Information in noncoding DNA

Ekta Khurana
Associate Research Scientist
Yale University
New Haven, CT, USA

AACR Annual Meeting 2014

Large scale sequencing consortia

International Cancer Genome Consortium

Cancer genome sequencing

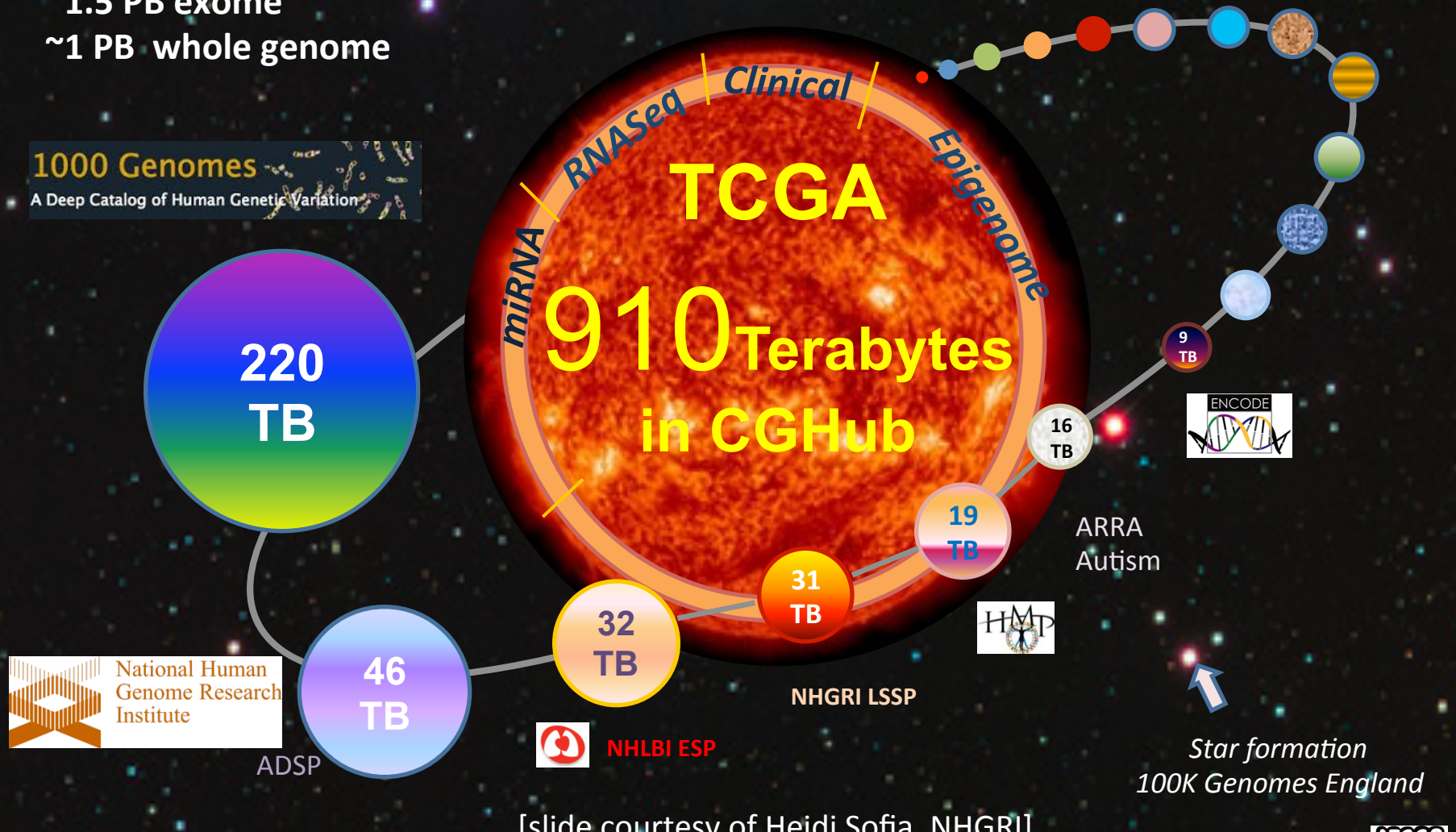
Population scale sequencing of healthy humans

Encyclopedia of DNA Elements, Genome annotation

Seq Universe

SRA >1 petabyte

TCGA endpoint: ~2.5 Petabytes
- **~1.5 PB exome**
- **~1 PB whole genome**



[slide courtesy of Heidi Sofia, NHGRI]

JESS3

Outline

Noncoding regions play an important role in gene regulation. Tumor genome contains thousands of somatic mutations in noncoding regions: how do we identify the functionally important ones

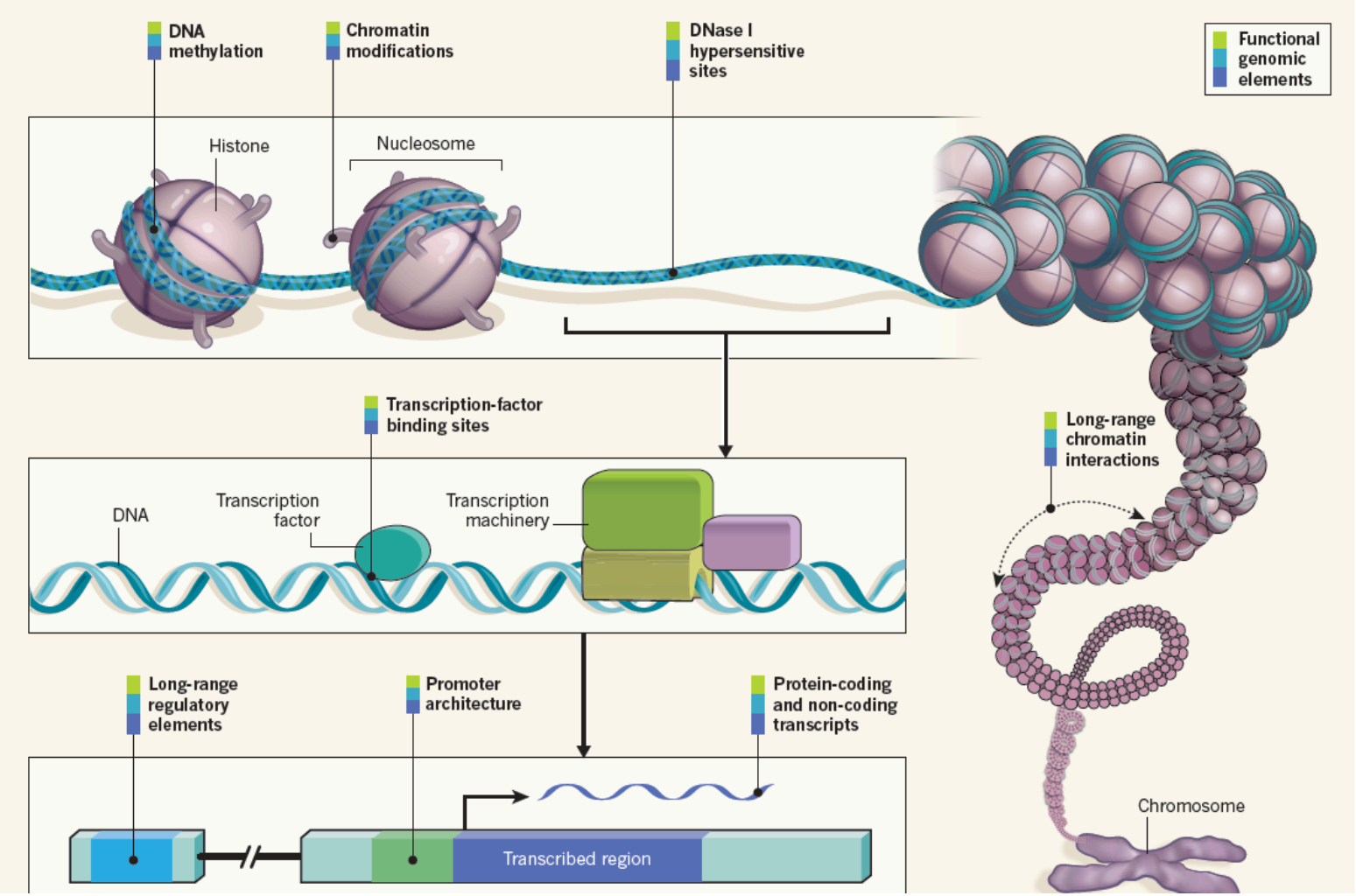
- Noncoding annotations
- Role of noncoding regions in cancer
- Our method for identifying noncoding candidate drivers

