FunSeq: Computational identification of cancer drivers from whole-genome sequencing data, using ENCODE functional annotations

Mark Gerstein

Yale

General Motivation:
Identifying damaging non-coding mutations

- Control elements for coding genes
- Most GWAS hits & many rare disease-causing mutations occur in regulatory regions

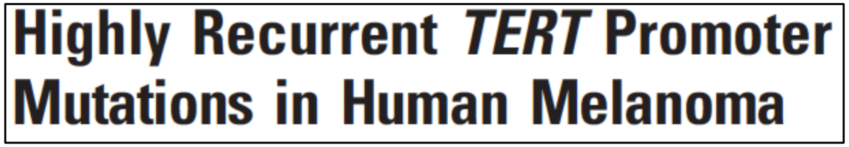
Encode integrative paper, Nature, 2012; Maurano et al, Science, 2012; Ward et al, Nature Biotech, 2012

- Most personal genome variants are non-coding
- Unlike for coding variants, no standard approaches exist to prioritize non-coding variants

- Similar thought process to GWAS group, but....

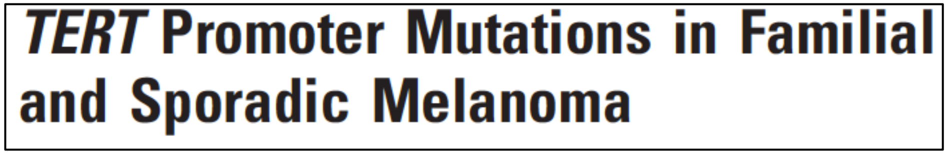
Most Cancer Mutations are Non-coding

- ~99% of somatic SNVs occur in non-coding regions, including TFBSs, ncRNAs and pseudogenes
- Nevertheless, cancer sequencing has been very exome focused
- Publicity for TERT promotor mutation – exception proves the rule!
- Somatic mutations very different from GWAS
  - GWAS is "common variants" – e.g. expected to follow LD
  - Somatic variations are not expected to follow patterns of natural variation (e.g. no LD), so can be contrasted with them



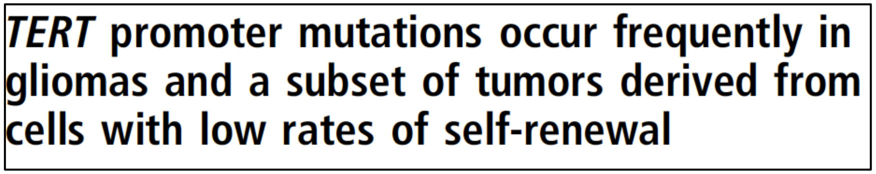
Highly Recurrent *TERT* Promoter Mutations in Human Melanoma

Science, 2013



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

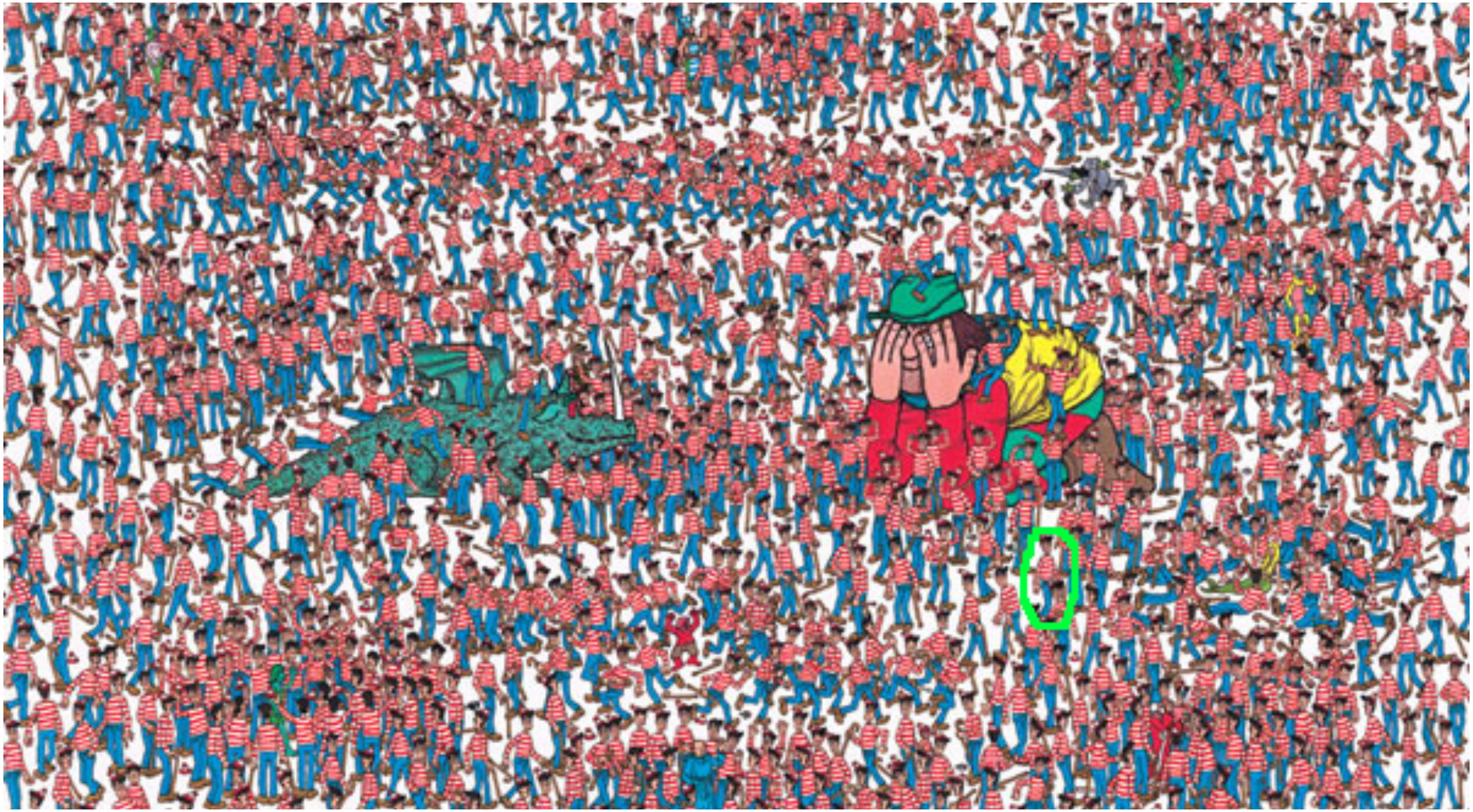
Science, 2013



***TERT* promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal**

PNAS, 2013

Where is Waldo?

