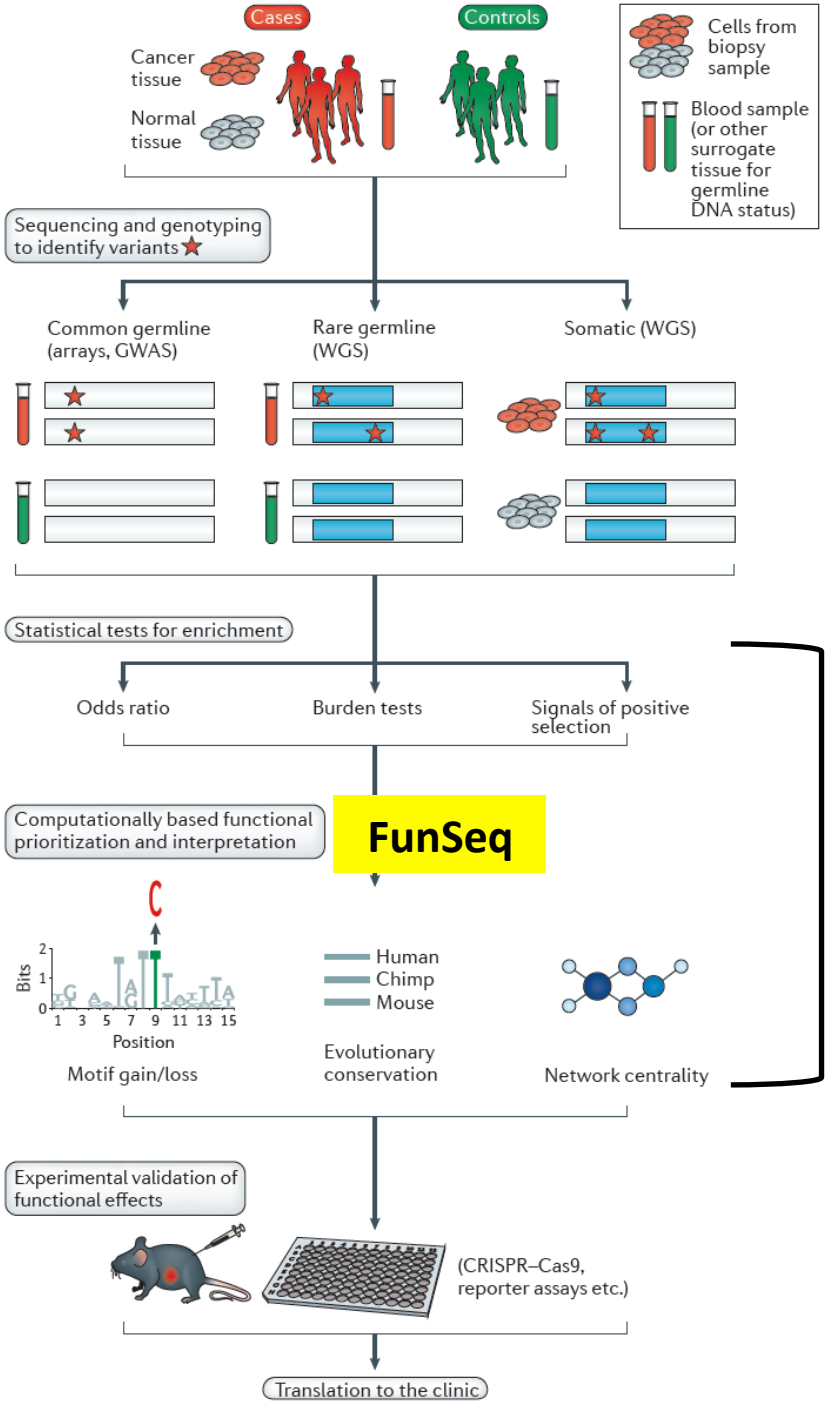
$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

