Weill Cornell Medicine

Tools for analyzing cancer variation

Ekta Khurana, PhD

Assistant Professor

Meyer Cancer Center

Englander Institute for Precision Medicine

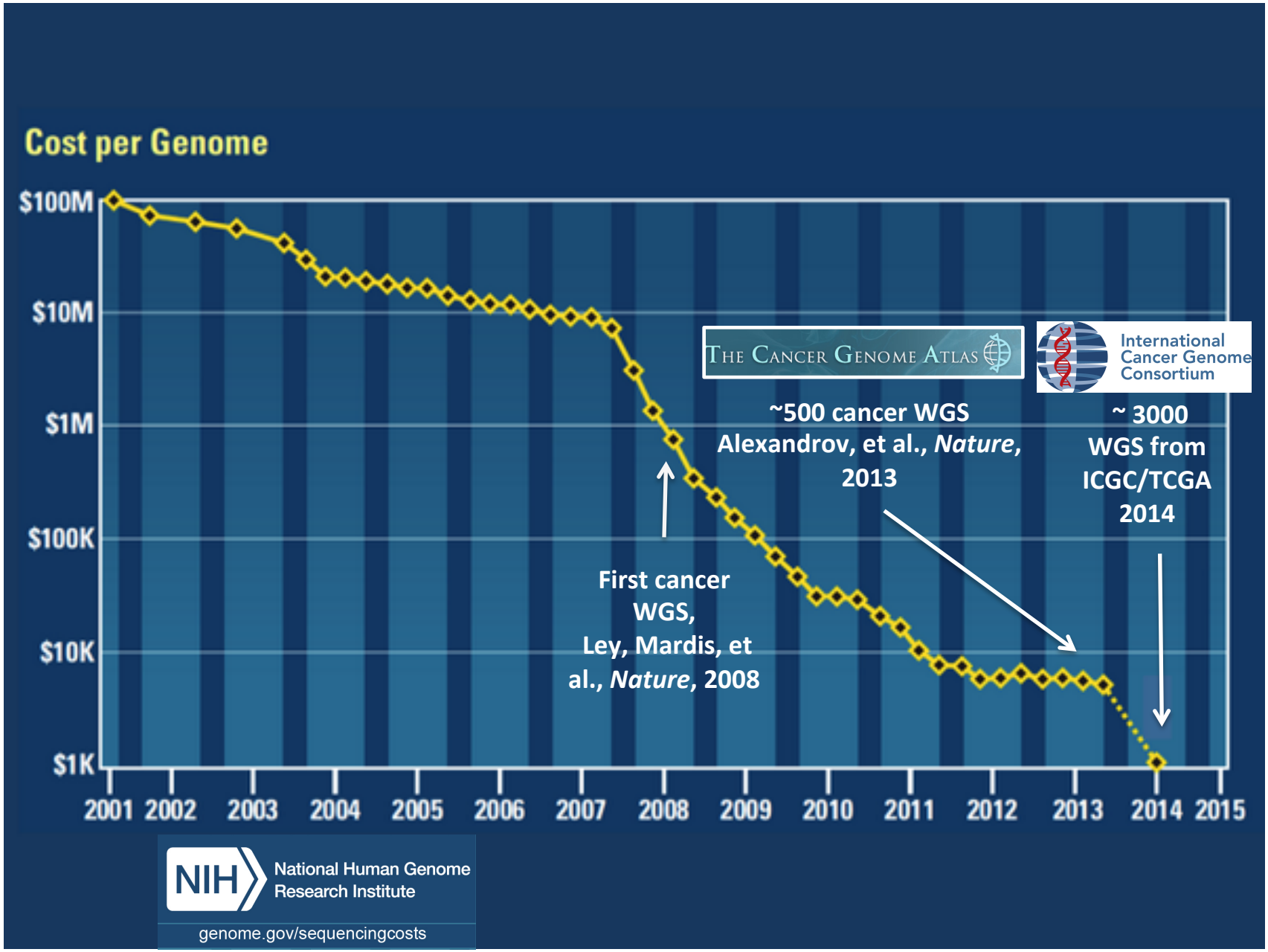
Institute for Computational Biomedicine

Department of Physiology and Biophysics

Weill Cornell Medicine, New York, NY

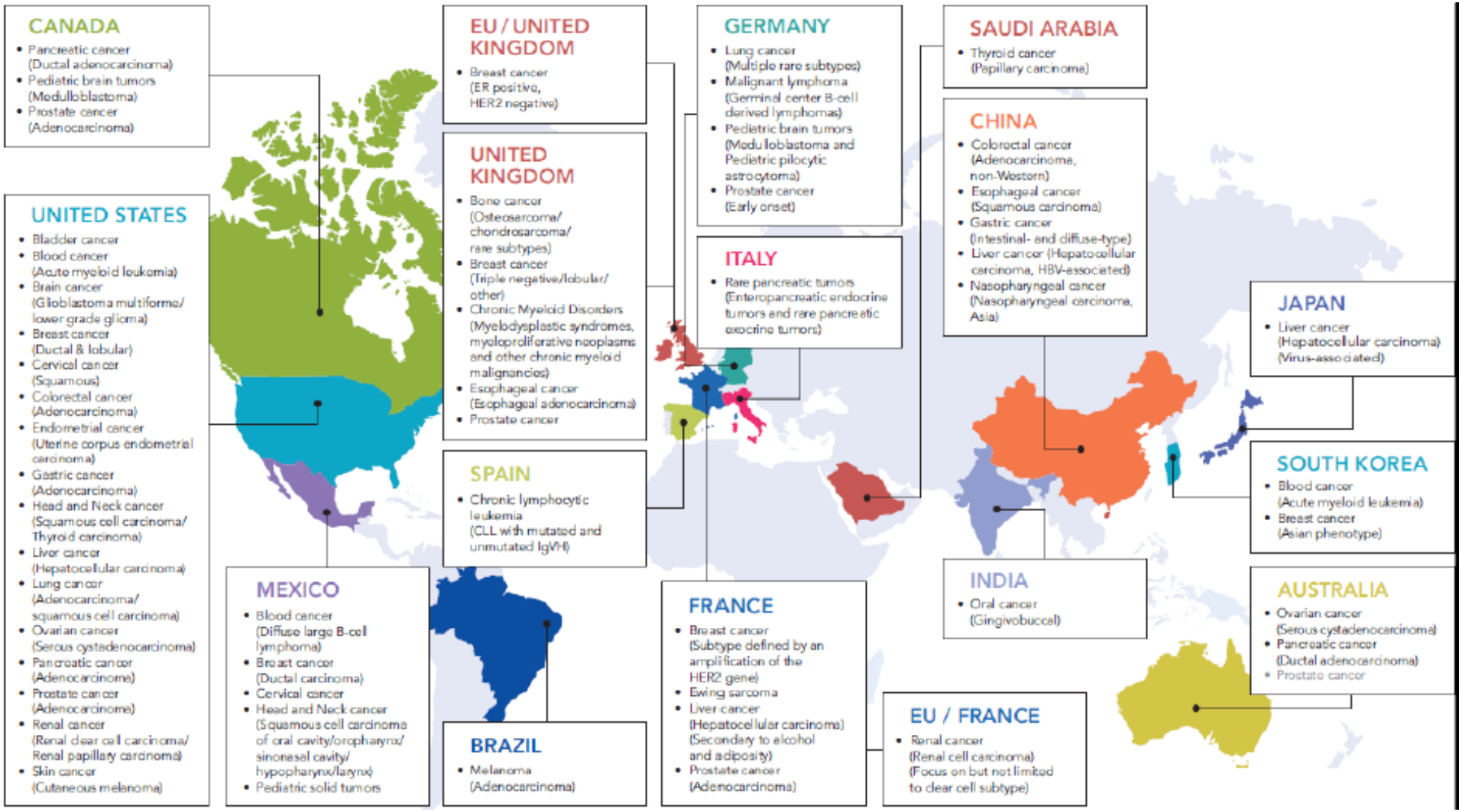
ekk2003@med.cornell.edu

 @ekta_khurana


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
THE CANCER GENOME ATLAS
International Cancer Genome Consortium
~500 cancer WGS
Alexandrov, et al., *Nature*, 2013
~ 3000 WGS from ICGC/TCGA 2014
First cancer WGS, Ley, Mardis, et al., *Nature*, 2008
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