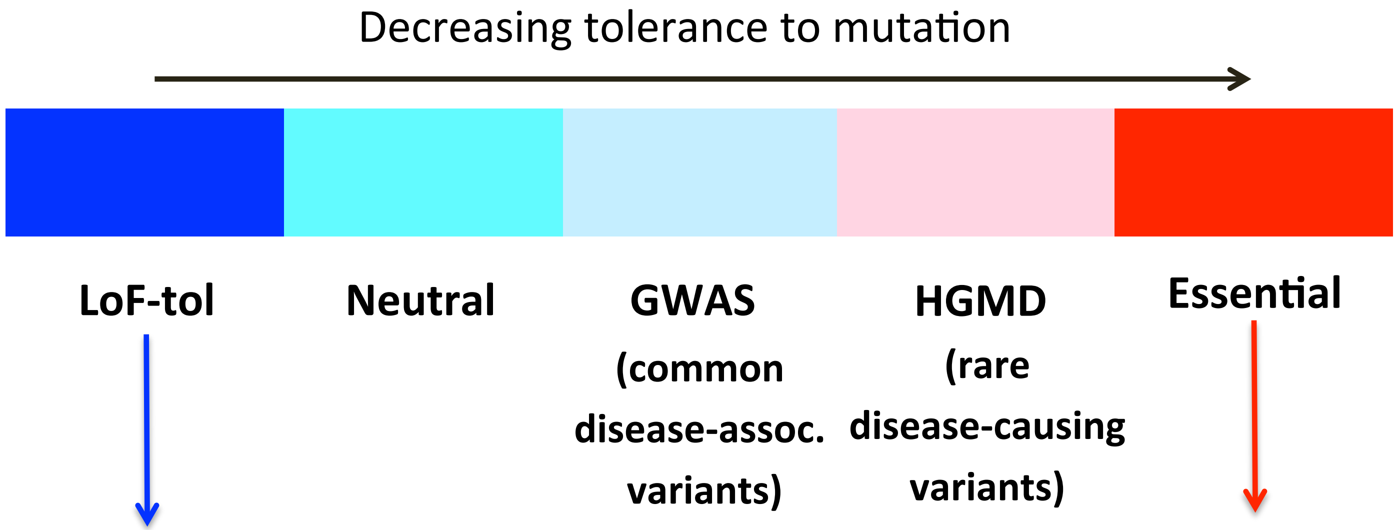


Outline

Our Approach : Use 1000G & ENCODE to characterize natural patterns of inherited variants in functional elements. Identify drivers as somatic variants breaking these patterns.

- Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
- Prioritizing based on network connectivity
- Building a workflow & software tool for cancer genomes

Gene categories with known phenotypic effects



Decreasing tolerance to mutation

LoF-tol

Neutral

GWAS (common disease-assoc. variants)

HGMD (rare disease-causing variants)

Essential

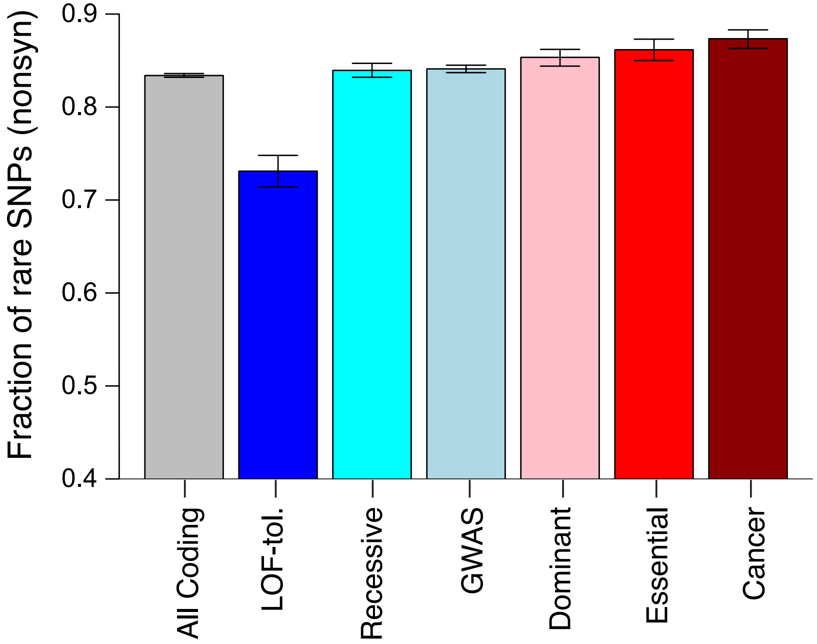
- Homozygous inactivation in at least one healthy 1000 Genomes individual
- Weak selection constraints

From MacArthur et al, Science, 2012

- Homozygous inactivation leads to clinical features of death before puberty or infertility
- Very strong selection constraints

From Liao et al, PNAS, 2008

Metric to estimate strength of negative selection amongst humans



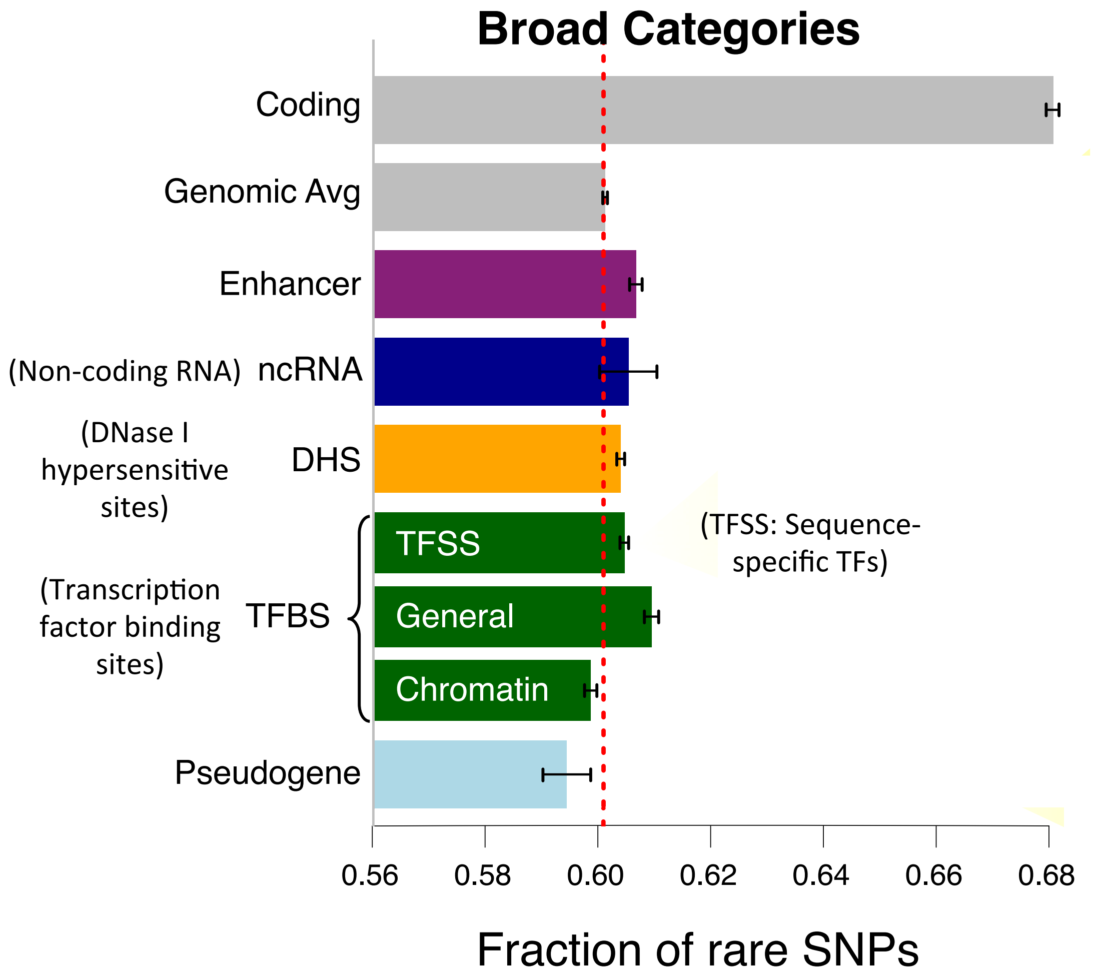
(rare=derived allele freq < 0.5%)

- SNP density, heterozygosity, **enrichment of rare SNPs**
- Negative selection restricts the allele frequency of deleterious mutations
- Results for protein-coding genes consistent with known phenotypic impacts