Outline

Noncoding regions play an important role in gene regulation. Tumor genome contains thousands of somatic mutations in noncoding regions: how do we identify the functionally important ones

- Noncoding annotations
- Role of noncoding regions in cancer
- Our method for identifying noncoding candidate drivers

Functional elements in the genome



Enhancers,
Insulators, Silencers

Ecker, *Nature* – News and Views, 2012

ChIP-Seq (Chromatin immunoprecipitation followed by sequencing) to identify TF (transcription factor) binding sites

Crosslink proteins and DNA

Sample fragmentation
- Sonication
- Endonuclease (ChIP–exo)

Immunoprecipitate and then purify DNA

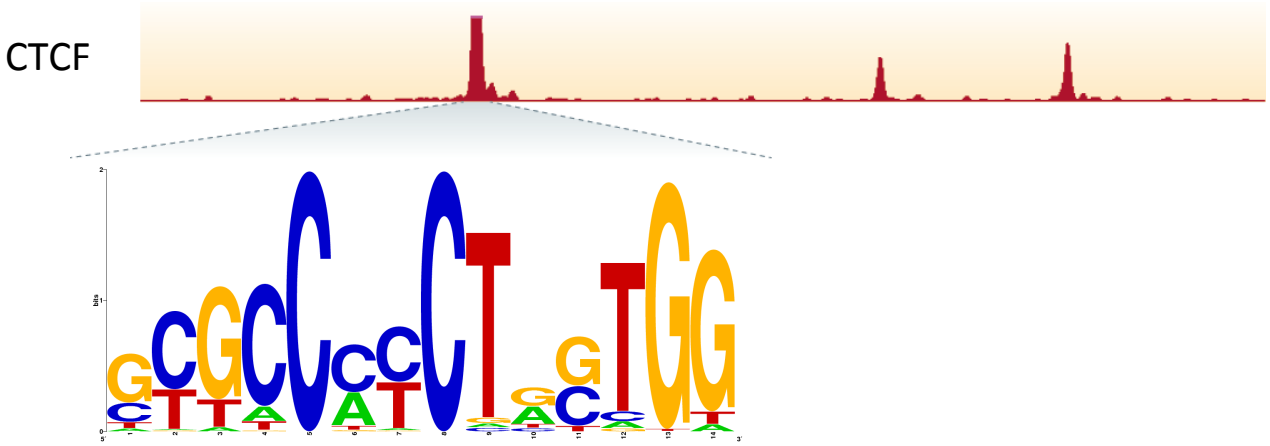
Amplify, if few cells
- LinDA

DNA library creation and sequencing →

→ Map reads to the genome

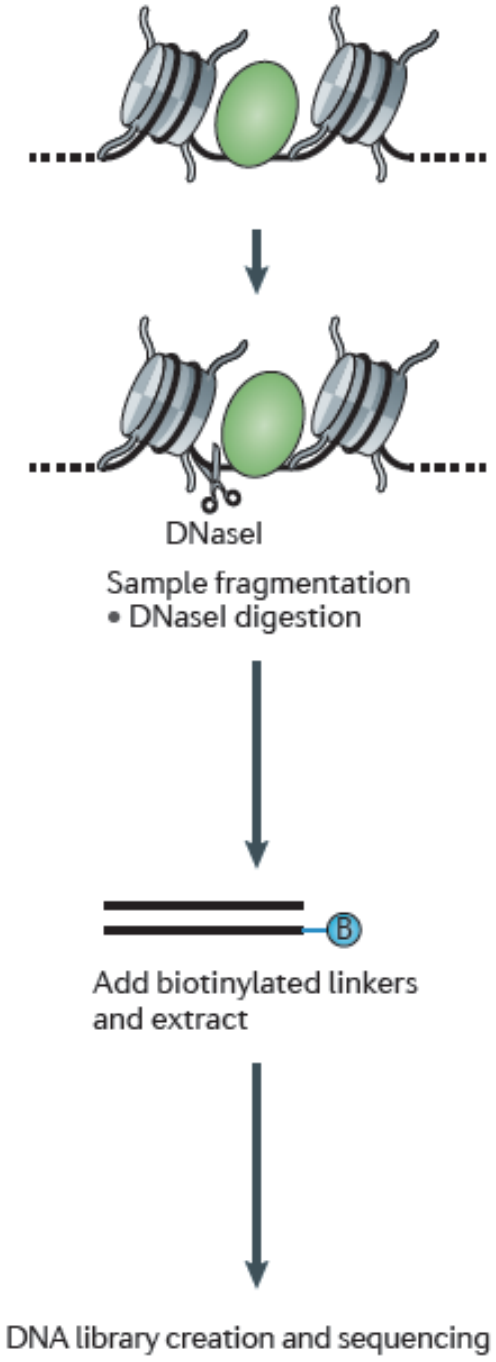
TF binding peak

CTCF



TF binding motif

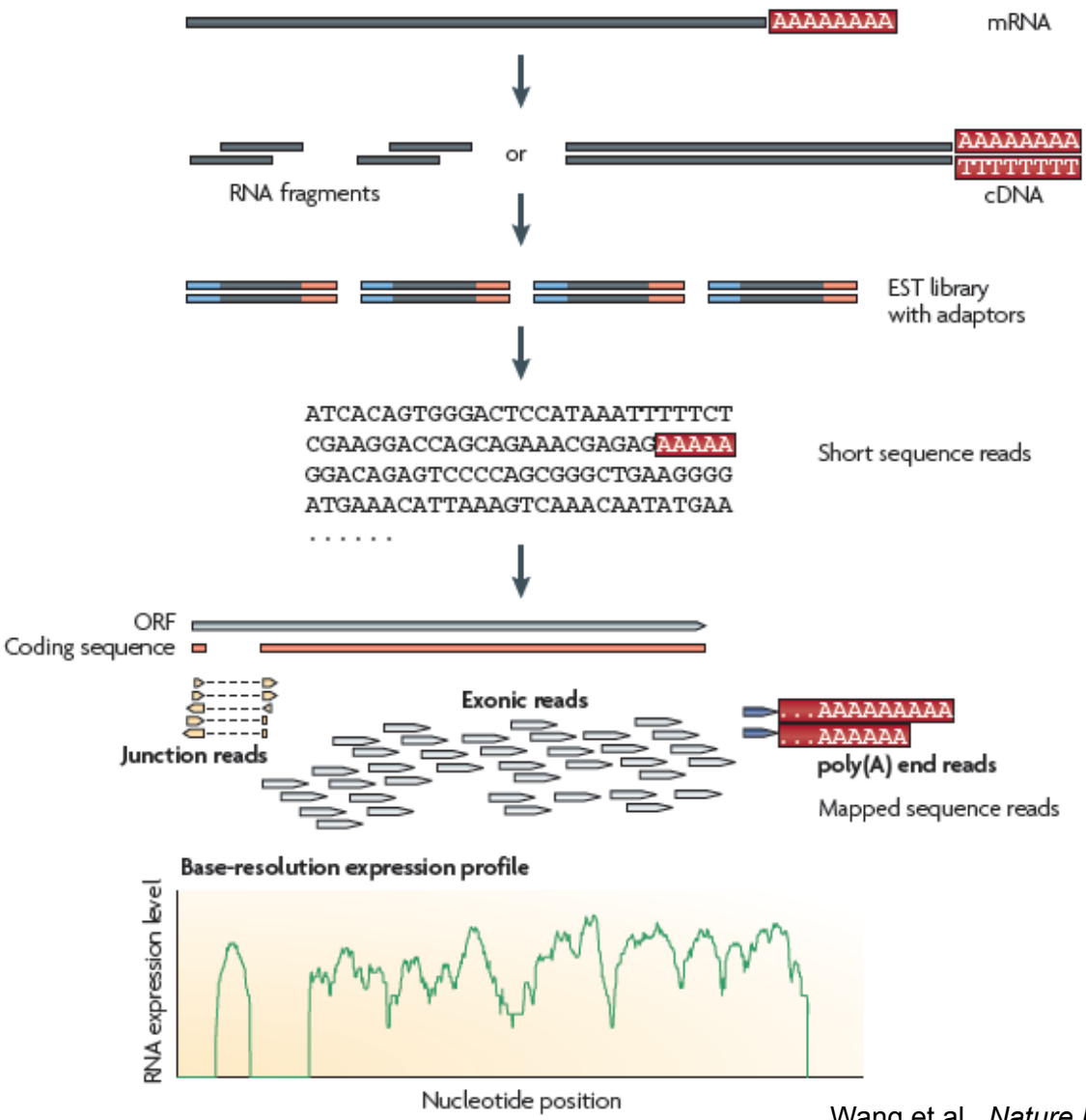
Furey, *Nature Rev. Genet.*, 2012



DNase-Seq to identify regions of open chromatin where any TF could be bound