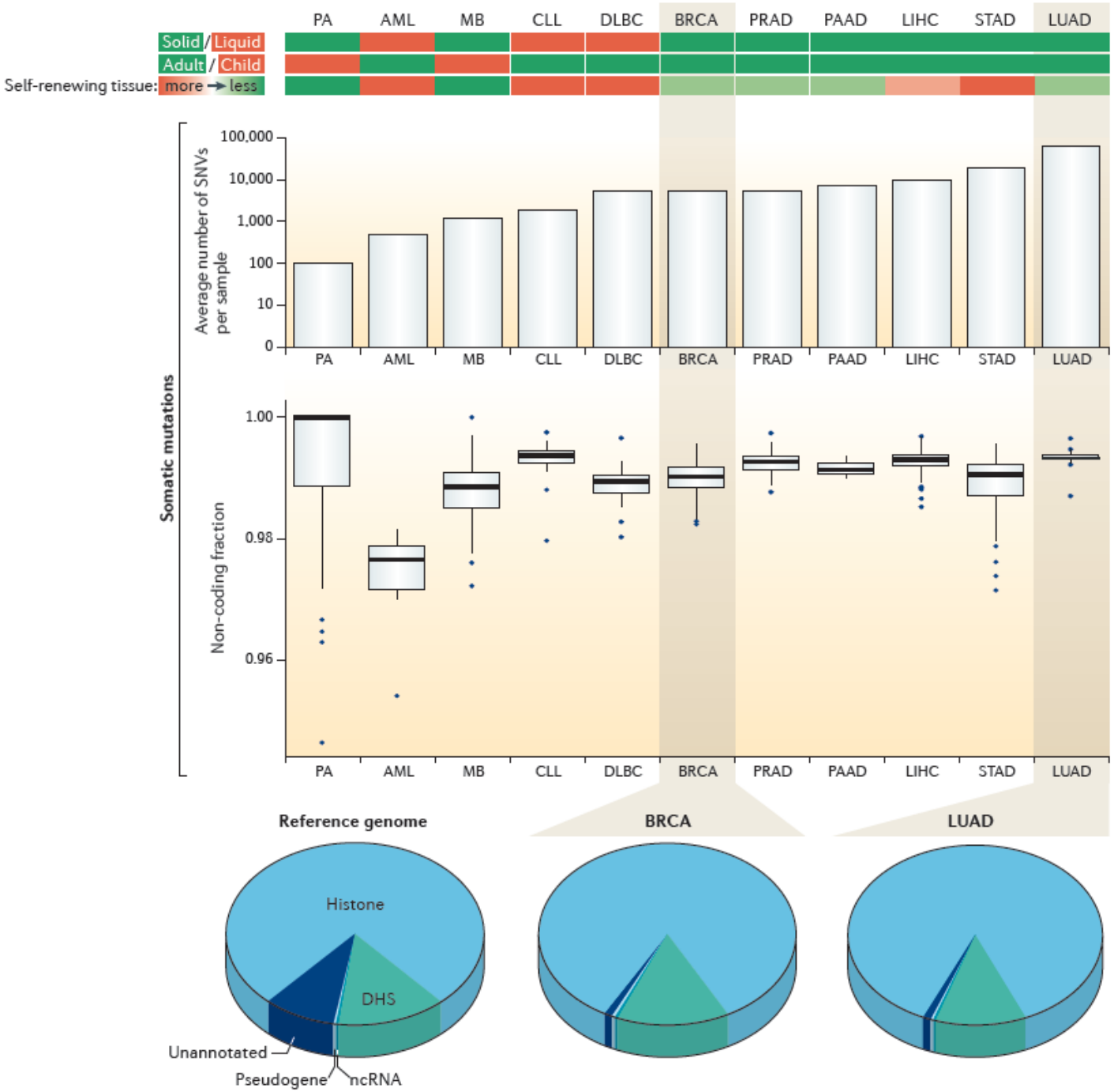
International Cancer Genome Consortium & The Cancer Genome Atlas

PCAWG PanCancer Analysis OF WHOLE GENOMES



~3000 WGS (tumor & normal), ~1600 RNA-Seq, ~1500 methylation

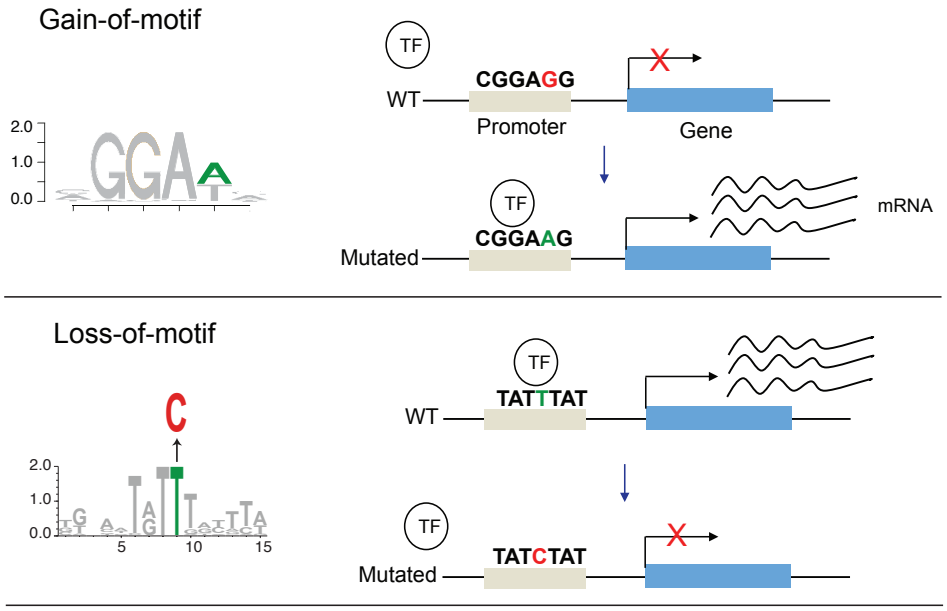
Most variants are in noncoding regions



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

Modes of action of noncoding variants:
transcription factor binding disruption



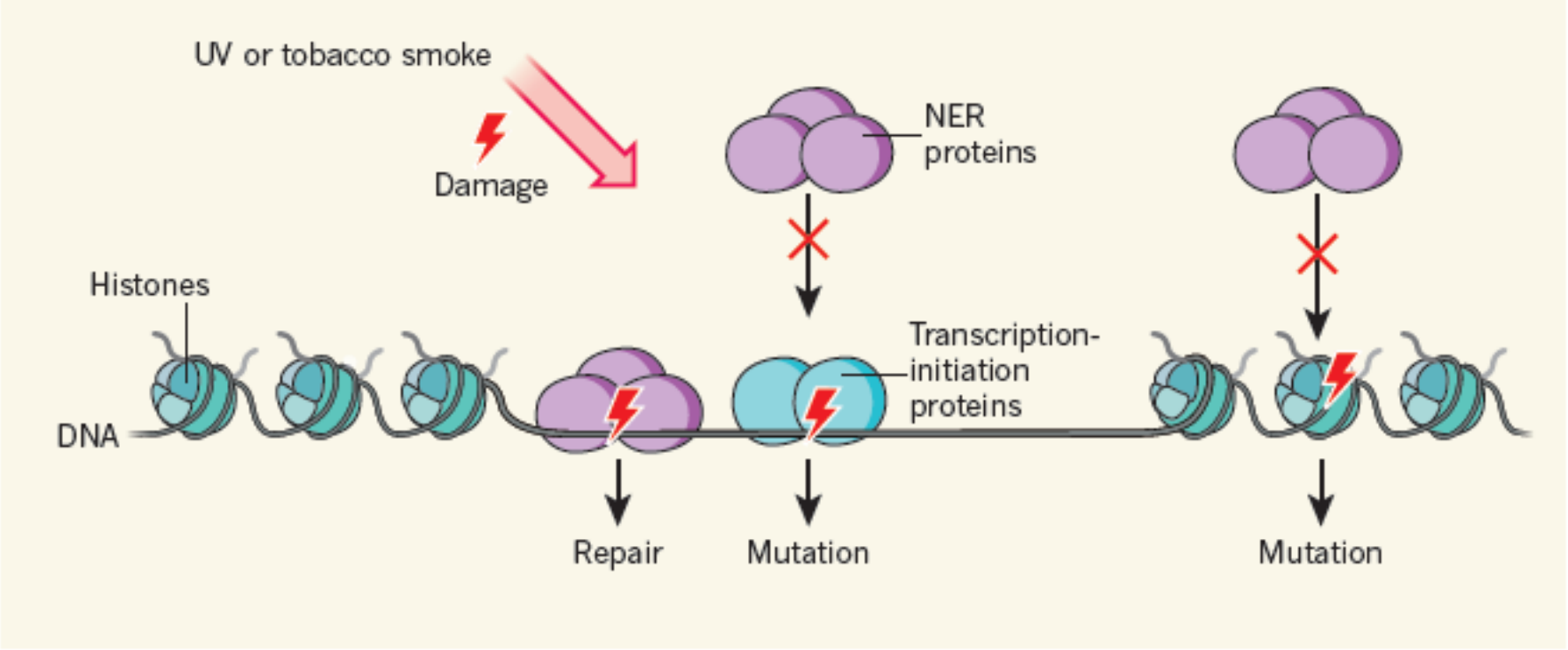
TERT promoter mutated in many different cancer types