For a variant: $Score = \sum w_d \quad of\ observed\ features$

https://github.com/khuranalab/FunSeq_PCAWG

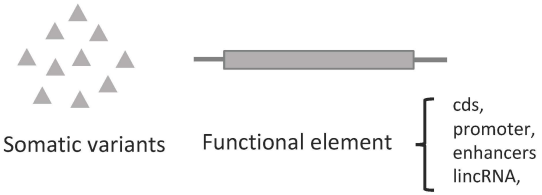
Fu et al., *Genome Biology*, 2014

Identifying noncoding variants associated with cancer



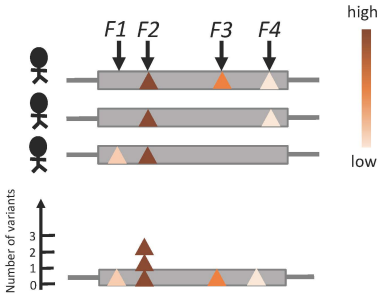
CNCDriver for detecting driver coding & noncoding elements

(1) Map somatic variants onto functional elements



(2) Calculate compositeDriver score (S)

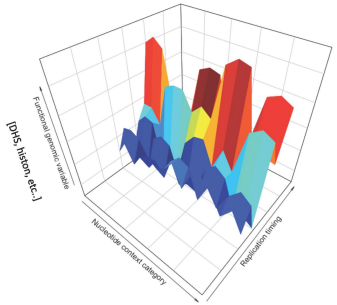
Fi **=** functional impact score

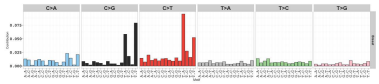


$$S_{r,k} = \sum_{i=0}^{n} W_i \times F_i$$

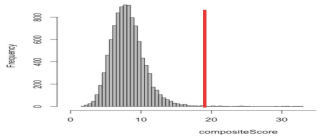
S **=** compositeDriver score
r **=** functional element
n **=** number of variants in r
W_i **=** number of samples with variant *i*

(3) Null model

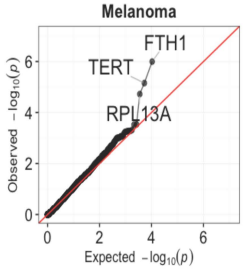




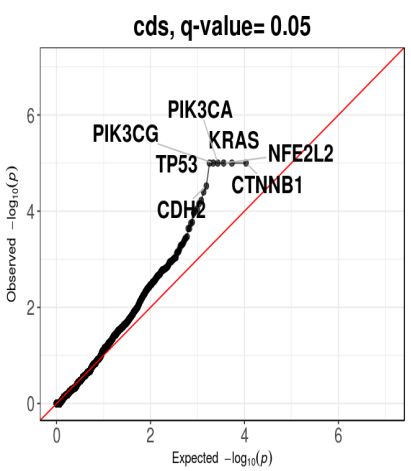
(4) Calculate p-value and multiple hypothesis correction