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

Khurana et al., *Science*, In press

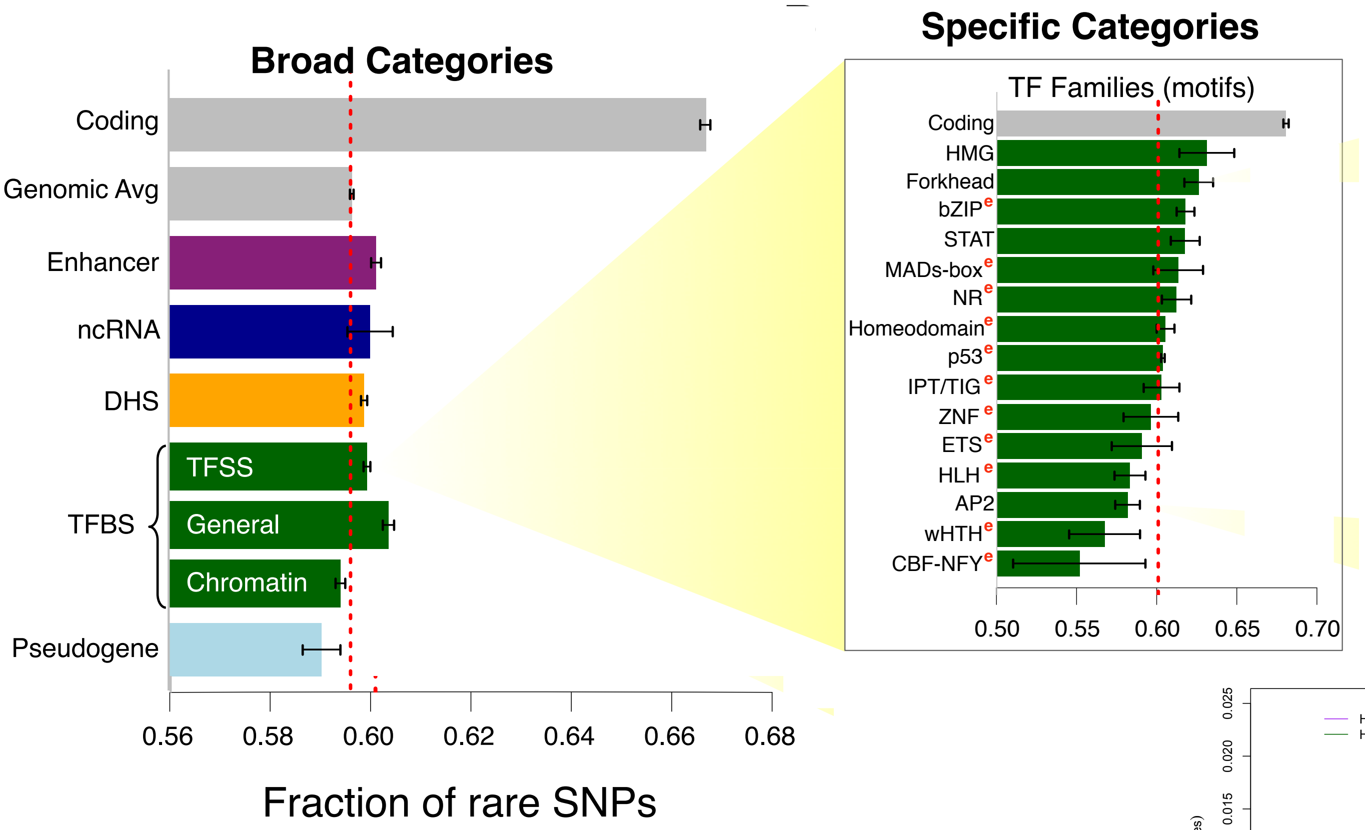
Negative selection in non-coding elements



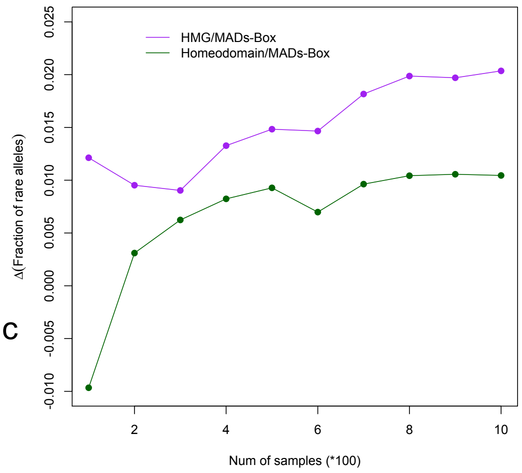
- Broad categories of regulatory regions under negative selection
- Consistent with previous studies

ENCODE, *Nature*, 2012
Ward & Kellis, Science, '12

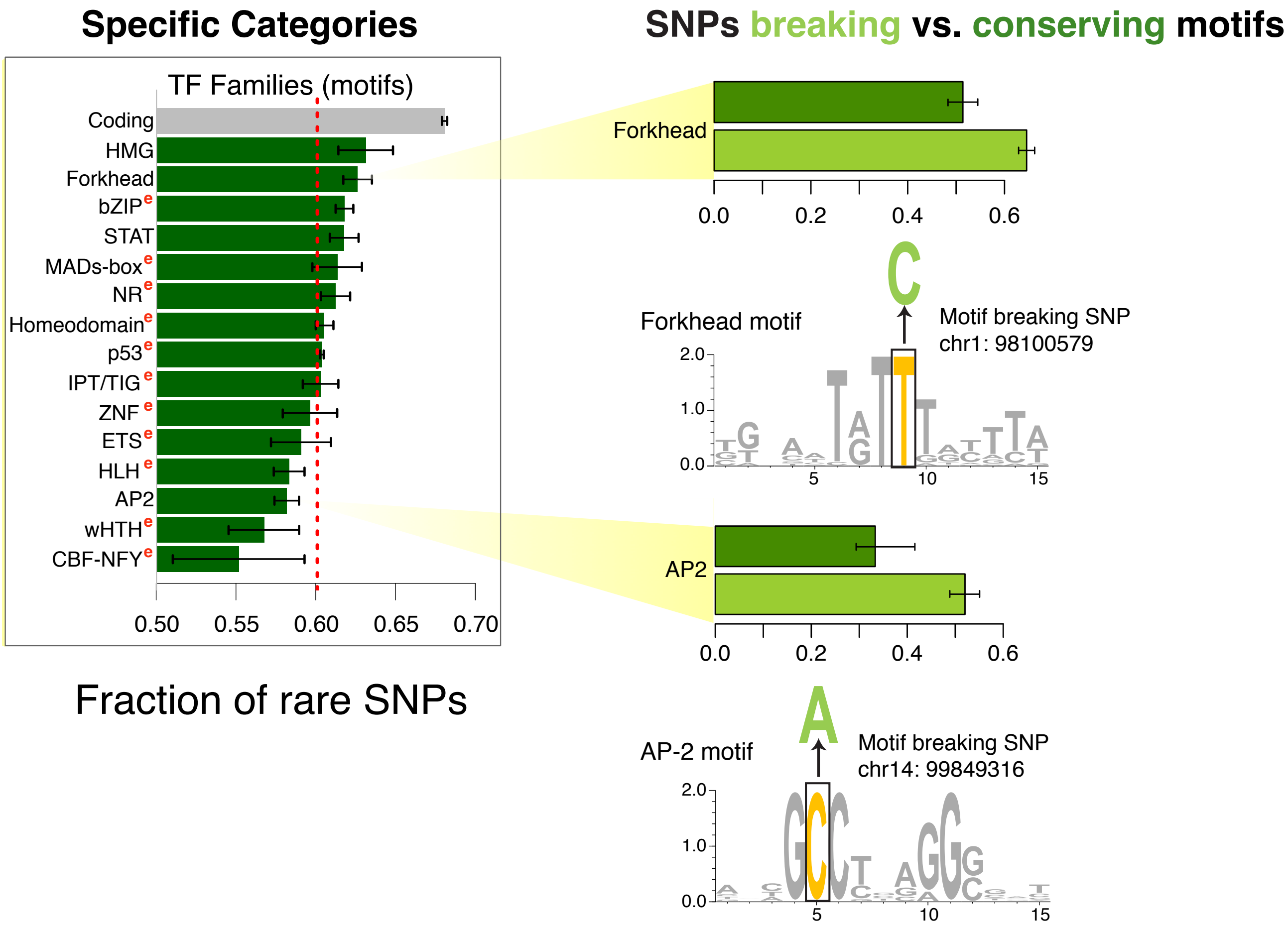
~700 specific sub-categories of broad non-coding categories; Possible to study now using 1000G Phase 1