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

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

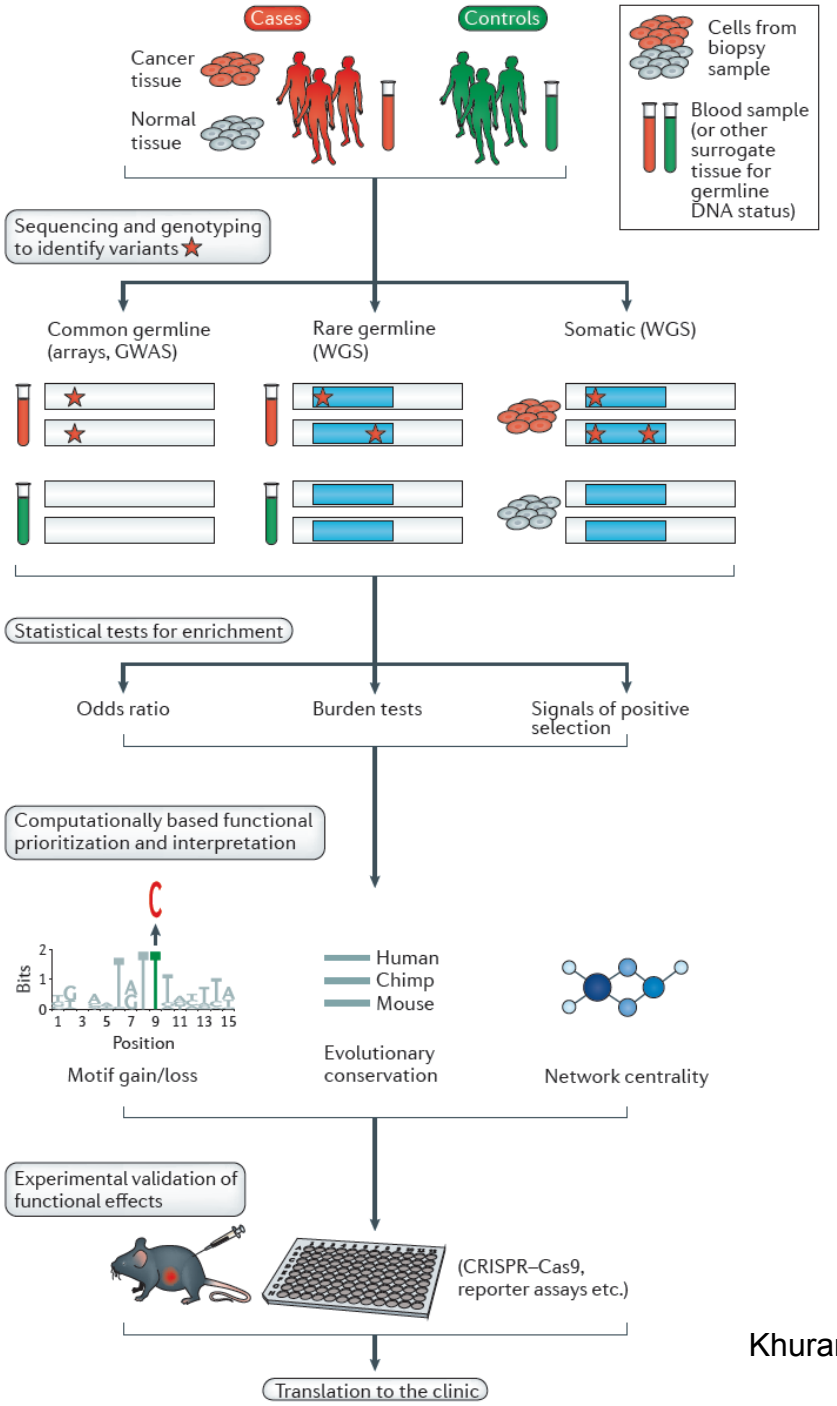


Perera et al, *Nature*, 2016
Sabarinathan et al, *Nature*, 2016
Khurana, *Nature News & Views*, 2016

Outline

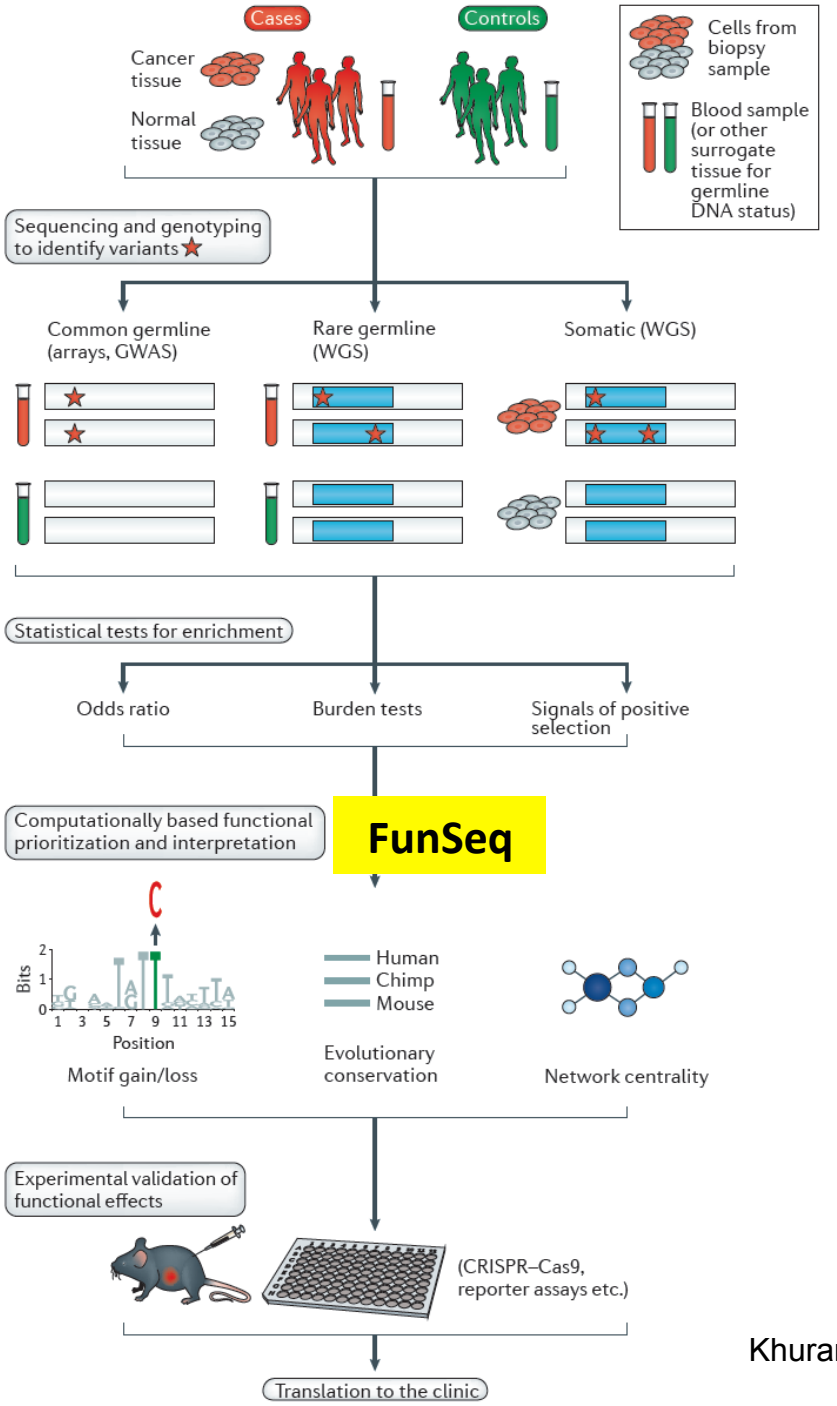
- Variants with high functional impact: FunSeq
- Driver elements w/ more recurrent & high functional impact mutations than expected randomly: CompositeDriver