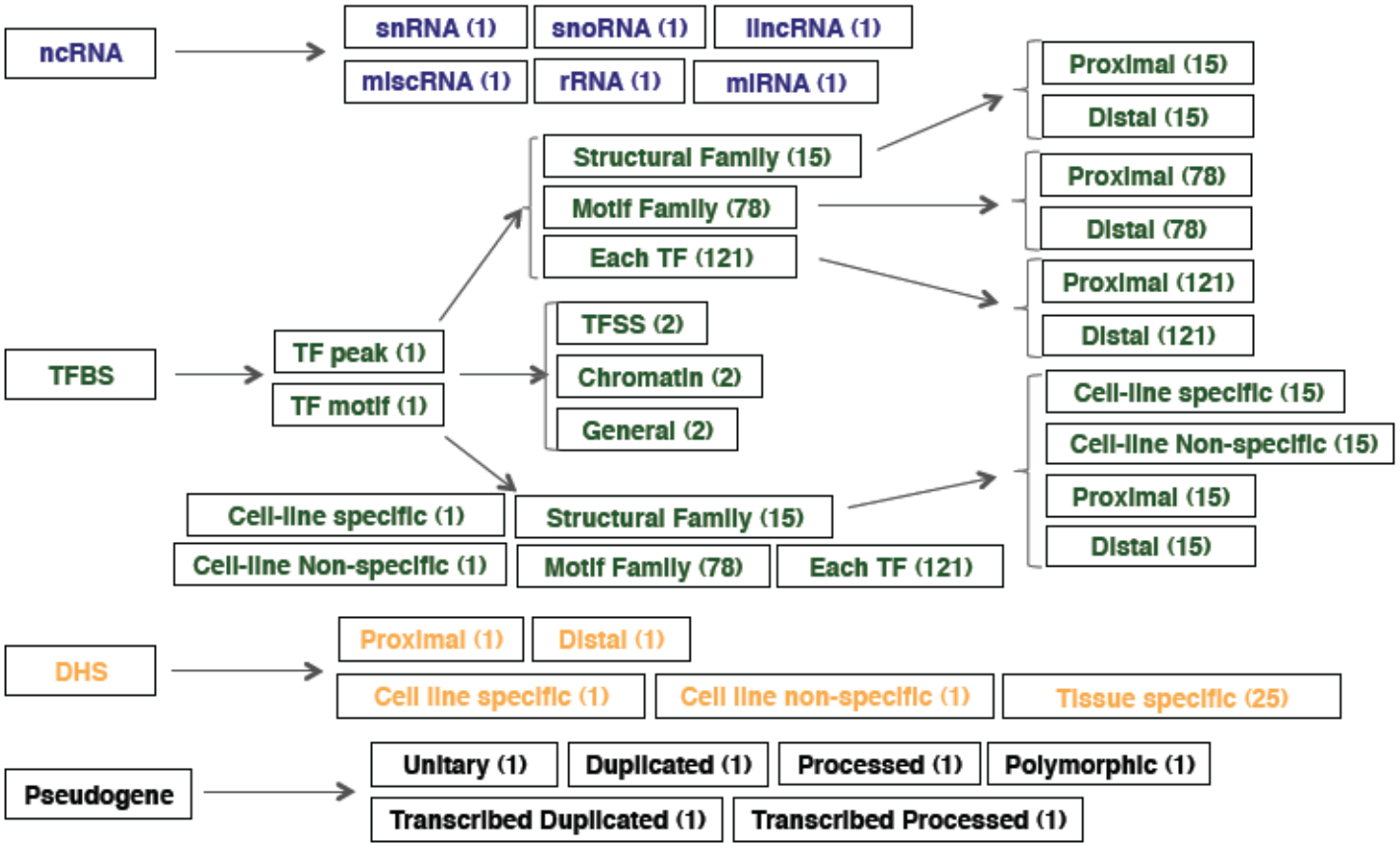
Furey, *Nature Rev. Genet.*, 2012

RNA-Seq to find transcribed regions

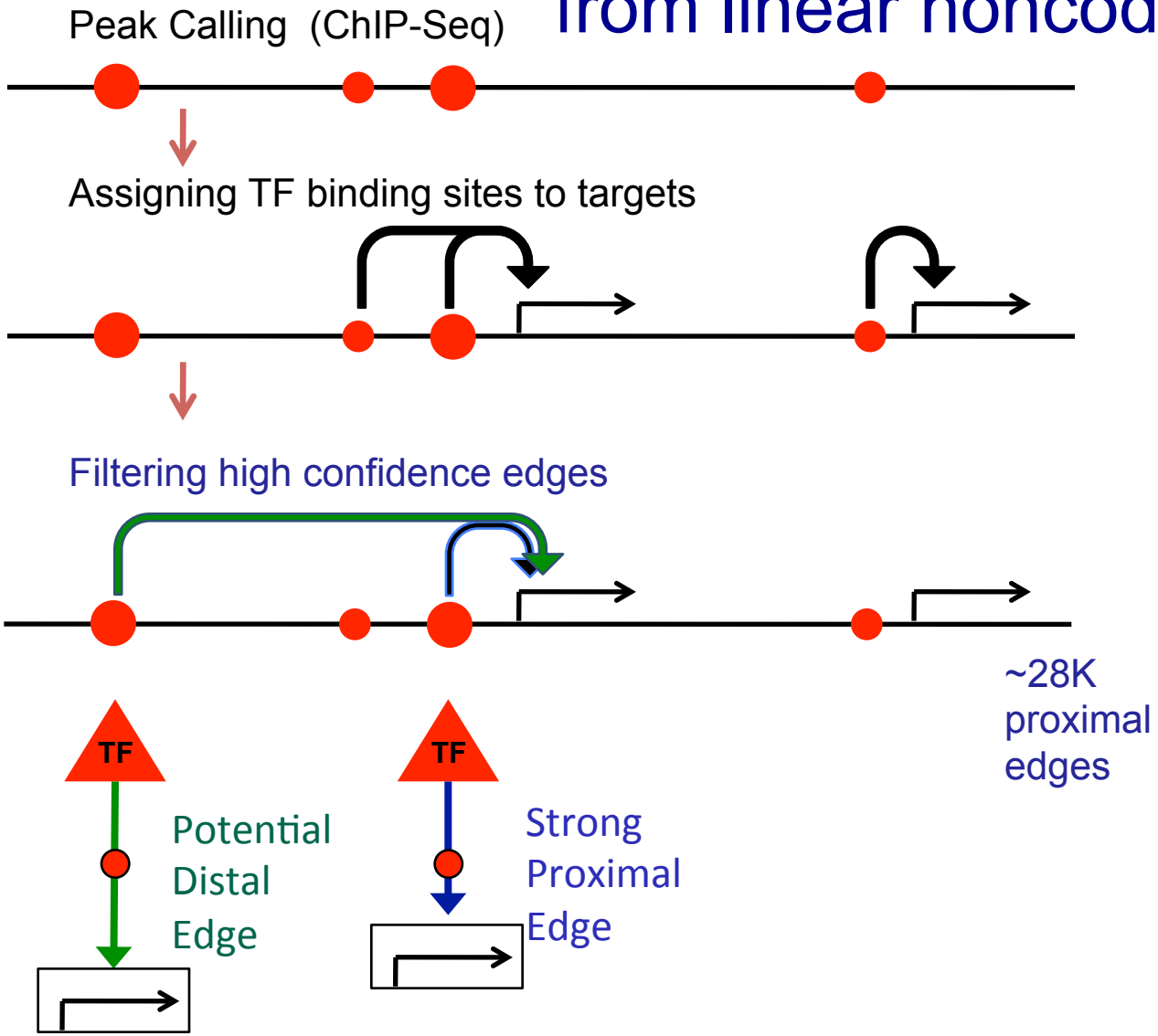


Wang et al., *Nature Rev. Genet.*, 2009

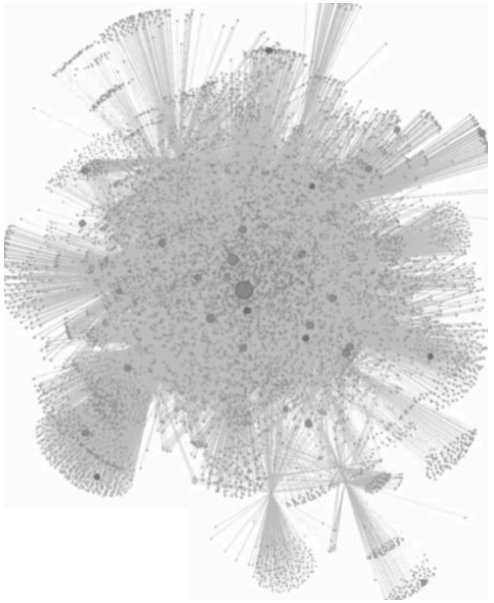
Complexity in levels of noncoding annotations



Building human regulatory network from linear noncoding annotations



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



Using correlation with expression data

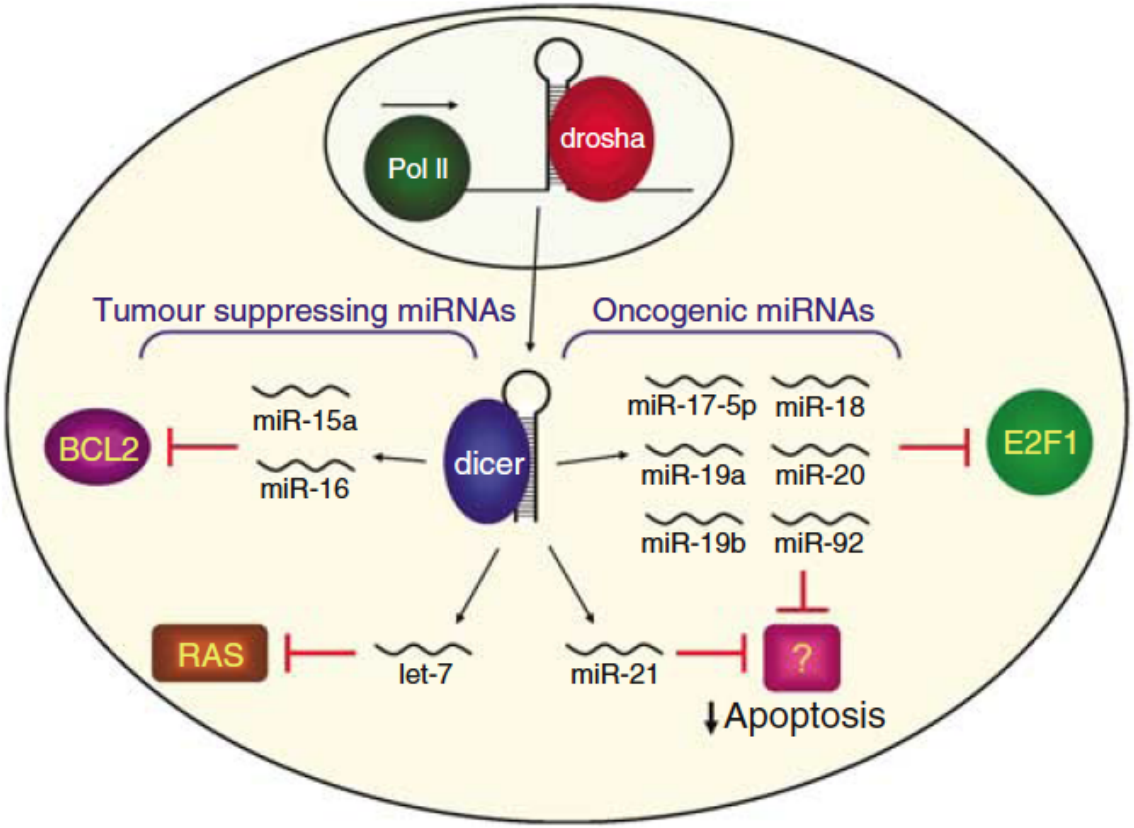
Gerstein¶…..Khurana¶…., *Nature*, 2012 (¶ co-first authors)
Yip et al, *Genome Res*, 2012

Outline

Noncoding regions play an important role in gene regulation. Tumor genome contains thousands of somatic mutations in noncoding regions: how do we identify the functionally important ones

- Noncoding annotations
- Role of noncoding regions in cancer
- Our method for identifying noncoding candidate drivers

miRNAs can act as oncogenes or tumor-suppressor genes



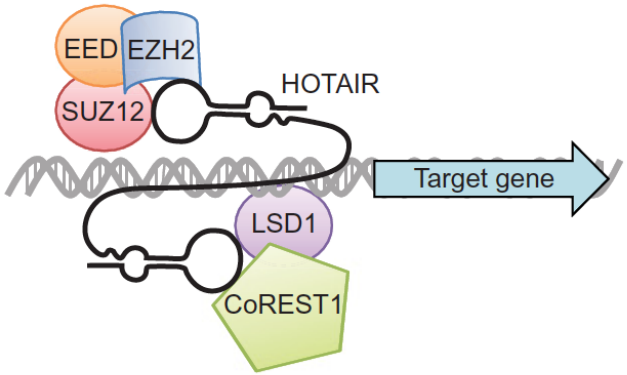
miRNA