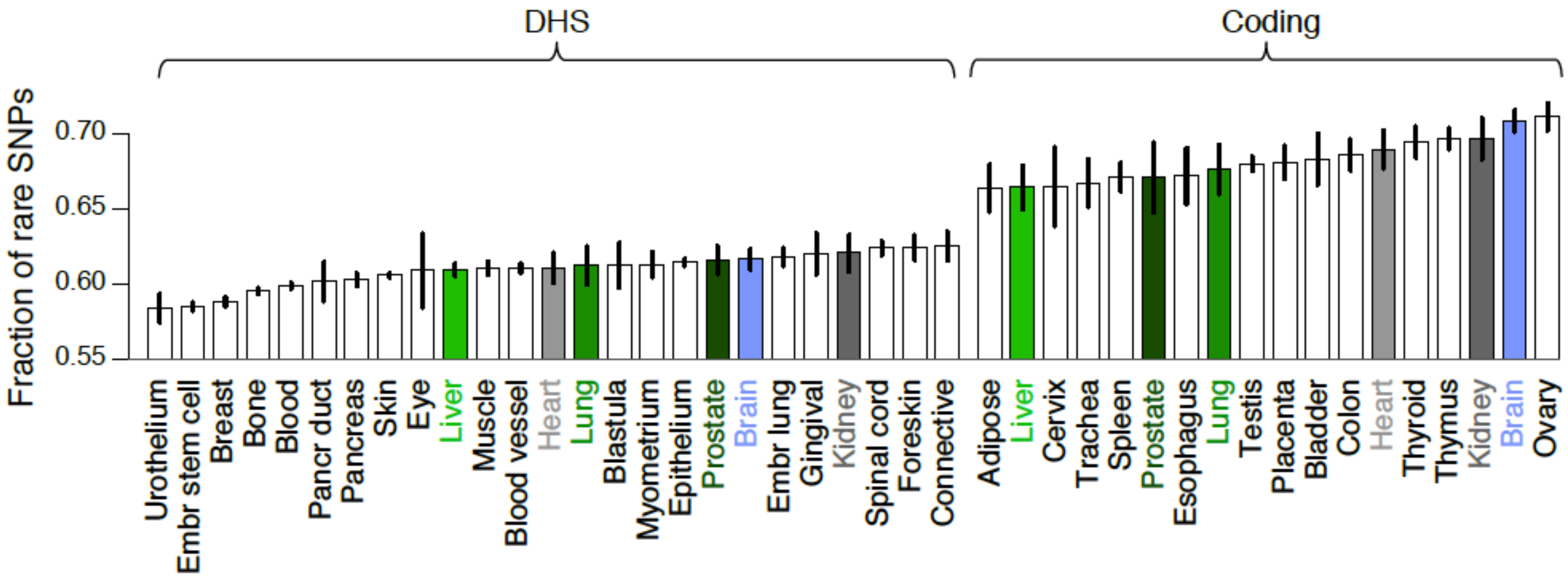
- ❑ ~ 700 specific non-coding categories
  - ❑ ncRNA: snRNA, snoRNA, miRNA, lincRNA ***
  - ❑ Motifs & binding sites of different TF families
  - ❑ TFBSs divide into proximal vs distal and cell-line–specific vs –non-specific
- ❑ Large sample size:1,092 humans compared to pilot ~180



SNPs which break TF motifs are under stronger selection



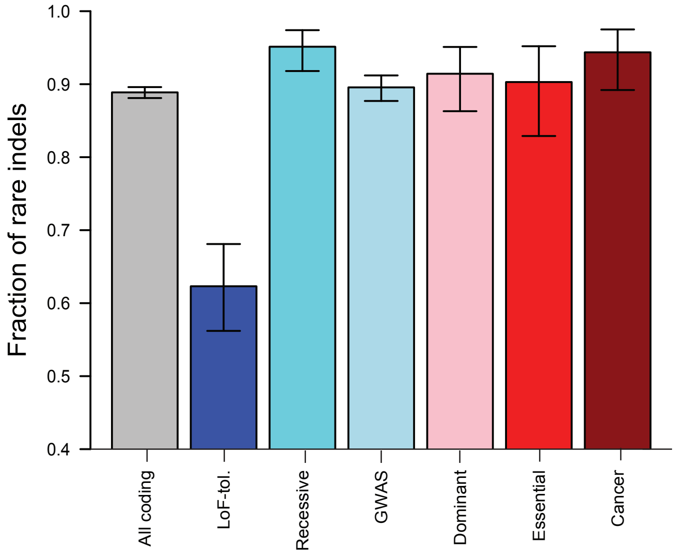
Negative selection and tissue-specificity of coding and non-coding regions



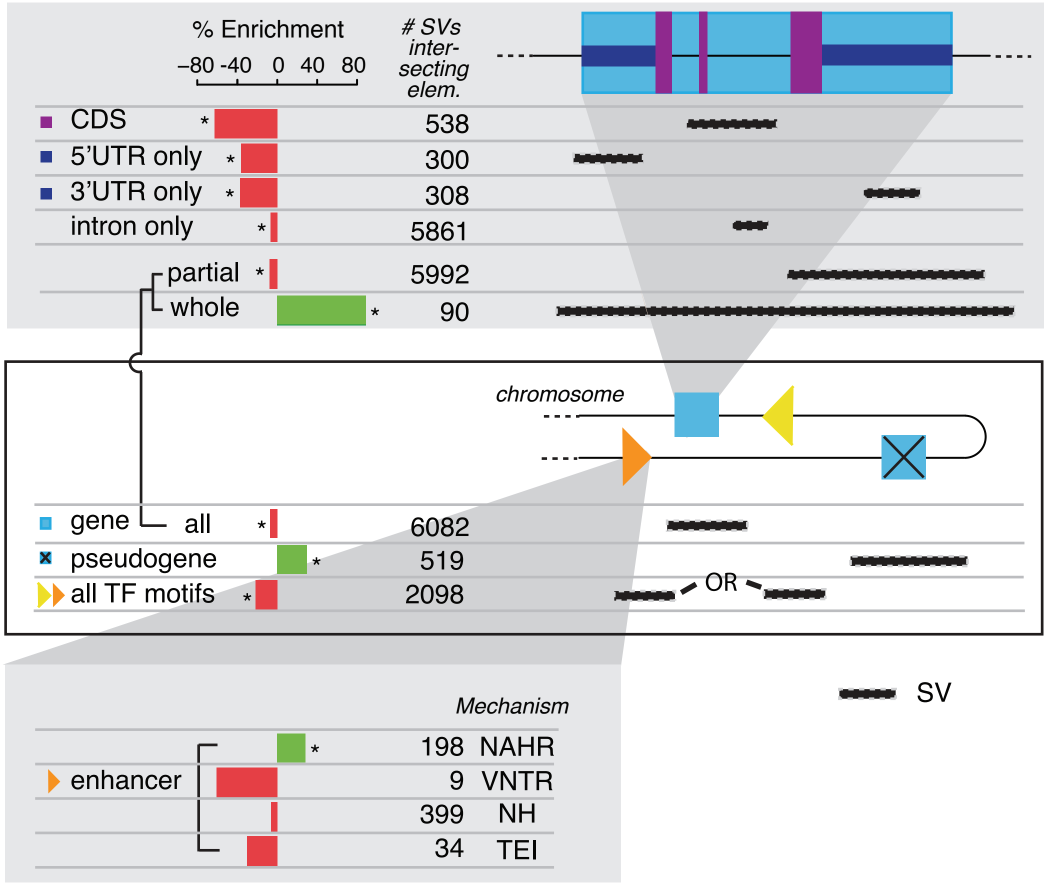
- ❑ Ubiquitously expressed genes and bound regions show stronger selection
- ❑ Differences in constraints amongst tissues
- ❑ Constraints in coding genes and regulatory genes are correlated across tissues