Identifying noncoding variants associated with cancer



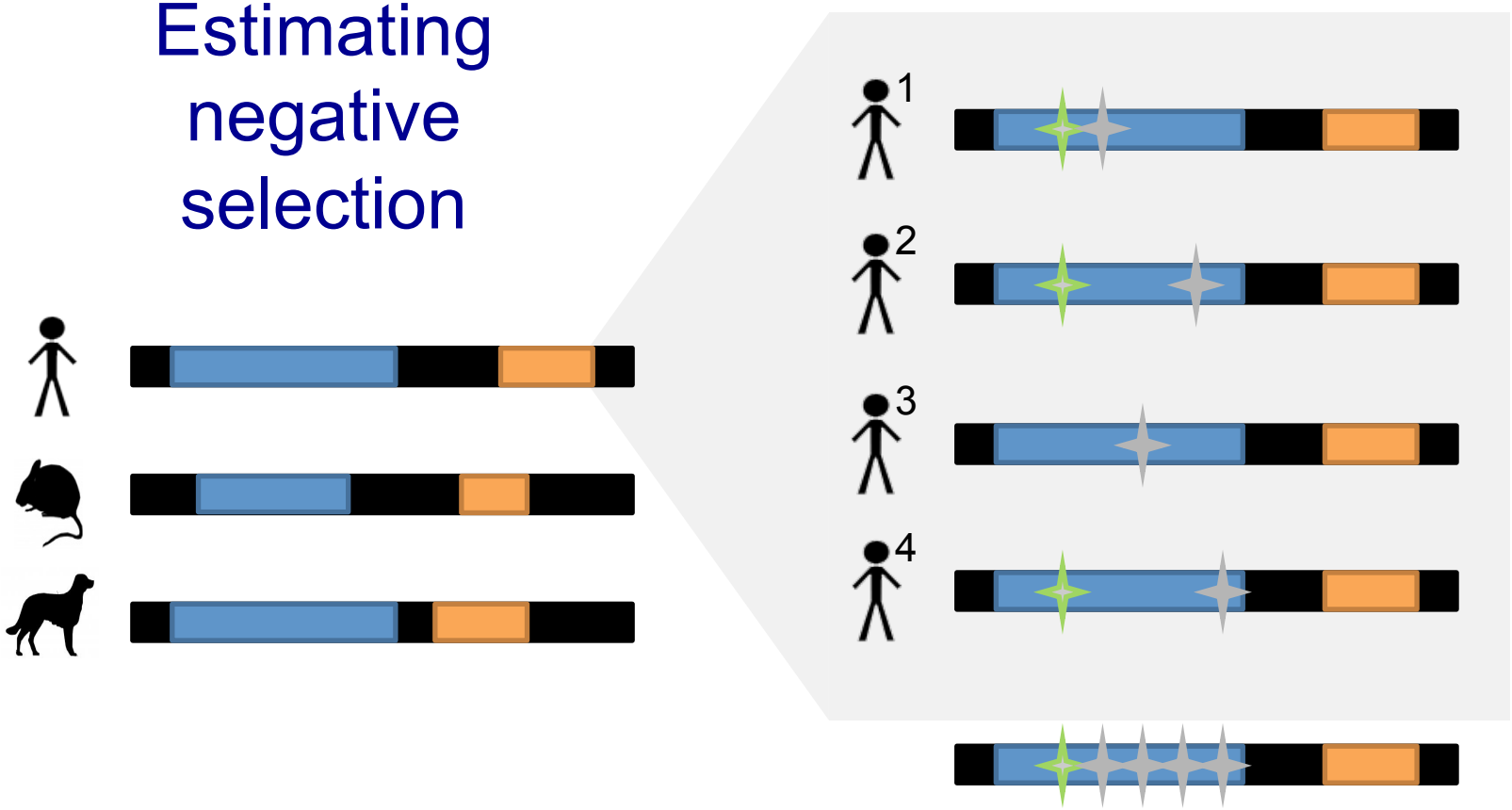
Khurana et al, *Nature Rev Genet*, 2016

Identifying noncoding variants associated with cancer



Khurana et al, *Nature Rev Genet*, 2016

Estimating negative selection



Evolutionary conservation

- Typically defined by comparison across species

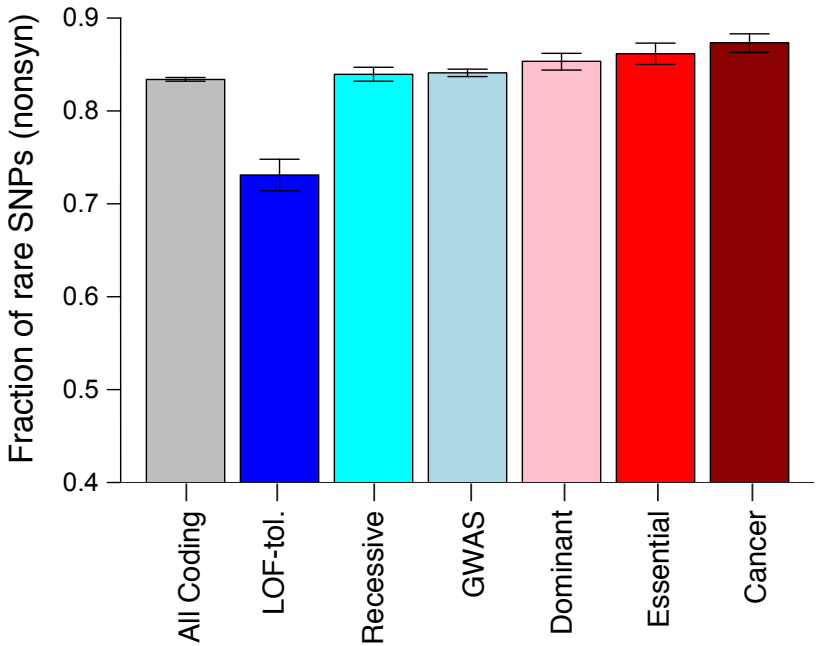
Conservation among humans

- Depletion of common variants/**Enrichment** 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



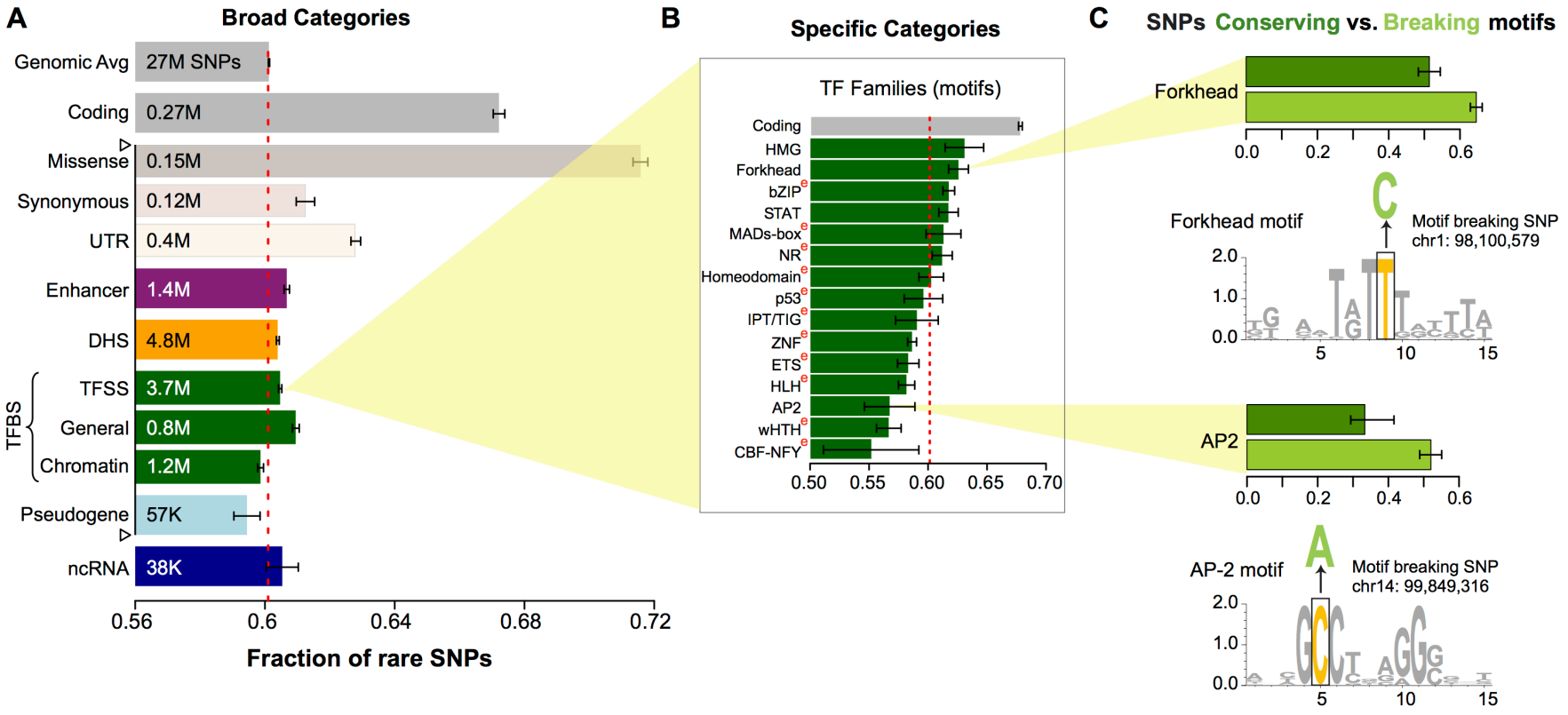
(rare=derived allele freq < 0.5%)