- **~22 nt long**
- **Regulate ~30% of mRNAs**
- **Negative regulation of gene expression**

Hwang & Mendell, *British Journal of Cancer*, 2006

Some lncRNA mechanisms

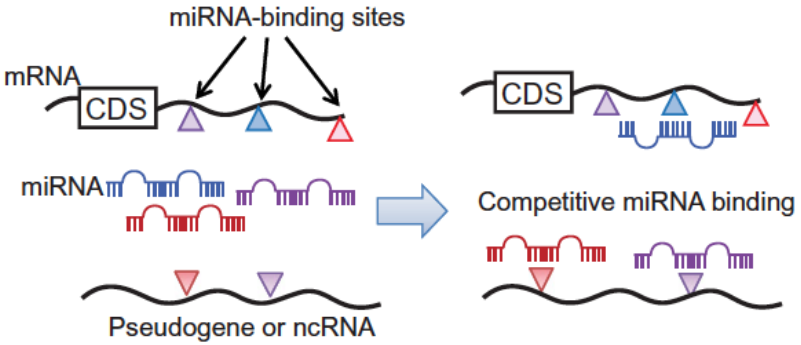
Flexible scaffold for chromatin-modifying complexes



HOTAIR
HOX antisense intergenic RNA

lncRNA
- **>200bp**
- **Include lincRNA and pesudogene derived transcripts**

miRNA sequestration



***PTEN* pseudogene loss in cancer related to reduced *PTEN* expression**

[Other mechanisms: Enhancer RNAs, Tumor suppressor signaling, RNA processing, RNA-RNA interactions]

Prensner & Chinnaiyan, *Cancer Discovery*, 2011
Poliseno et al., *Nature*, 2010
Gupta et al, *Nature*, 2010

Mutations affecting TF binding

Huang et al, *Science*, 2013

Highly Recurrent *TERT* Promoter Mutations in Human Melanoma

Horn et al., *Science*, 2013

***TERT* Promoter Mutations in Familial and Sporadic Melanoma**

***TERT* (Telomerase Reverse Transcriptase) promoter mutation frequency**

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)