Functional annotation of indels and larger structural variants

Indels show similar patterns as SNPs

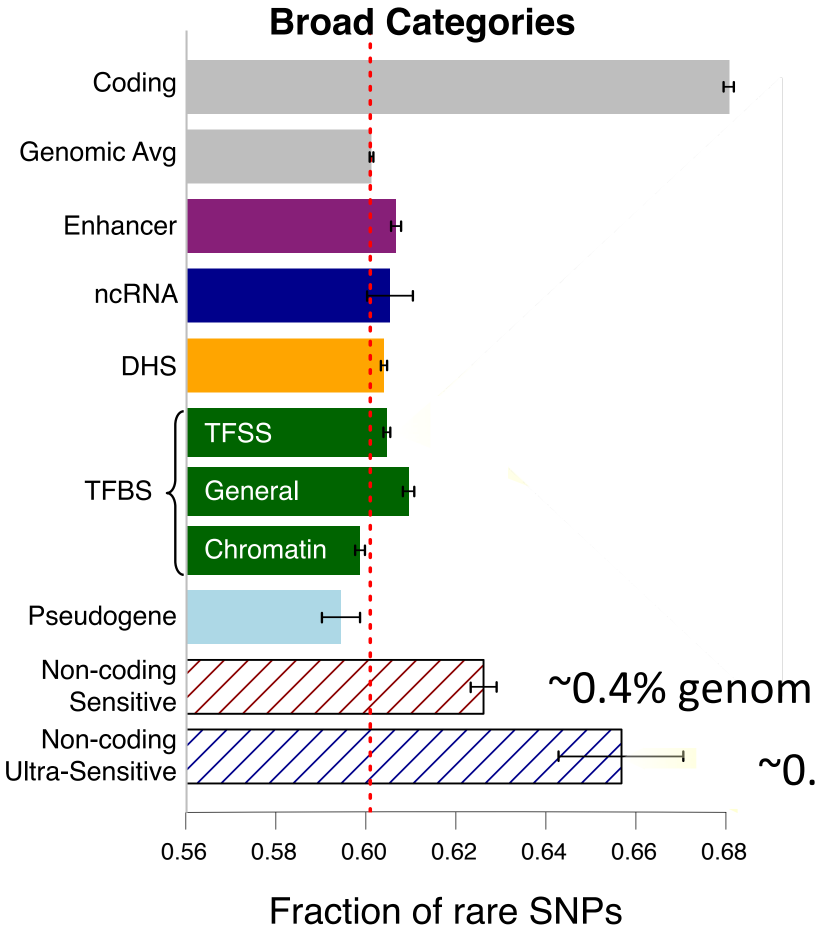


Structural variants are generally depleted for functional elements



🟦 gene ▮ CDS ▬ UTR ☒ pseudogene ▶ TF motif ▶ enhancer ▬ SV

Can we identify which non-coding elements are under very strong "coding-like" selection ?



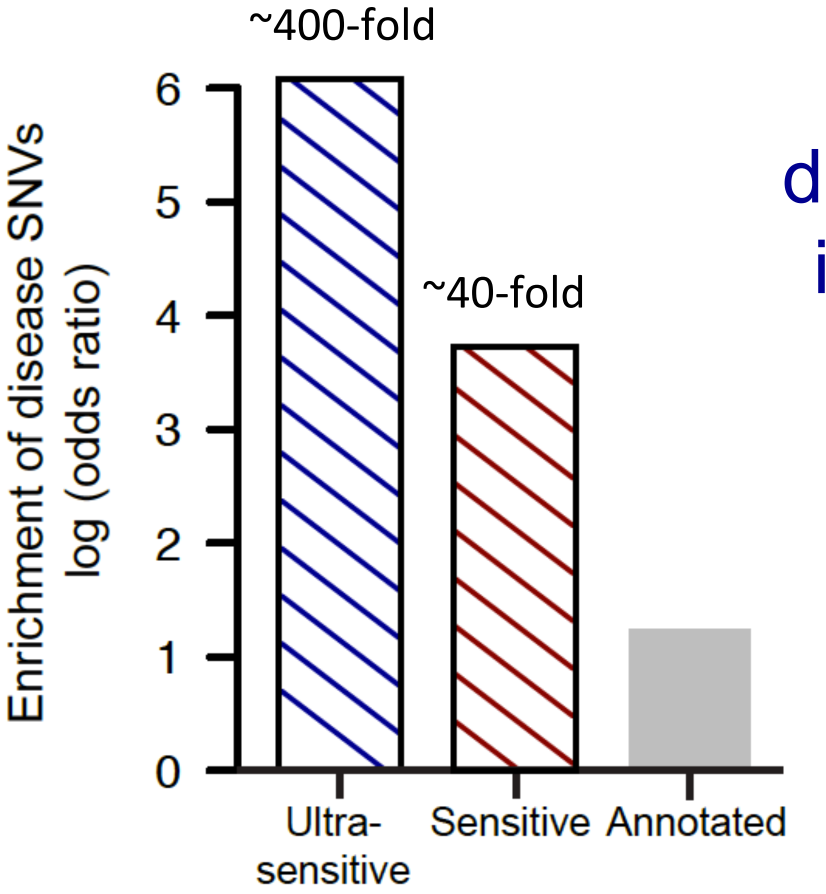
~0.4% genomic coverage

~0.02% genomic coverage

- ❑ Start 677 high-resolution non-coding categories; Rank & find those under strongest selection
- ❑ Pick useful subsets of these – e.g. a similar fraction to exome & Top-5 -- to define sensitive & ultra-sens.

- ❑ Binding peaks of some general TFs (eg *FAM48A*)
- ❑ Core motifs of some TF families (eg *JUN, GATA*)
- ❑ Proximal but not distal sites of ZNF274

Enrichment of know disease-causing mutations from Human Gene Mutation database validates functional indispensability of sensitive and ultra-sensitive regions



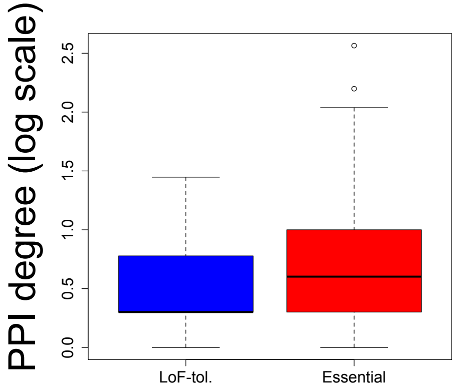
Outline

Our Approach : Use 1000G & ENCODE to characterize natural patterns of inherited variants in functional elements. Identify drivers as somatic variants breaking these patterns.