- 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

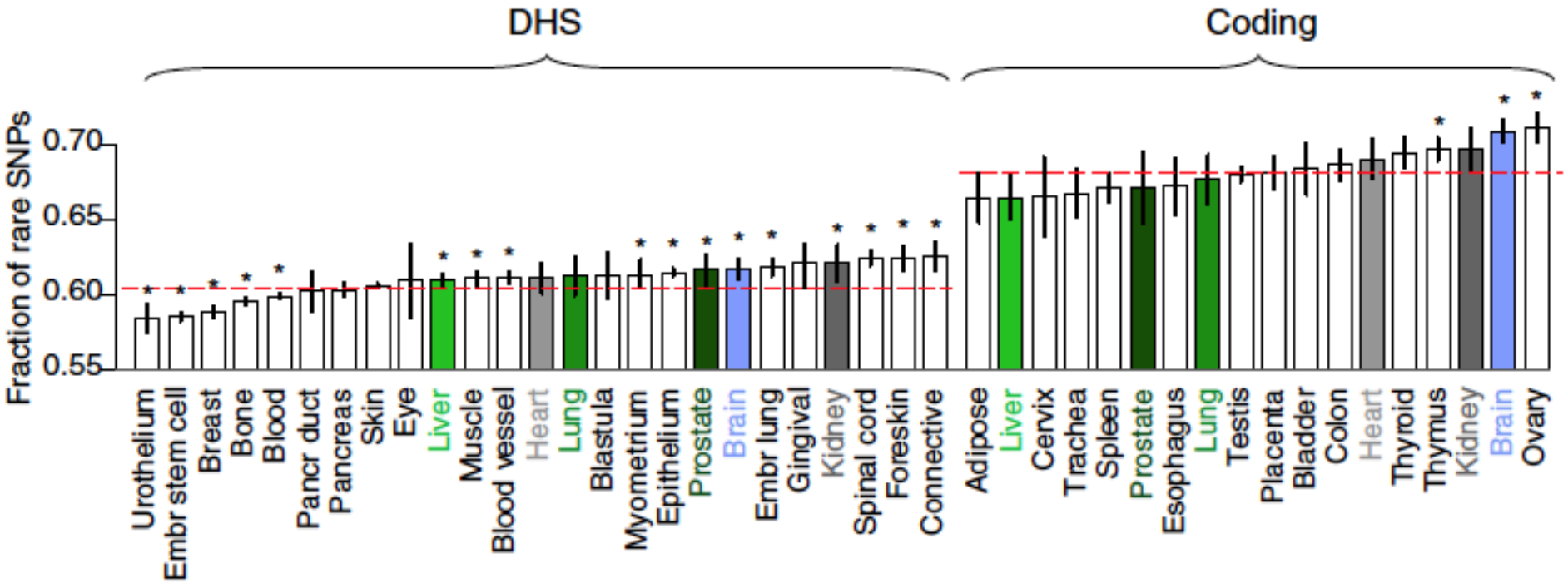
Khurana et al., *Science*, 2013

Organism-level negative selection in noncoding elements



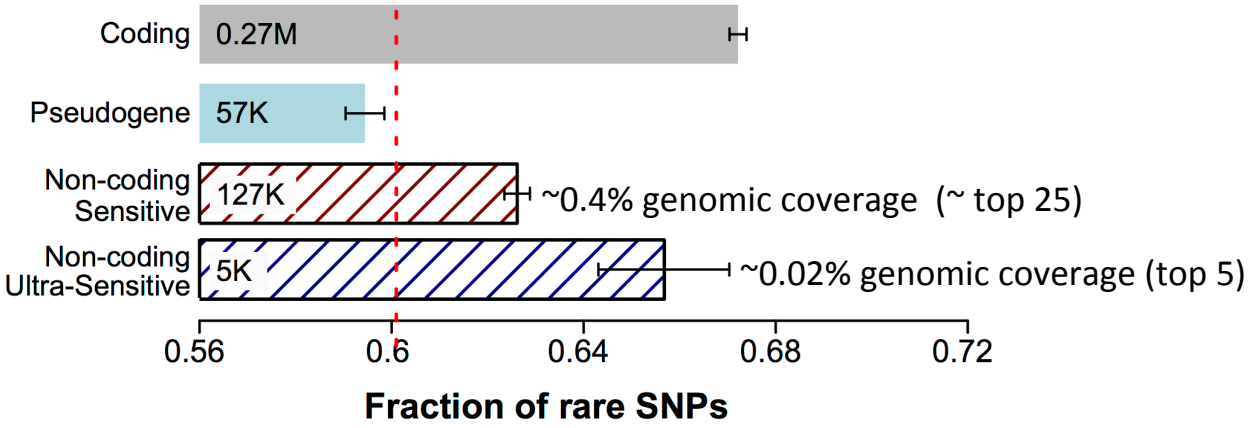
Khurana et al., *Science*, 2013

Negative selection and tissue-specificity of coding and noncoding regions

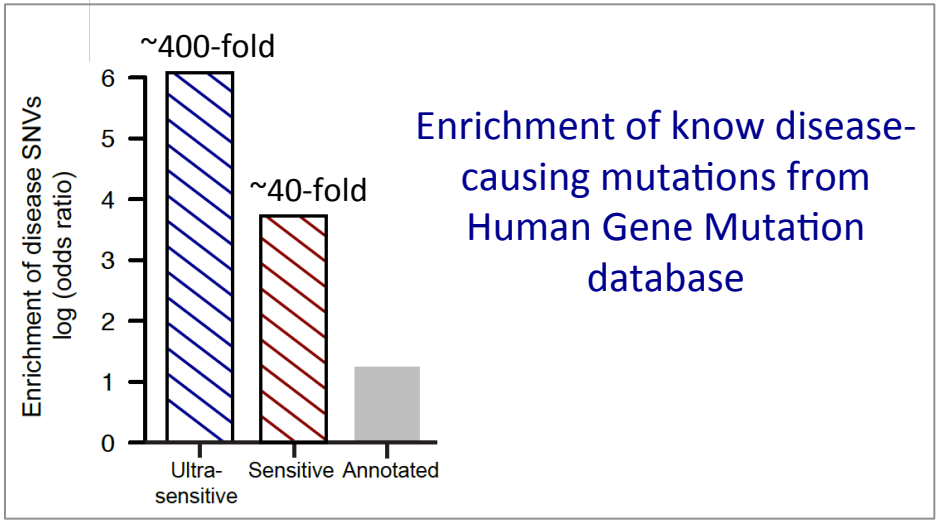


- Ubiquitously expressed genes and bound regions show stronger selection
- Differences in constraints amongst tissues
- Constraints in coding genes and regulatory genes are correlated across tissues

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

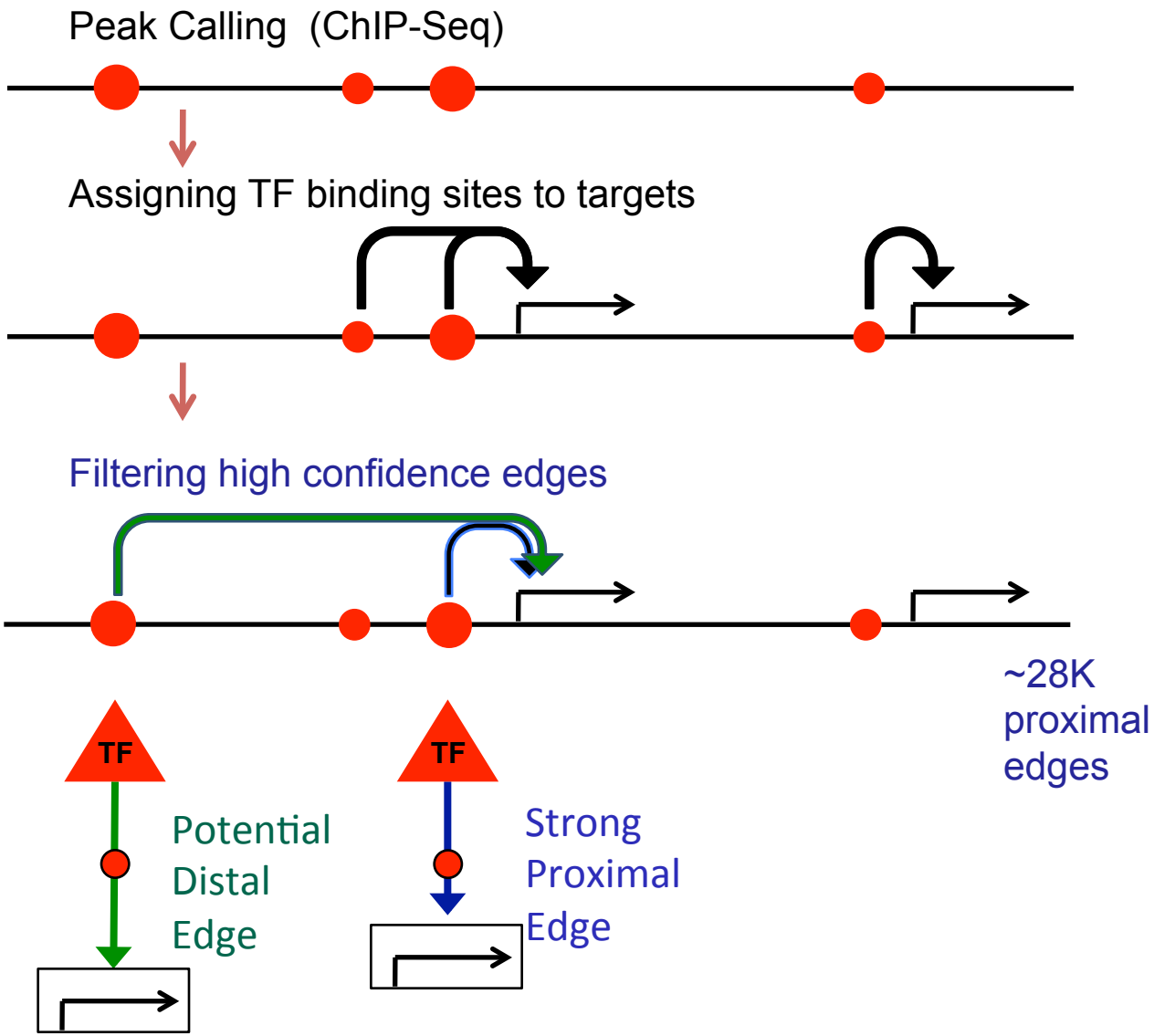


- Top categories among ranked 102 categories
- Binding peaks of some general TFs (eg *FAM48A*)
- Core motifs of some TF families (eg *JUN*, *GATA*)
- DHS sites in spinal cord and connective tissue

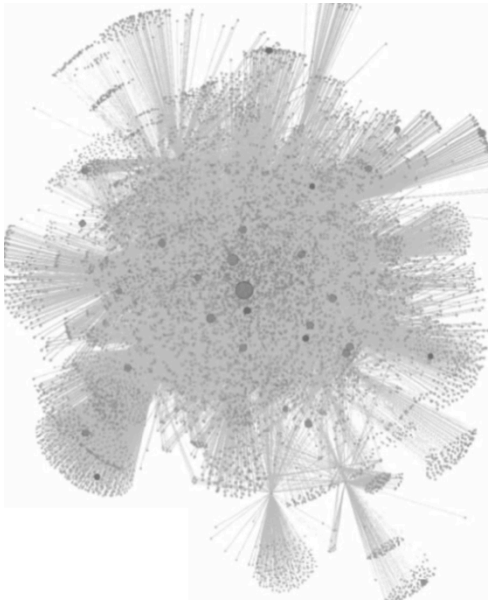


Enrichment of know disease-causing mutations from Human Gene Mutation database

Human regulatory network from ENCODE ChIP-Seq



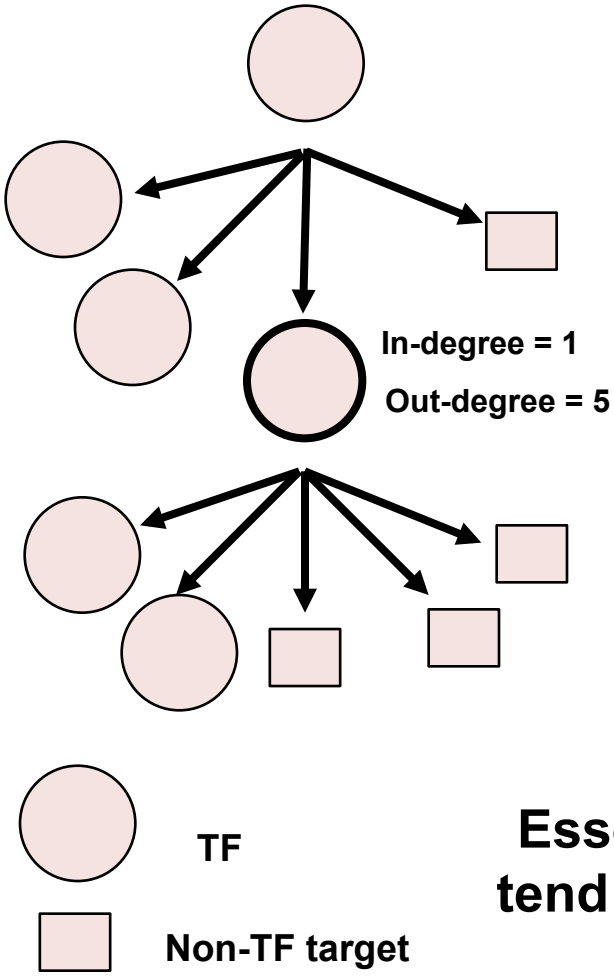
Nodes
119 TFs and ~9000 target genes
Edges
28,000 interactions