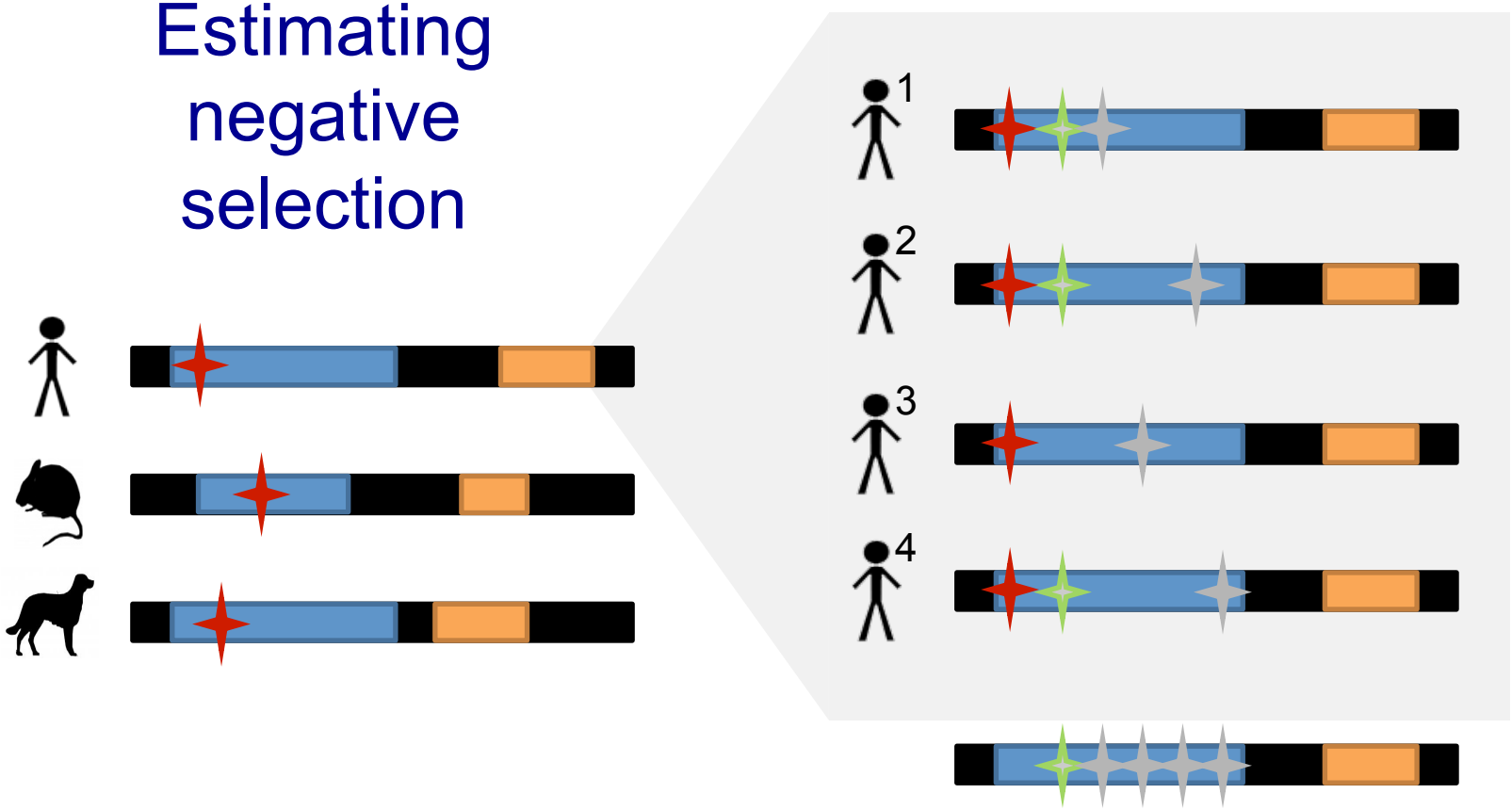
Killela et al, *PNAS*, 2013

Outline

Noncoding regions play an important role in gene regulation. Tumor genome contains thousands of somatic mutations in noncoding regions: how do we identify the functionally important ones

- Noncoding annotations
- Role of noncoding regions in cancer
- Our method for identifying noncoding candidate drivers

Estimating negative selection



Evolutionary conservation

- Typically defined by comparison across species

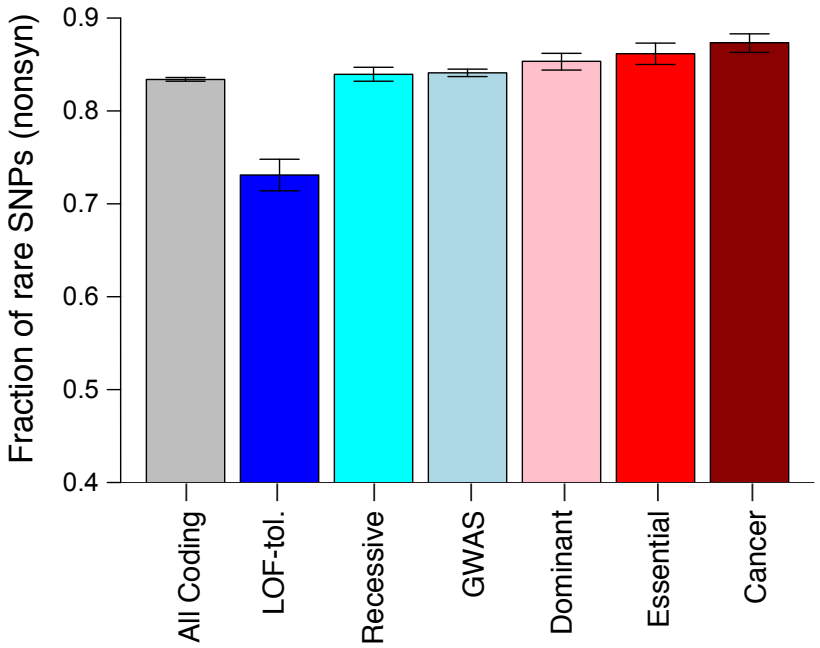
Conservation among humans

- 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 in protein-coding genes



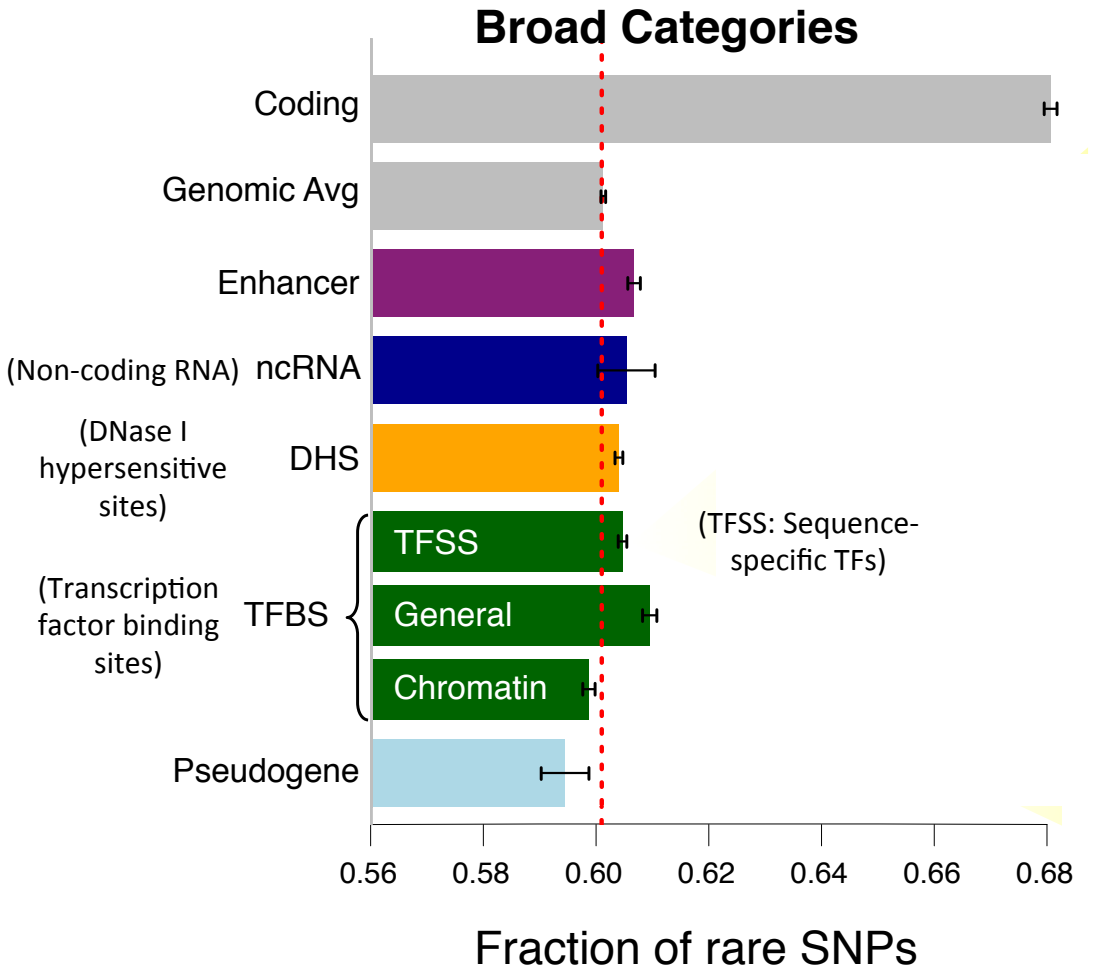
(rare=derived allele freq < 0.5%)