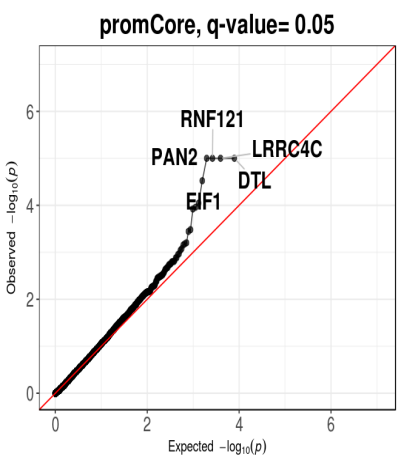


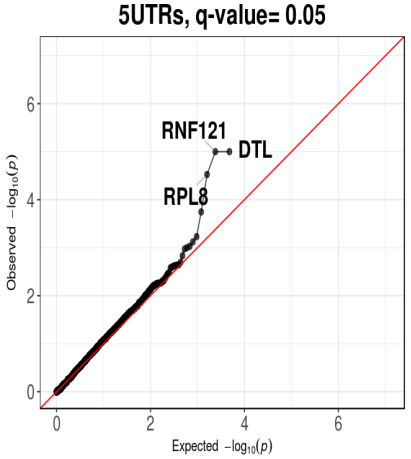
(5) QQ-plot for each functional element

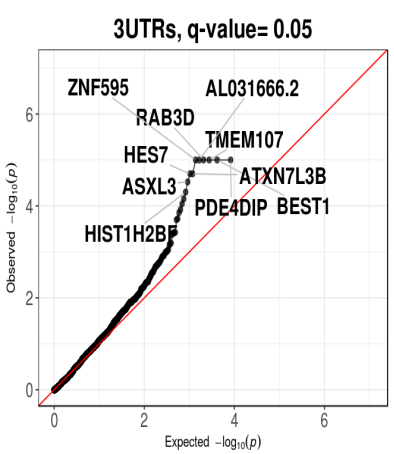


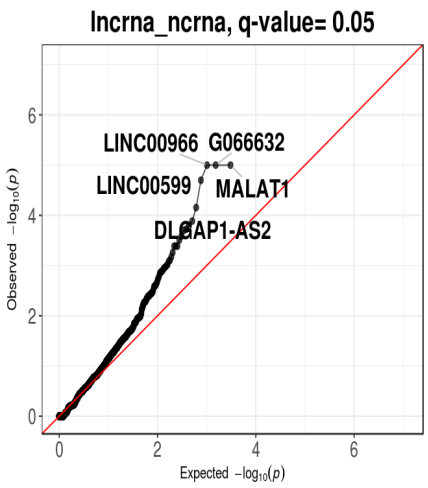
CNCDriver results in lung cancer (n=84)

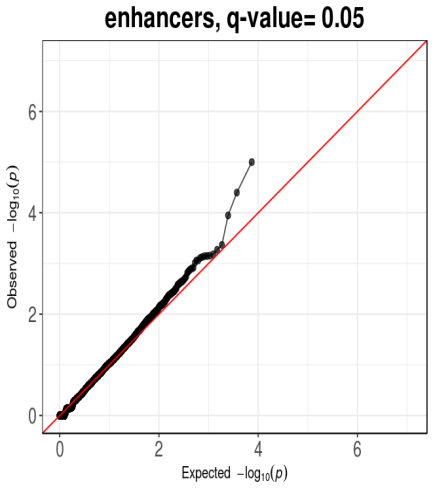








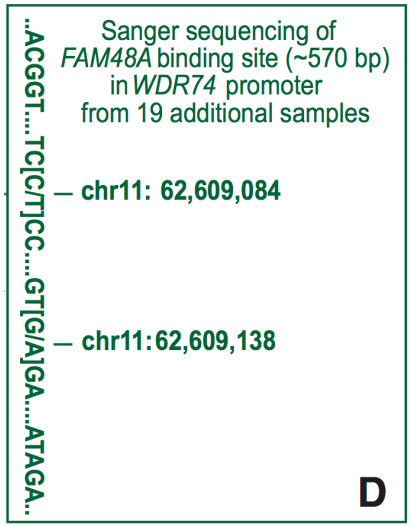




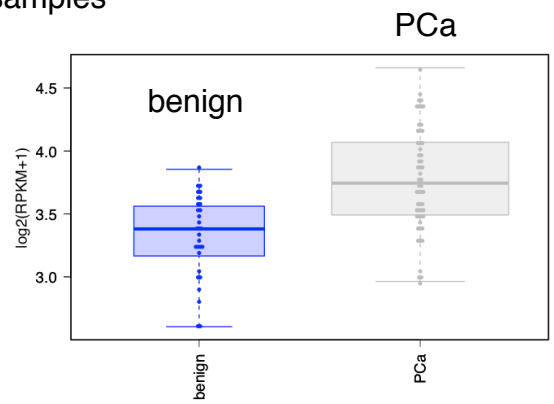
Functional validation of candidates in prostate cancer