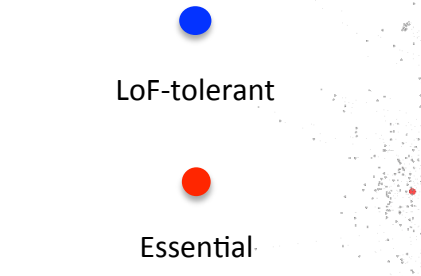
Using correlation with expression data

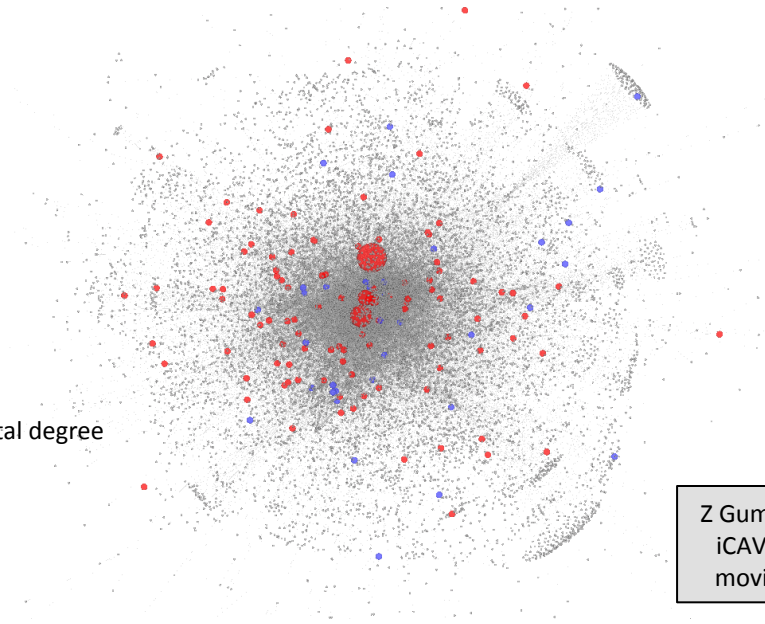
Gerstein¶…..Khurana¶…., *Nature*, 2012 (¶ co-first authors)
Yip et al, *Genome Res*, 2012

Gene essentiality and human regulatory network

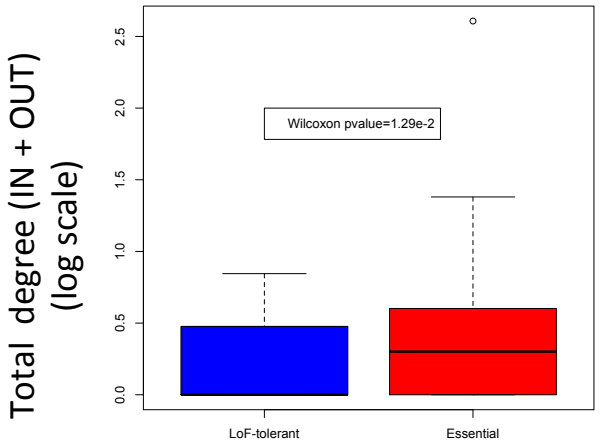


Essential genes tend to be central





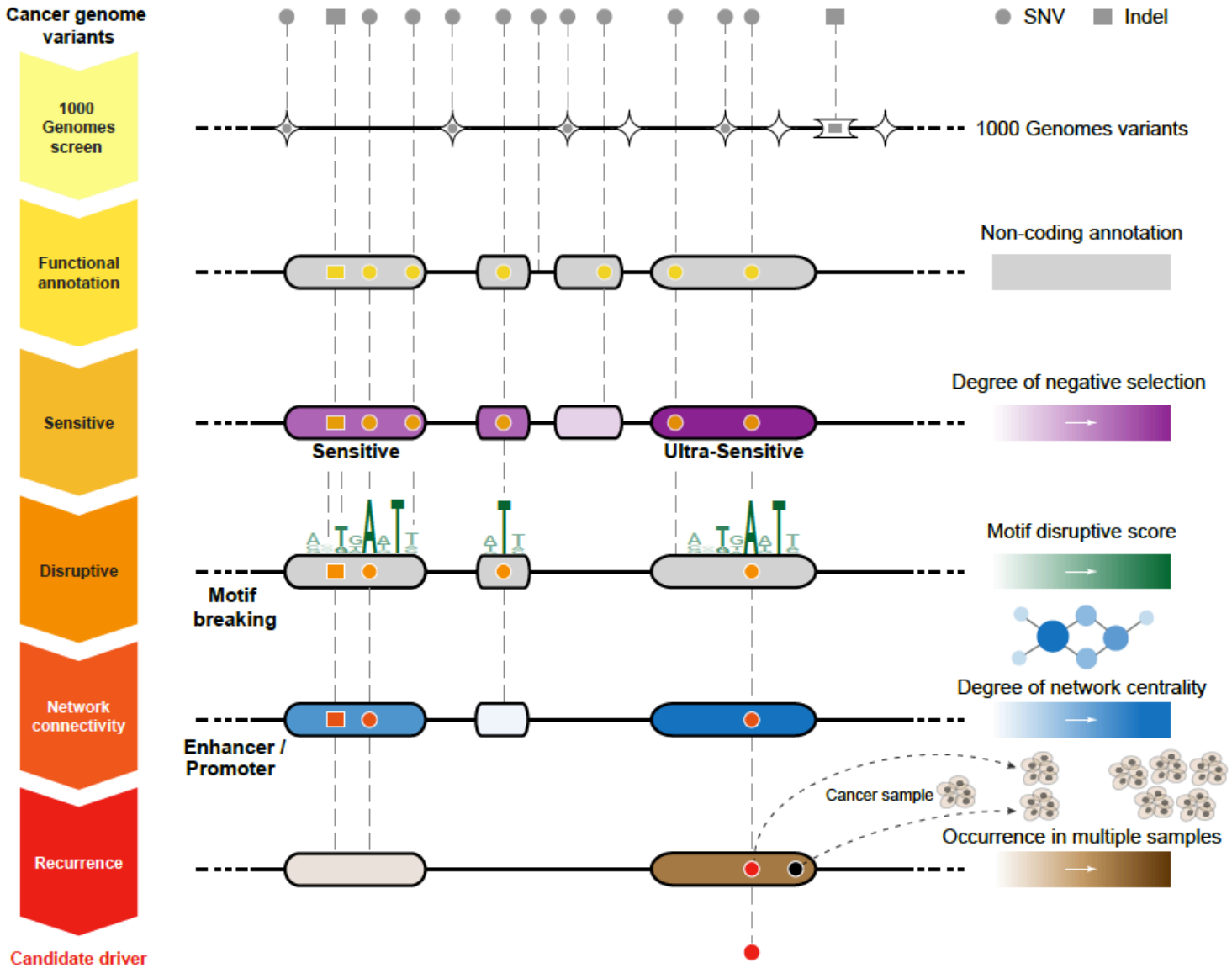
Z Gumus iCAVE movie



Gerstein¶…..Khurana¶…., *Nature*, 2012 (¶ co-first authors)