- 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
- Cancer drivers under strongest negative selection

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

Khurana et al., *Science*, 2013

Negative selection in noncoding elements

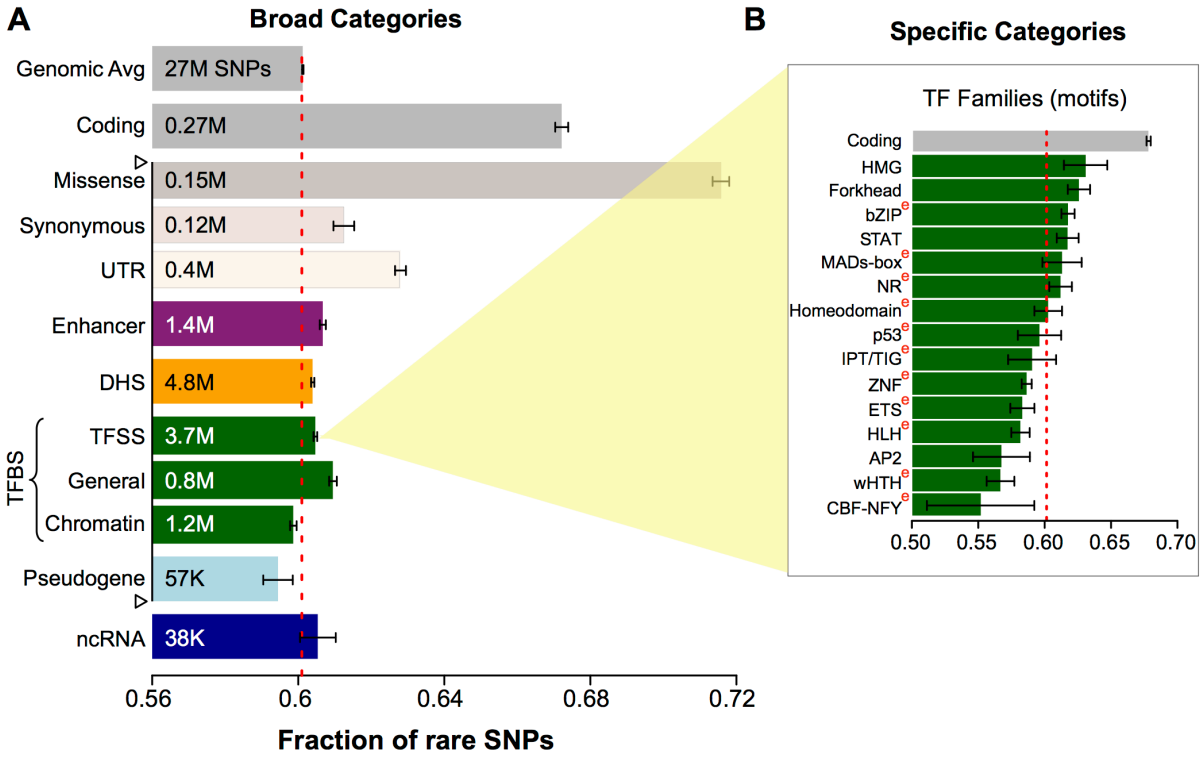


- Broad categories of regulatory regions under negative selection
- Consistent with previous studies

ENCODE, *Nature*, 2012

Khurana et al., *Science*, 2013

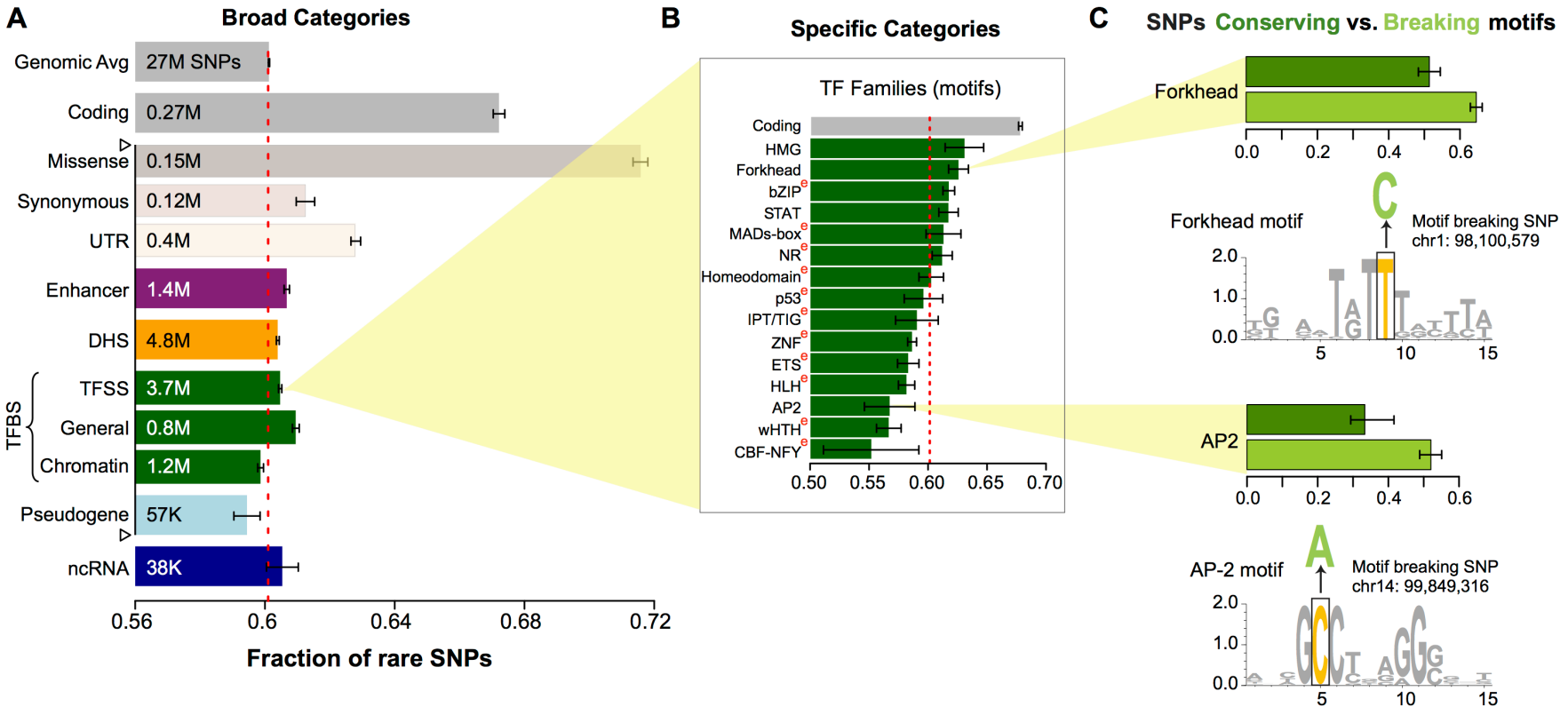
Differential selective constraints among sub-categories