WDR74 promoter

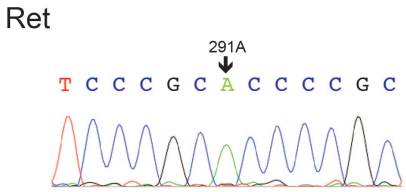
- Sanger sequencing in 19 additional samples confirms the recurrence

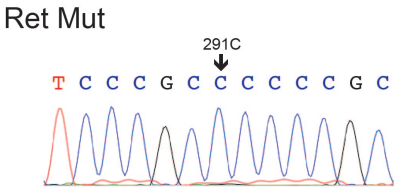


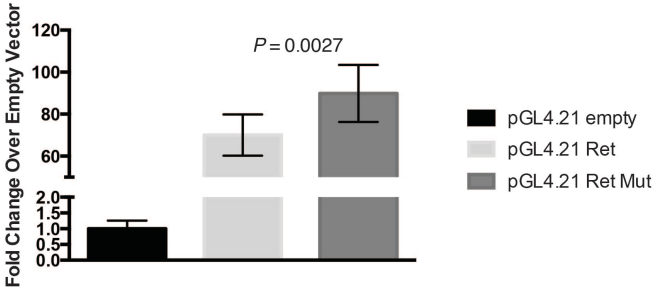
- *WDR74* shows increased expression in tumor samples