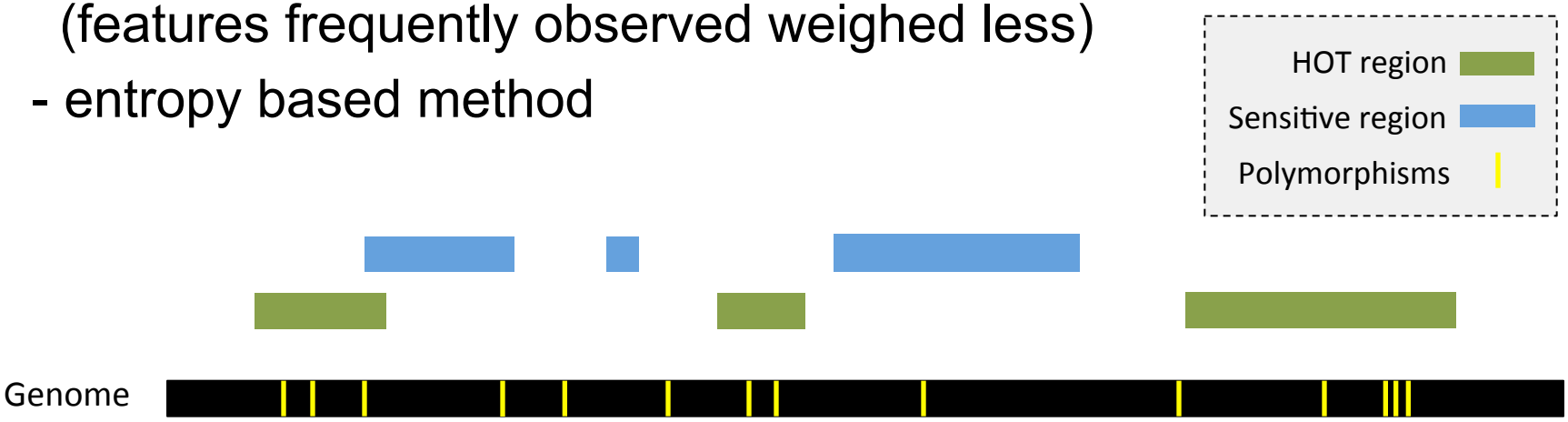
Khurana et al., *PLoS Comp. Bio.*, 2013

Identification of noncoding mutations with high impact: FunSeq



FunSeq2: weighted scoring scheme

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



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

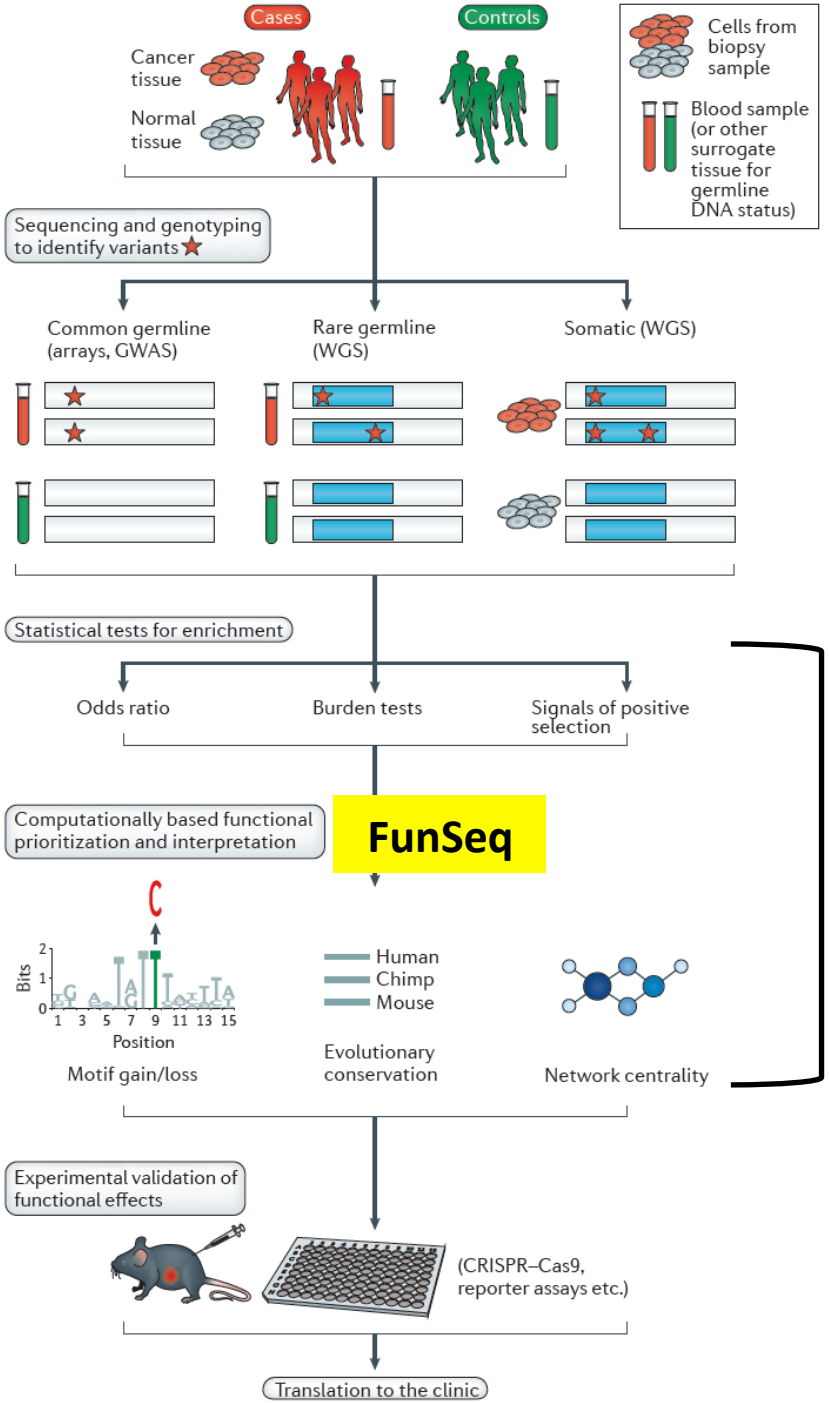
$p \uparrow \quad w_d \downarrow$ *p = probability of the feature overlapping natural polymorphisms*

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

https://github.com/khuranalab/FunSeq_PCAWG
http://funseq2.gersteinlab.org

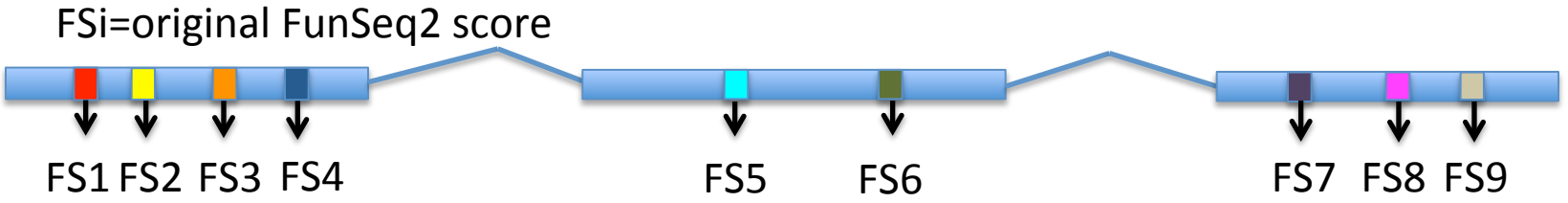
Fu et al., *Genome Biology*, 2014

Identifying noncoding variants associated with cancer

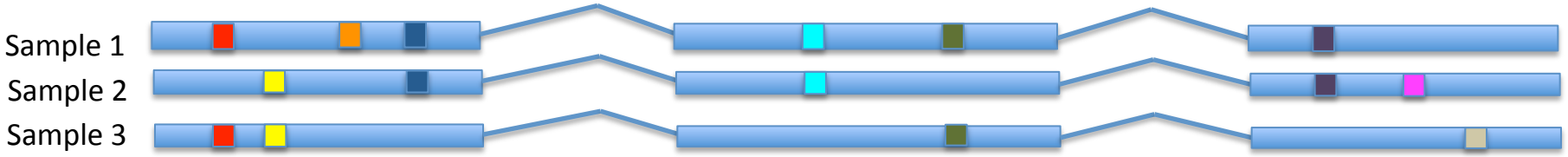


CompositeDriver for detecting driver coding & noncoding elements

(A) Alterations are functionally annotated by FunSeq2 pipeline



(B) Calculate positional recurrence of each mutation in the cohort



(C) Within each functional region, composite functional score (CFSr) is sum of recurrence multiplied by FunSeq2 score in each position with alteration.

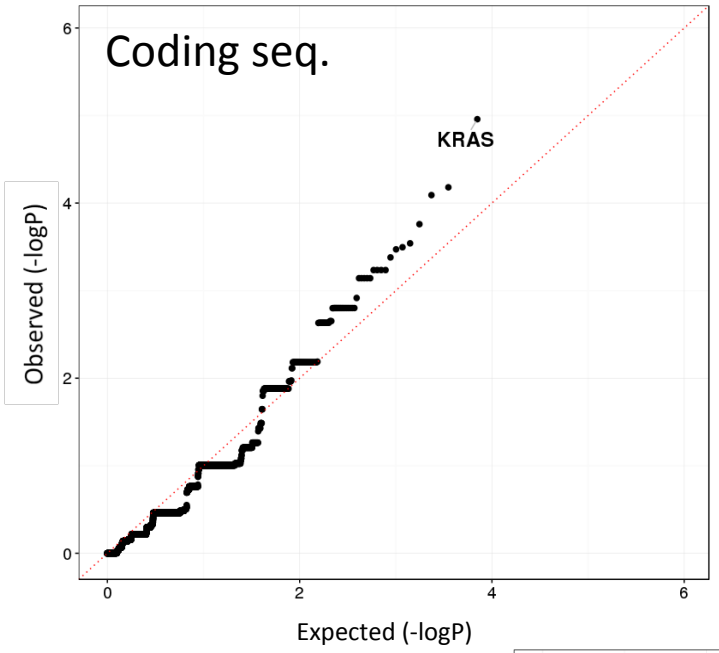
$$CFS_r = \sum_{i=0}^{n} W_i \times FS_i$$

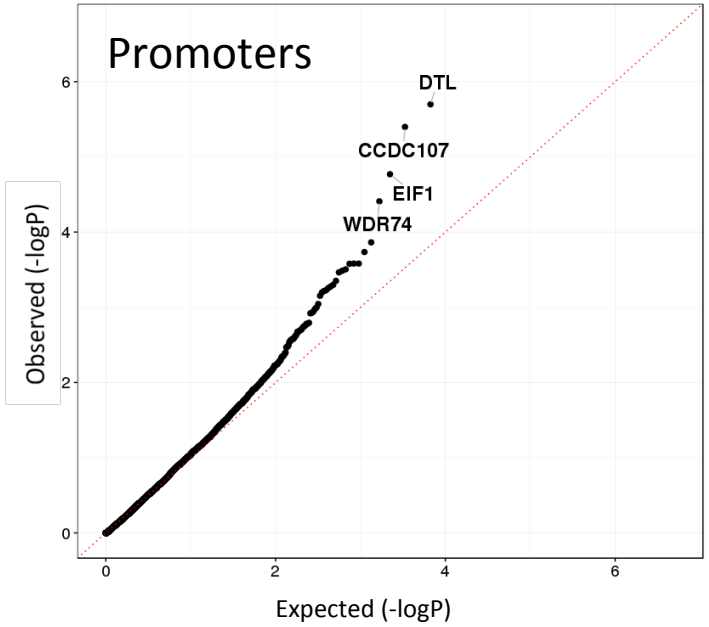
r = region (cds, promoter, enhancer and lincRNA)
n = number of variants in r
W_i = number of samples with variant i