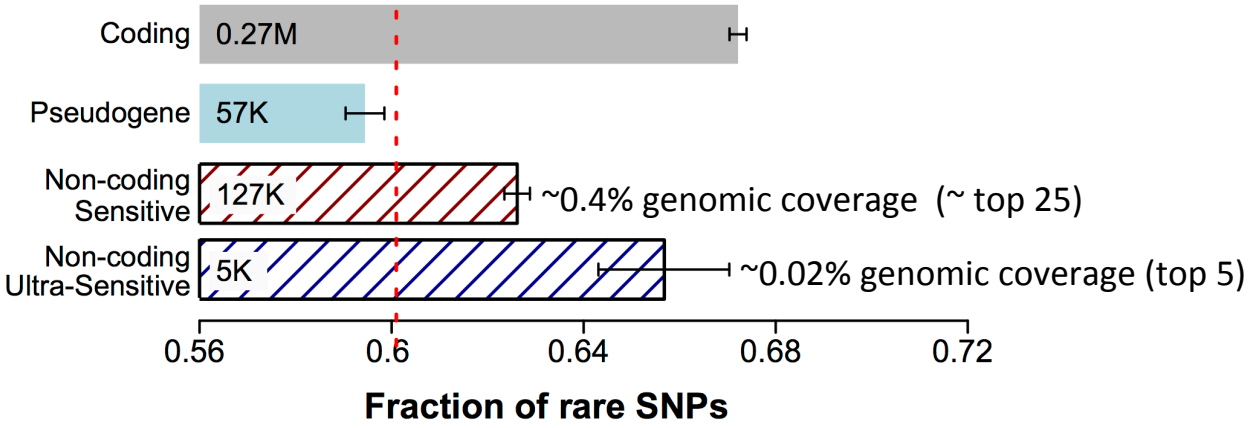
- Divide broad categories
  - ncRNA: snRNA, snoRNA, miRNA, lincRNA
  - Motifs & binding sites of different TF families
  - TFBSs divide into proximal vs distal and cell-line–specific vs – non-specific

Khurana et al., *Science*, 2013

SNPs which break TF motifs are under stronger selection



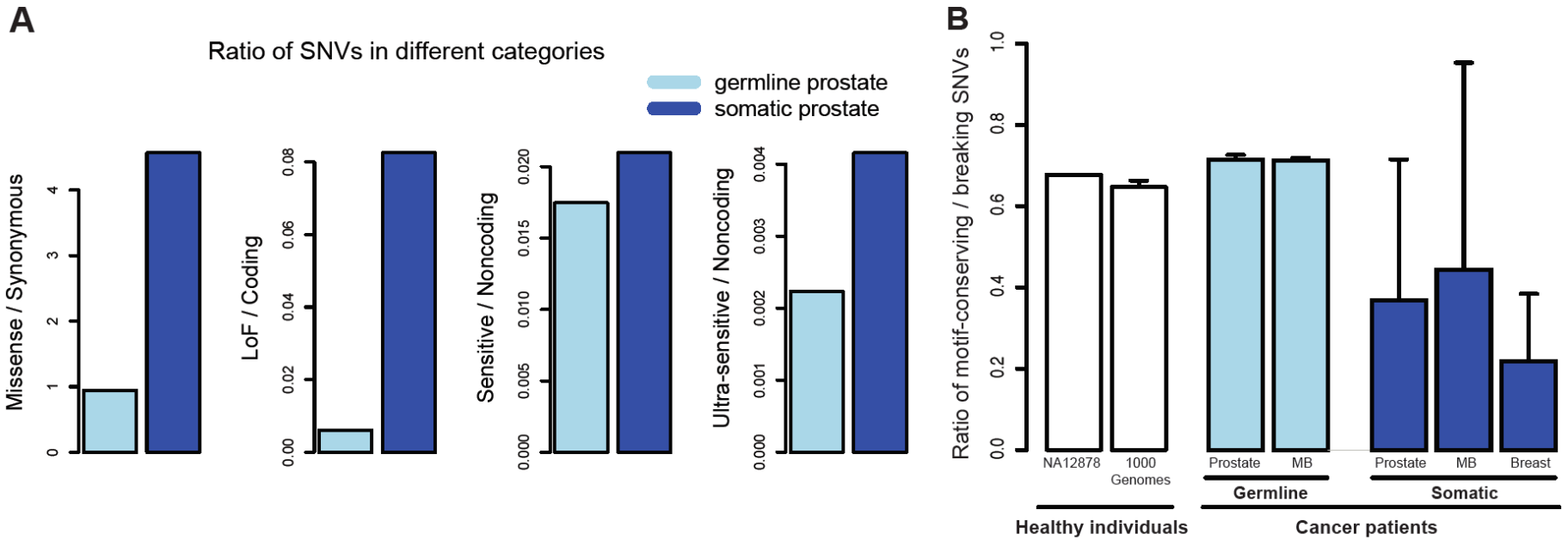
Which noncoding categories are under very strong “coding-like” selection ?



- Identify the top categories under strong negative selection
- 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