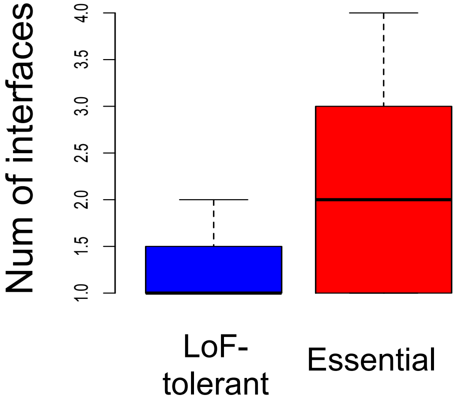
- Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
- Prioritizing based on network connectivity
- Building a workflow & software tool for cancer genomes

Gene essentiality and protein-protein interaction network

Essential genes



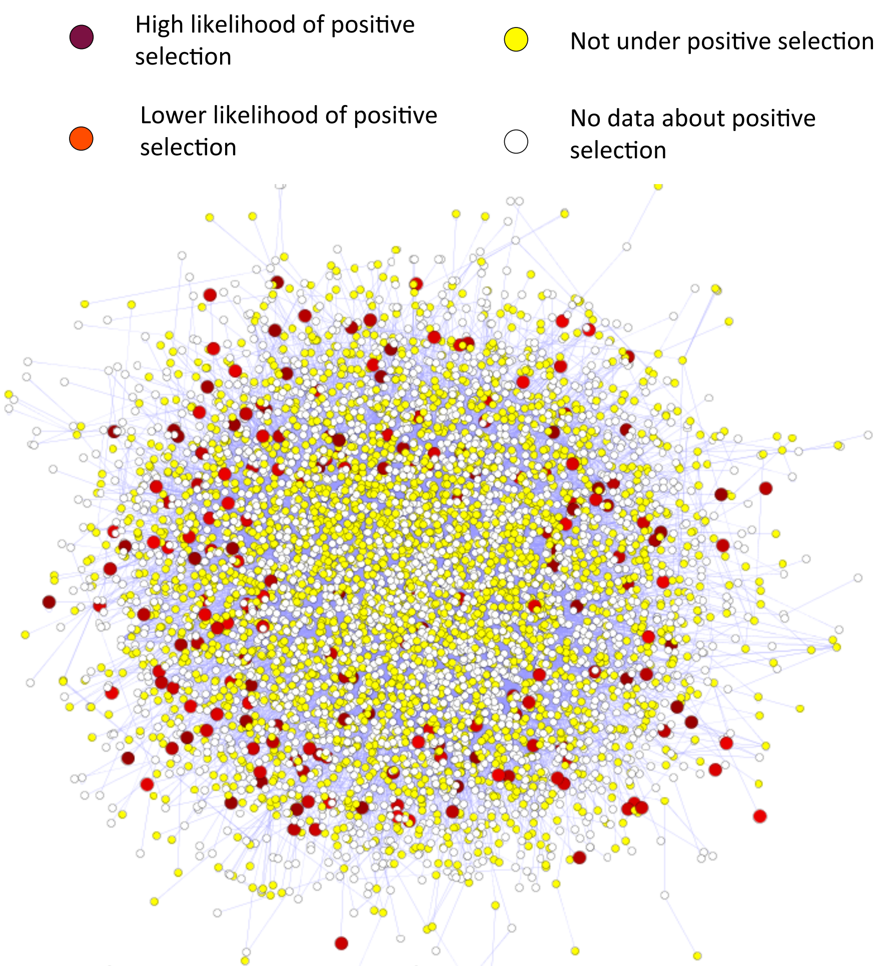
Higher Centrality



More interaction interfaces

Khurana et al., *PLoS Comp. Bio.*, 2013
Wang et al, *Nature Biotech*, 2012

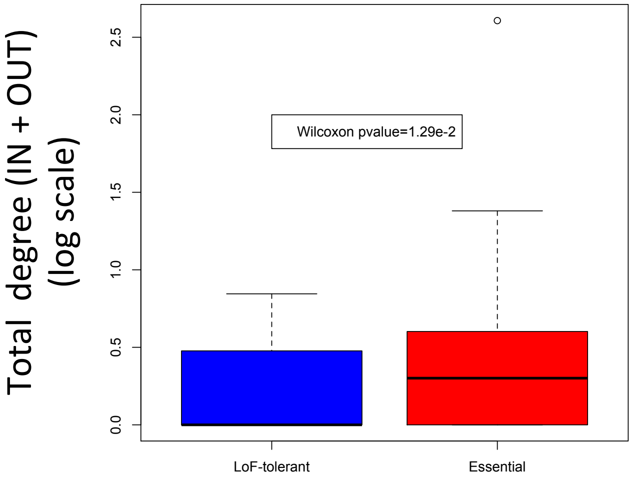
More Connectivity, More Constraint : A theme borne out in many studies


High likelihood of positive selection
Not under positive selection
Lower likelihood of positive selection
No data about positive selection

[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. PNAS (2007)]

Similar Results for ENCODE Human Regulatory Network to PPI

- Essential genes tend to be central



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

TF **target in-degree**

Neg. corr. with

(SCC=-.2, P<0.5)

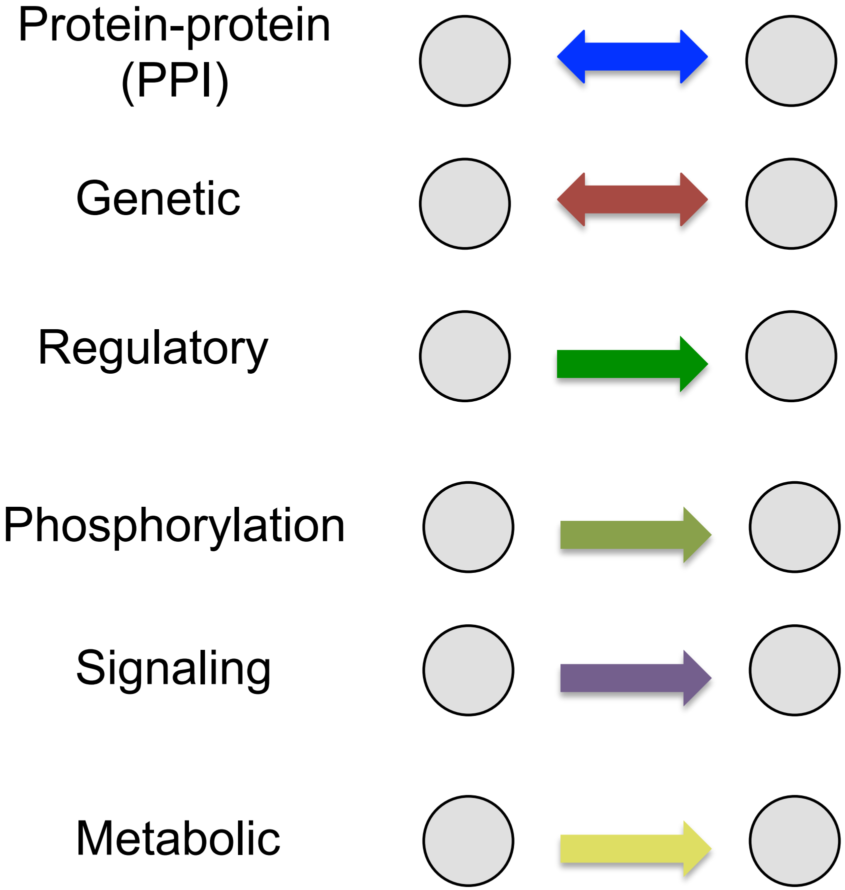
dN/dS
(from chimp alignments)

TF **target in-degree**
&
TF out-degree

Neg. corr. with

ns SNP density, pN/pS, avg. DAF

Genes interact using many different modes



Protein-protein (PPI)