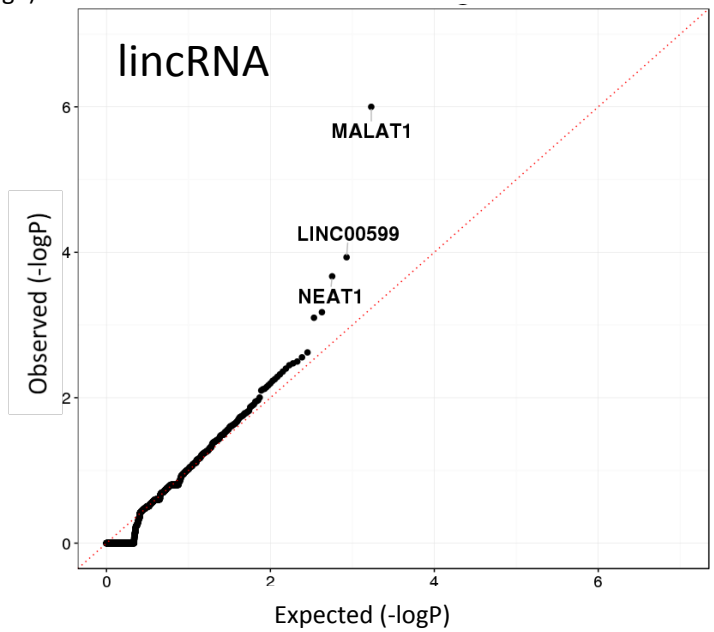
(D) P-value for each region is produced from permutation test and Benjamini and Hochberg method to correct multiple hypothesis testing.

Eric Minwei Liu

Results from 40 lung adenocarcinoma samples

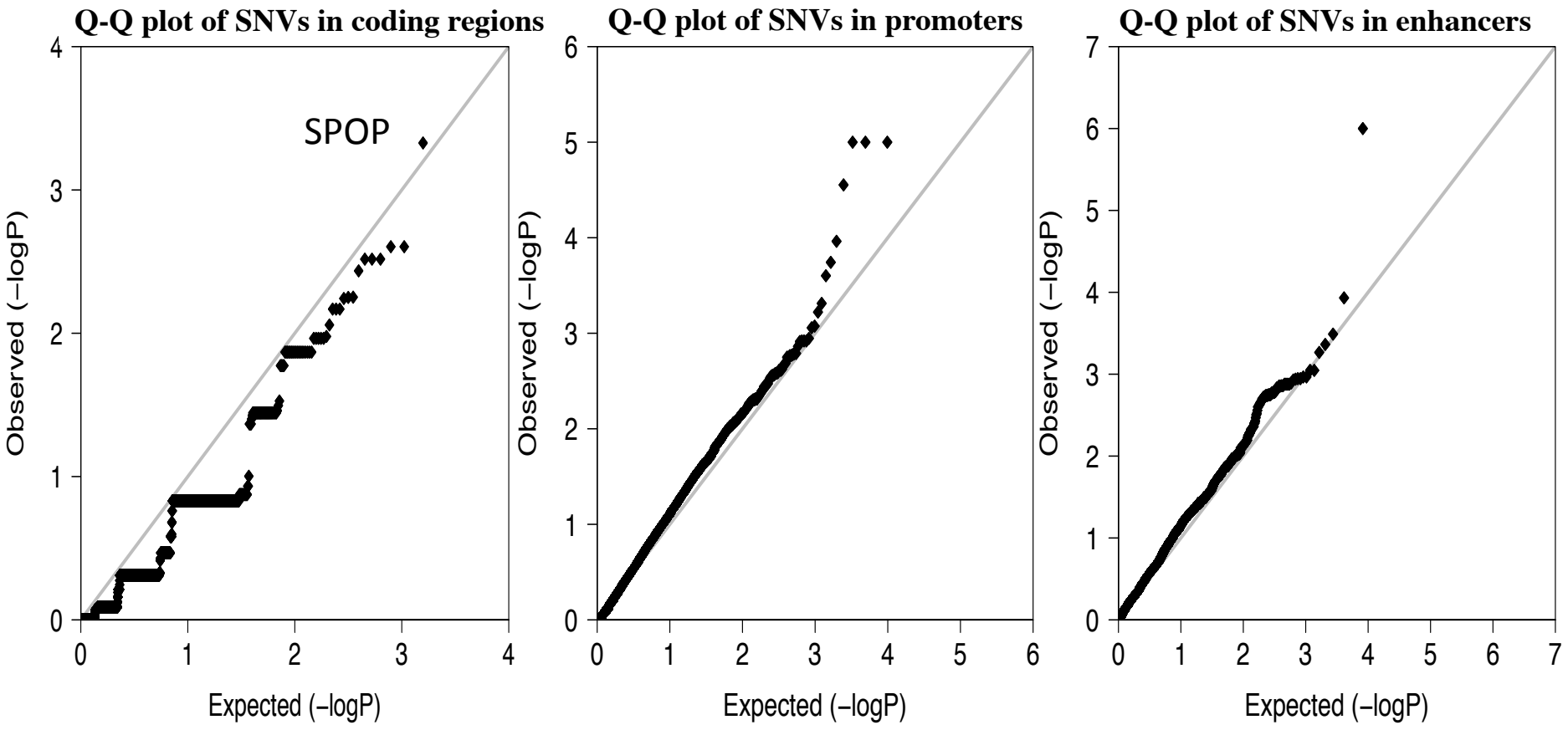






Data from TCGA

Results from 188 prostate cancer samples

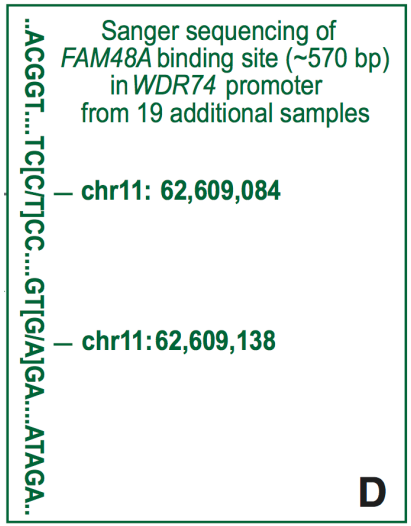


Data from ICGC, Baca et al Cell 2013, Berger et al Nature 2011

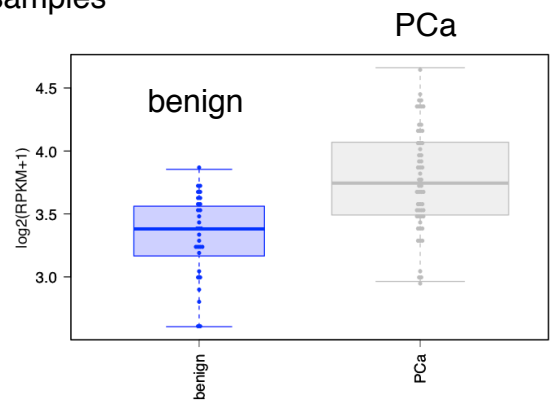
Functional validation of candidates in prostate cancer

WDR74 promoter

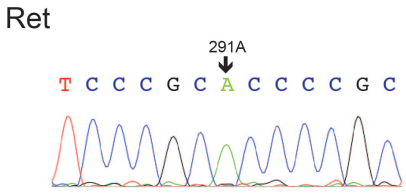
- Sanger sequencing in 19 additional samples confirms the recurrence

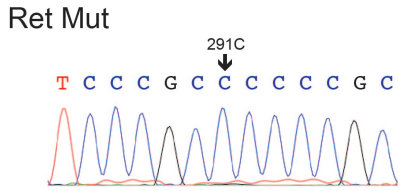


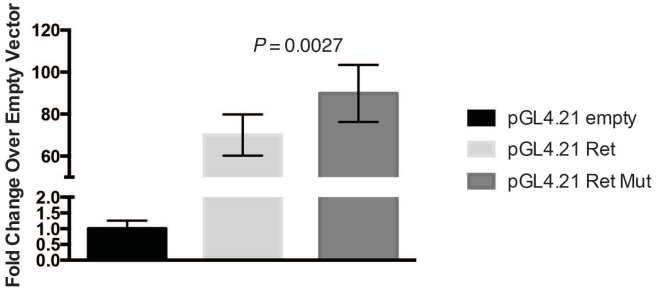
- *WDR74* shows increased expression in tumor samples



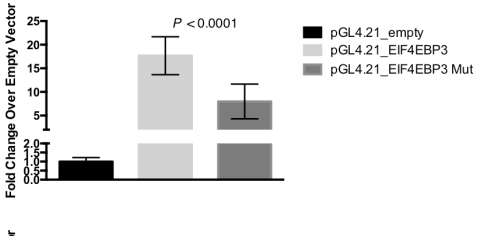
RET promoter
Increased activity







EIF4EBP3 promoter
Reduced activity



In collaboration w/ Mark Rubin

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

Khurana lab

Eric Minwei Liu

Priyanka Dhingra

Alexander Fundichely

Tawny Cuykendall

Andrea Sboner

Mark Rubin

Dimple Chakravarty

Kellie Cotter

Steve Lipkin

Chason Lee

Weill Cornell Medicine

Sandra and Edward
Meyer Cancer Center

Englander Institute for
Precision Medicine

Institute for Computational
Biomedicine