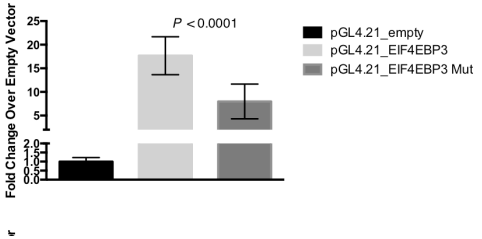
RET promoter
Increased activity





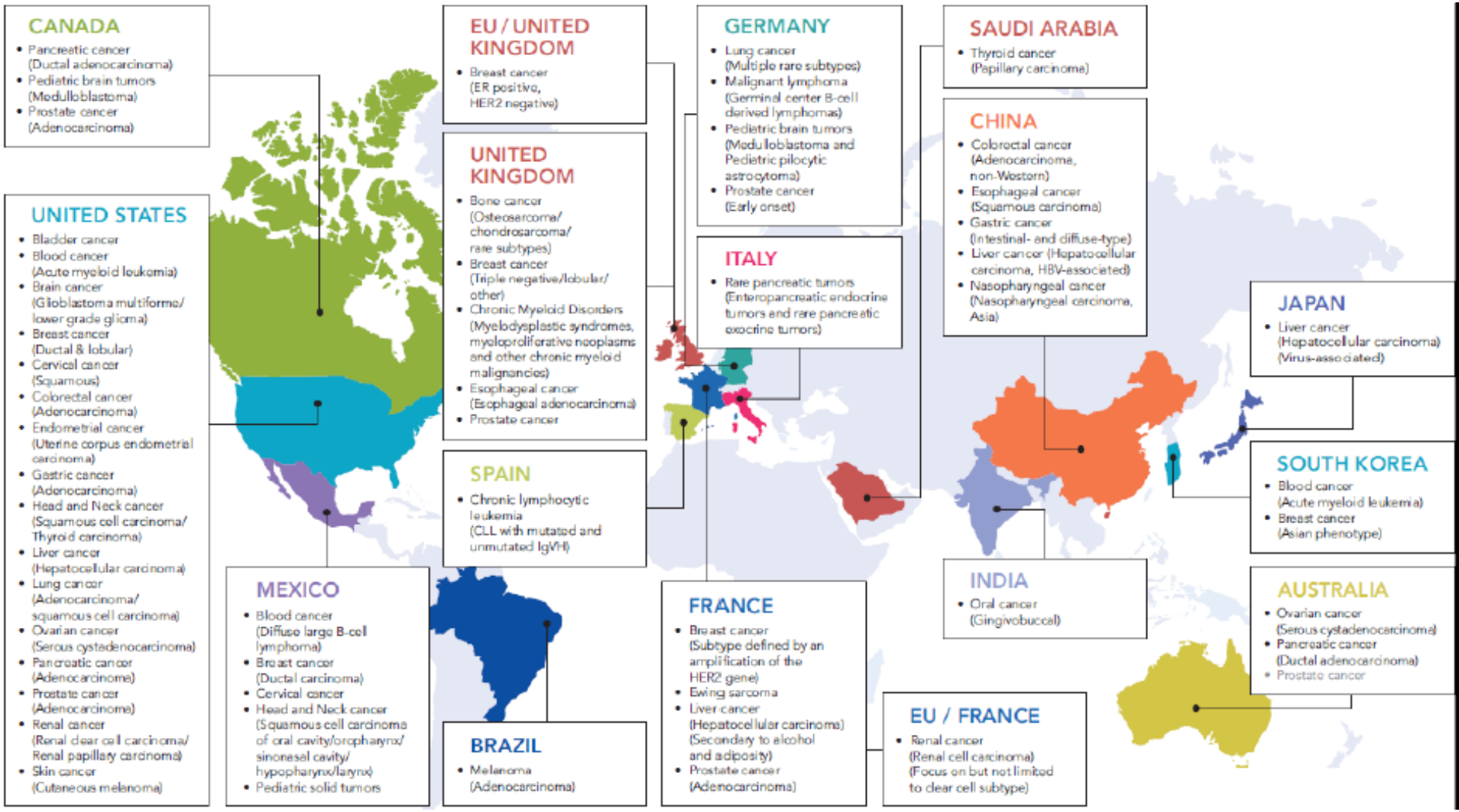


EIF4EBP3 promoter
Reduced activity



International Cancer Genome Consortium & The Cancer Genome Atlas

PCAWG PanCancer Analysis OF WHOLE GENOMES



~2800 WGS (tumor & normal), ~1500 RNA-Seq, ~1400 methylation

Acknowledgements

~40 Institutes
~550 participants

Functional Interpretation Group

~50 participants

Yale

Yao Fu (now at Bina), Xinmeng Mu (now at Broad), Jieming Chen, Lucas Lochovsky, Arif Harmanci, Alexej Abyzov, Suganthi Balasubramanian, Cristina Sisu, Declan Clarke, Mike Wilson, Yong Kong, **Mark Gerstein**

Sanger

Vincenza Colonna, Yuan Chen, Yali Xue, **Chris Tyler-Smith**

Cornell

Steven Lipkin, Jishnu Das, Robert Fragoza, Xiaomu Wei, Haiyuan Yu

Andrea Sboner, Dimple Chakravarty, Naoki Kitabayashi, Vaja Liluashvili, Zeynep H. Gümüş, Kellie Cotter, Mark A. Rubin

U of Michigan

Hyun Min Kang

U of Geneva

Tuuli Lappalainen (NYGC), Emmanouil T. Dermitzakis

Baylor

Daniel Challis, Uday Evani, Donna Muzny, Fuli Yu, Richard Gibbs

EBI

Kathryn Beal, Laura Clarke, Fiona Cunningham, Paul Flicek, Javier Herrero, Graham R. S. Ritchie

Boston College

Erik Garrison, Gabor Marth

Mass Gen Hospital

Kasper Lage, Daniel G. MacArthur, Tune H. Pers

Rutgers

Jeffrey A. Rosenfeld

Acknowledgements

Khurana lab
Eric Minwei Liu
Priyanka Dhingra
Alexander Fundichely
Tawny Cuykendall
Andre Forbes

Mark Rubin
Dimple Chakravarty
Kellie Cotter
David Rickman
Adeline Berger
Andrea Sboner
Deli Liu
Steve Lipkin

PCAWG (ICGC/TCGA) collaboration (~100)

Weill Cornell Medicine

NIH NATIONAL CANCER INSTITUTE

NIH National Human Genome Research Institute