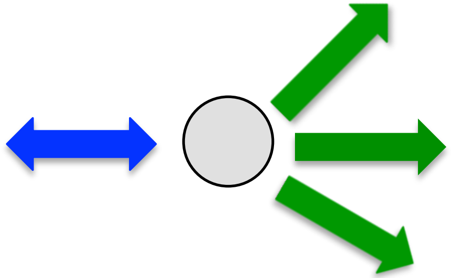
Genetic

Regulatory

Phosphorylation

Signaling

Metabolic



E.g. *SIX5*

□ Interacts with one protein

□ Regulates 360 genes

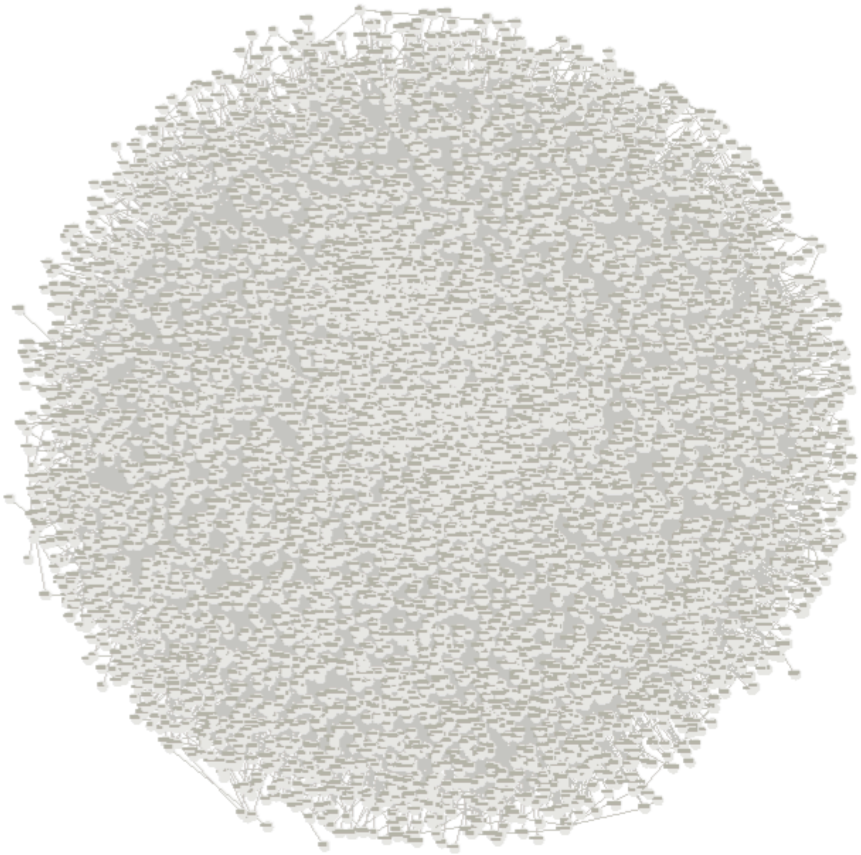
□ HGMD, Branchio-oto-renal syndrome

Hoskins et al, AJHG, 2007

Multinet – the ultimate hairball!

Genes participate in many networks and no single network captures the global picture of gene interactions

Combine **regulatory** interactions with other networks : **physical protein-protein**, **signaling**, **metabolic**, **phosphorylation** and **genetic** to create a **unified network (Multinet)**

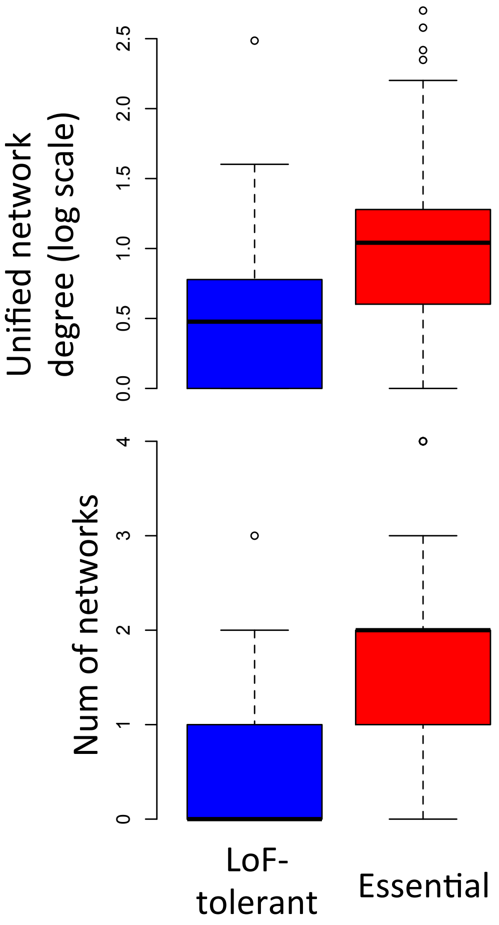


Nodes: ~15,000 genes
Edges: ~110,000 interactions

Edges shown in gray

[Khurana et al., *PLOS Comp. Bio.* '13]

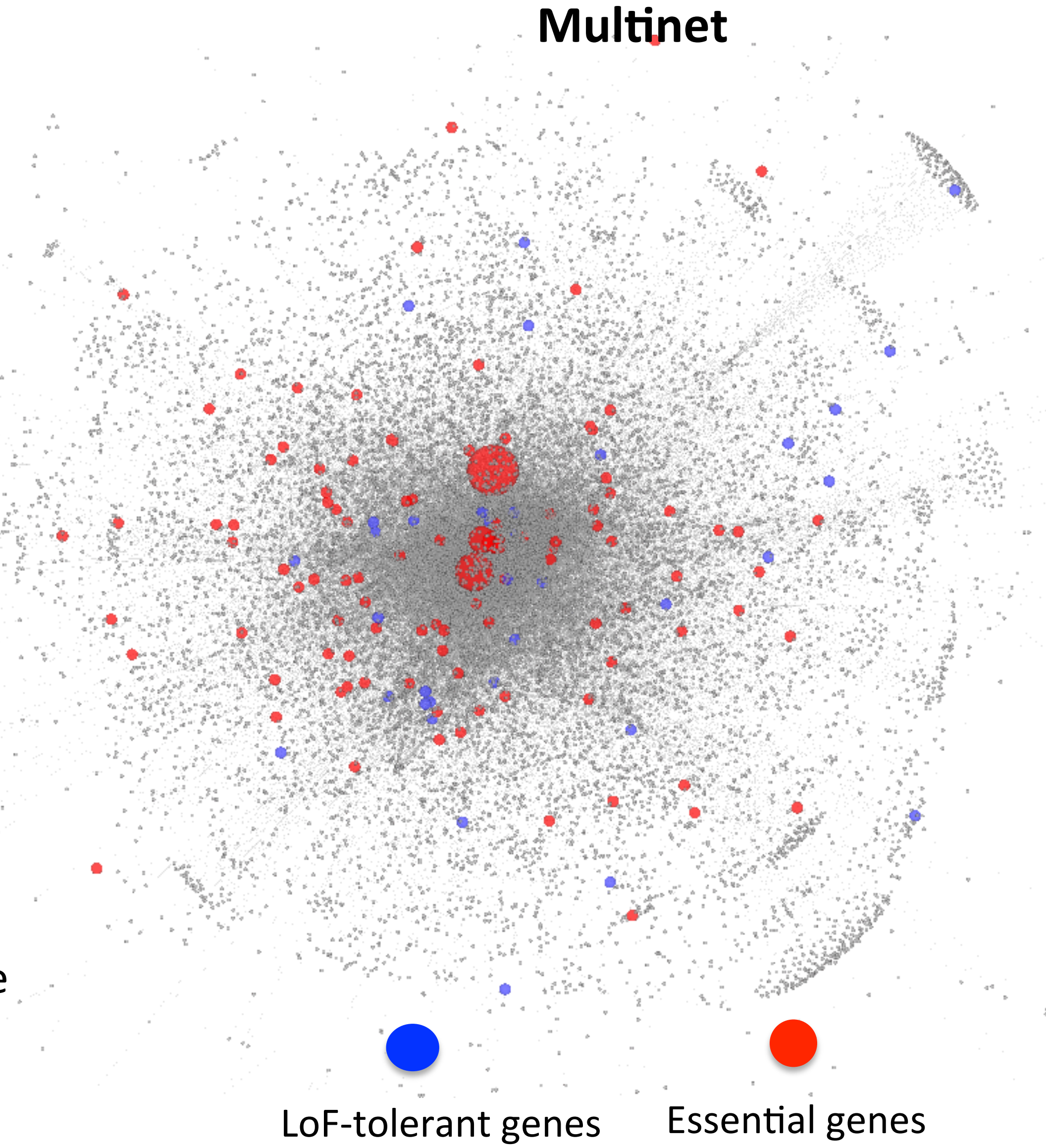
Gene properties in Multinet



Essential genes are

Connected to more genes

Involved in more networks



Z Gumus iCAVE movie

Size of nodes scaled by total degree

[Khurana et al., *PLOS Comp. Bio.* '13]