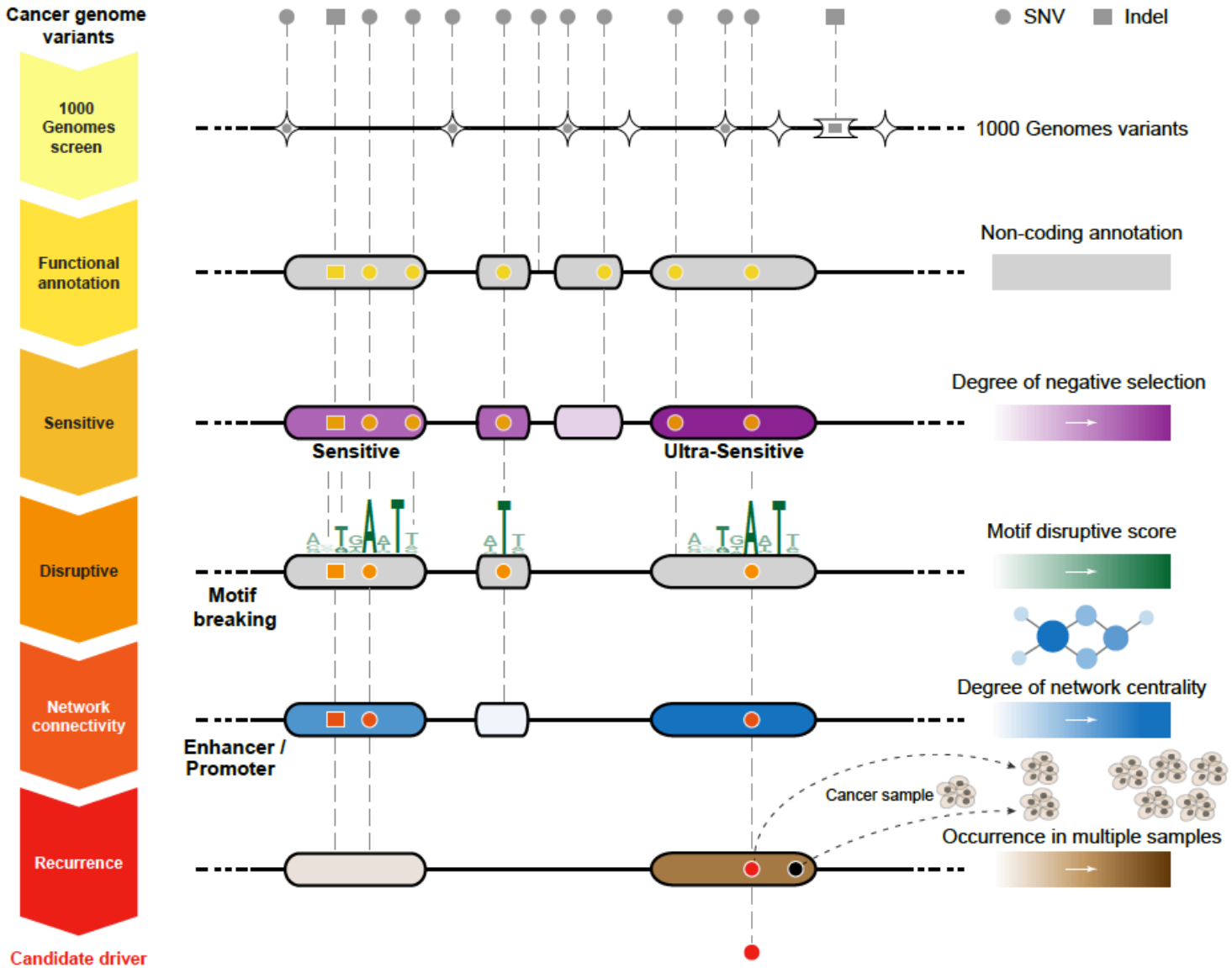
Germline vs somatic variants



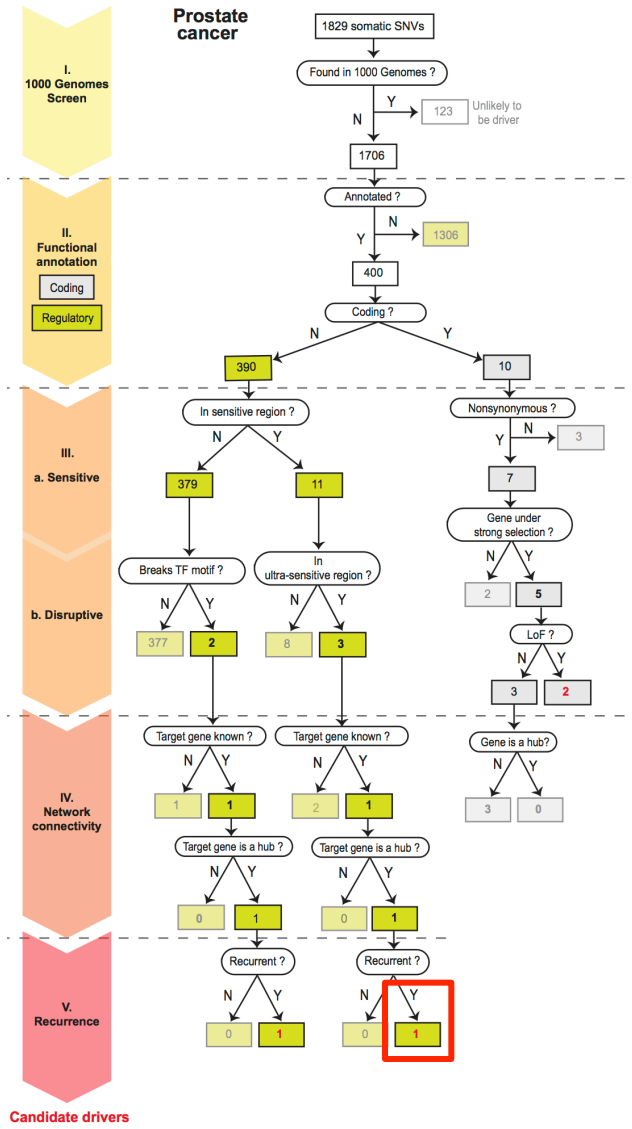
- Somatic mutations do not follow patterns of natural polymorphisms
- Those deviating the most from these patterns are most likely to be cancer drivers providing selective advantage to the tumor cells (confirmed for protein-coding genes)
- Look for mutations in elements under strong negative selection

Khurana et al., *Science*, 2013

Identification of noncoding candidate drivers among somatic variants: Scheme

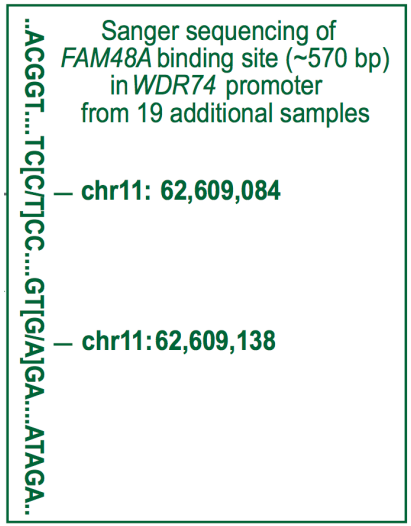


Identification of noncoding candidate drivers among somatic variants: Example

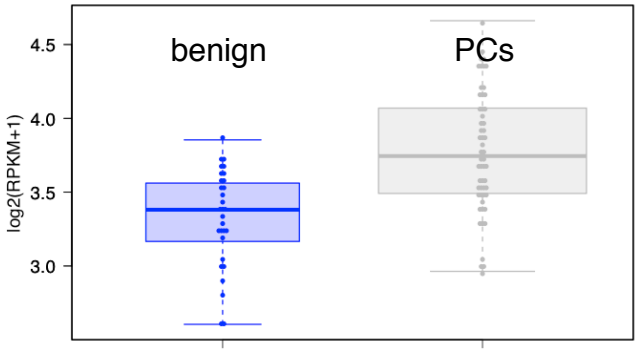


Validation of a candidate driver identified in prostate cancer sample in *WDR74* gene promoter

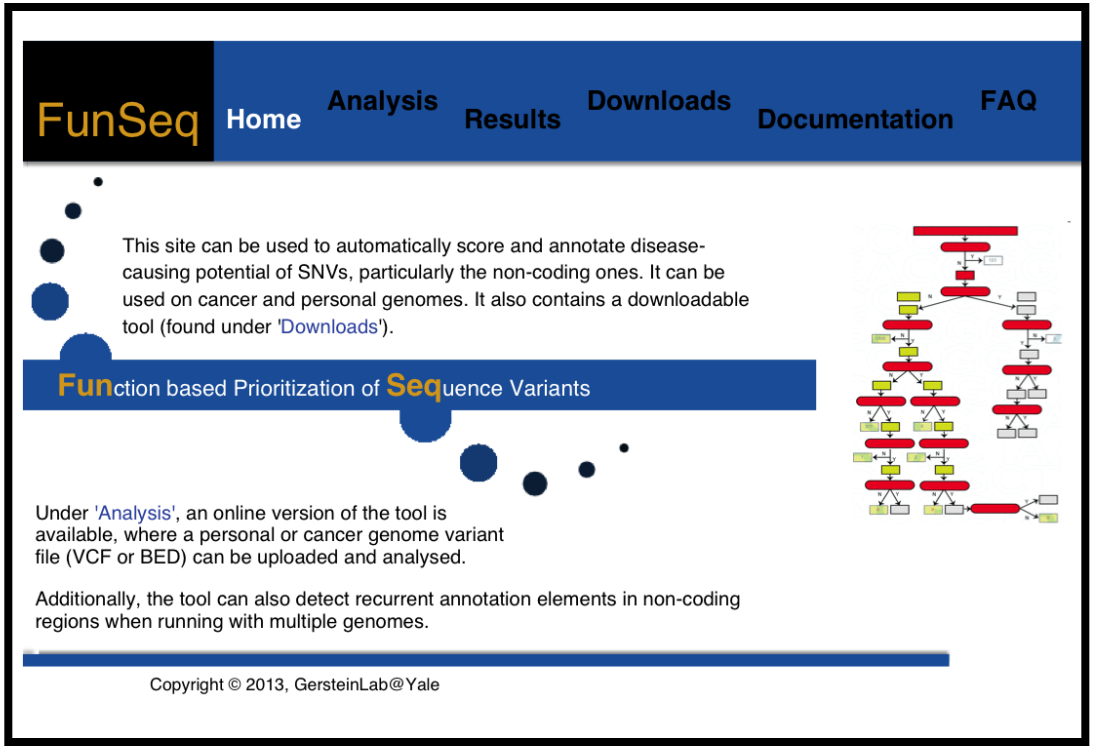
- Sanger sequencing in 19 additional samples confirms the recurrence in one more sample



- *WDR74* shows increased expression in tumor samples



FunSeq.GersteinLab.org : webserver & code download



Acknowledgements

~40 Institutes
~550 participants

Functional Interpretation Group

~50 participants

Yale
Yao Fu, Xinmeng Mu, 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, **Mark A. Rubin**

U of Michigan
Hyun Min Kang

U of Geneva
Tuuli Lappalainen, 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