Outline

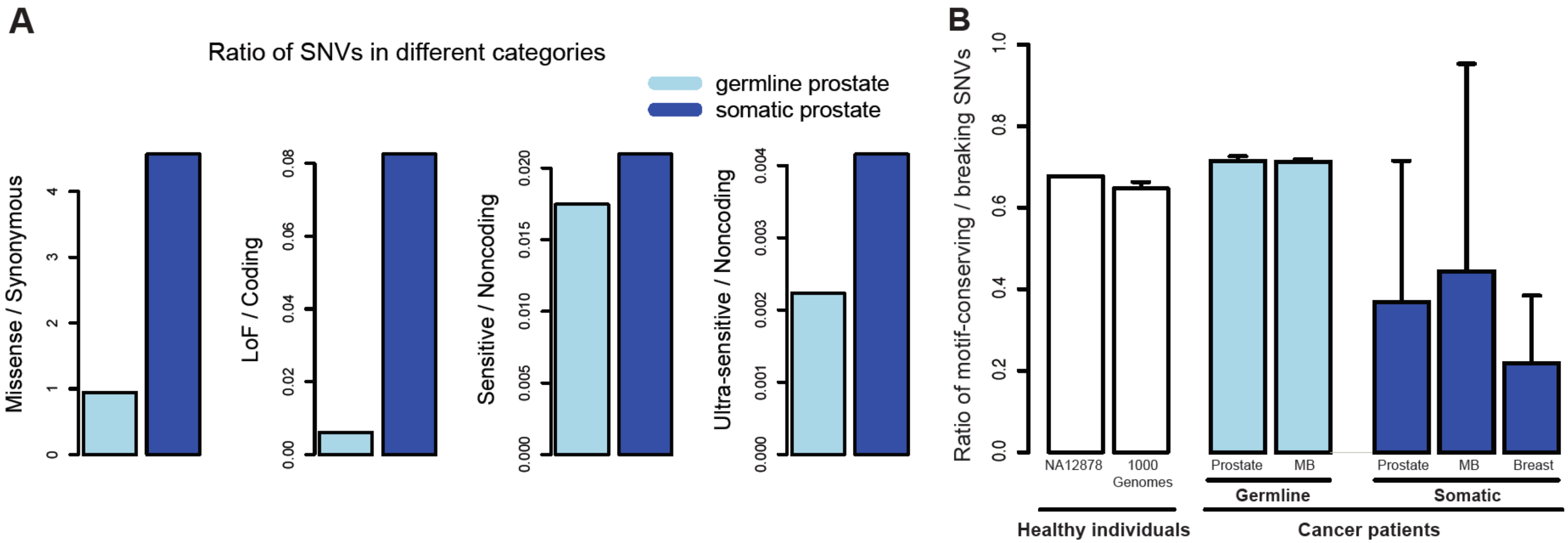
Our Approach : Use 1000G & ENCODE to characterize natural patterns of inherited variants in functional elements. Identify drivers as somatic variants breaking these patterns.

- Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
- Prioritizing based on network connectivity
- Building a workflow & software tool for cancer genomes

We have learned for non-coding regions…..

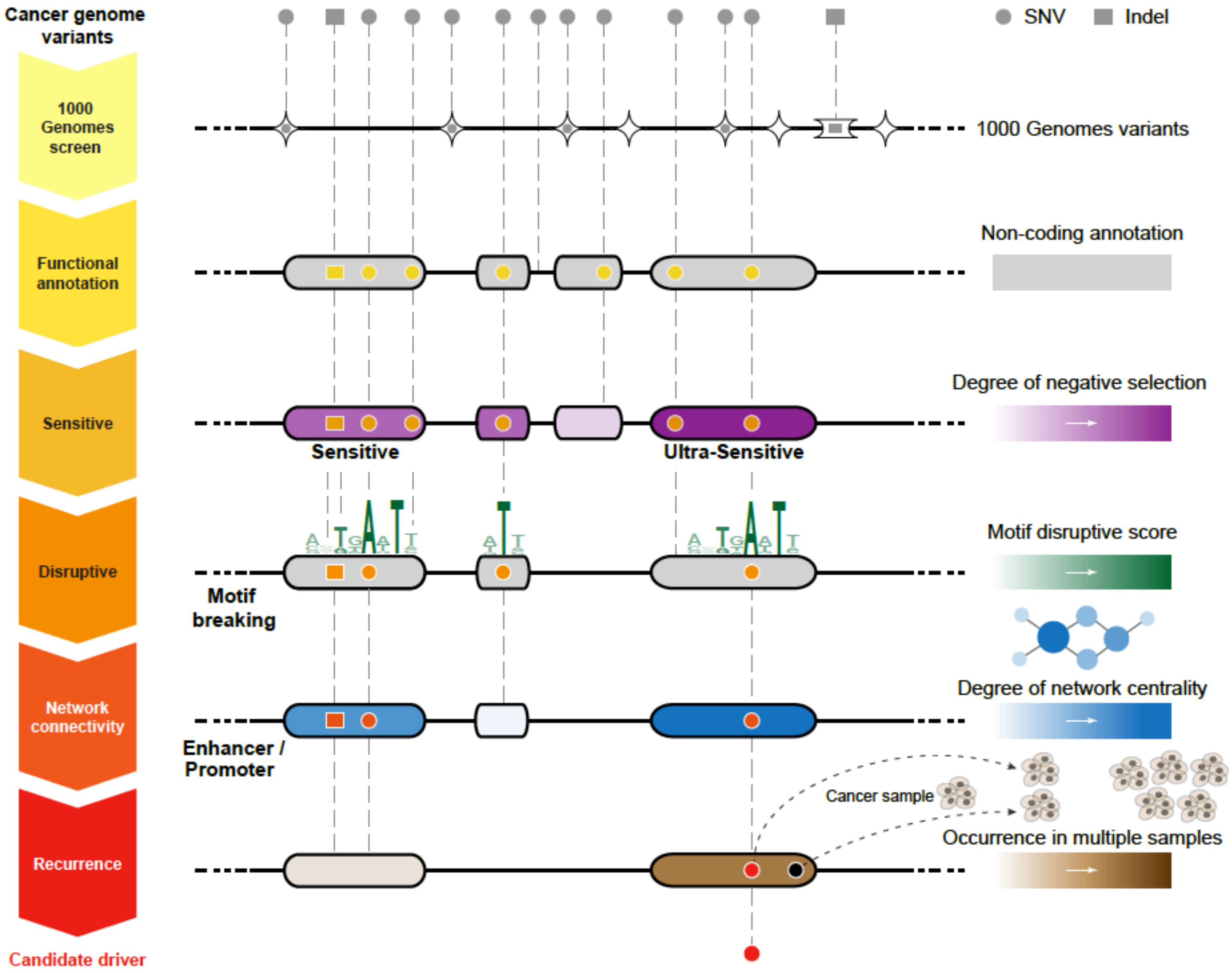
- Ultra-sensitive and sensitive regions are under strong selection
- Variants which break TF motifs are selected against
- Variants in promoters or enhancers of highly connected genes are selected against
- Can we combine all these features to prioritize damaging non-coding variants?

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

Identification of non-coding candidate drivers amongst somatic variants: Scheme



Identification of non-coding candidate drivers amongst somatic variants: Examples

Identified ~100 non-coding driver mutations

- 64 prostate cancer samples
- 21 breast cancer samples
- 3 medulloblastoma samples

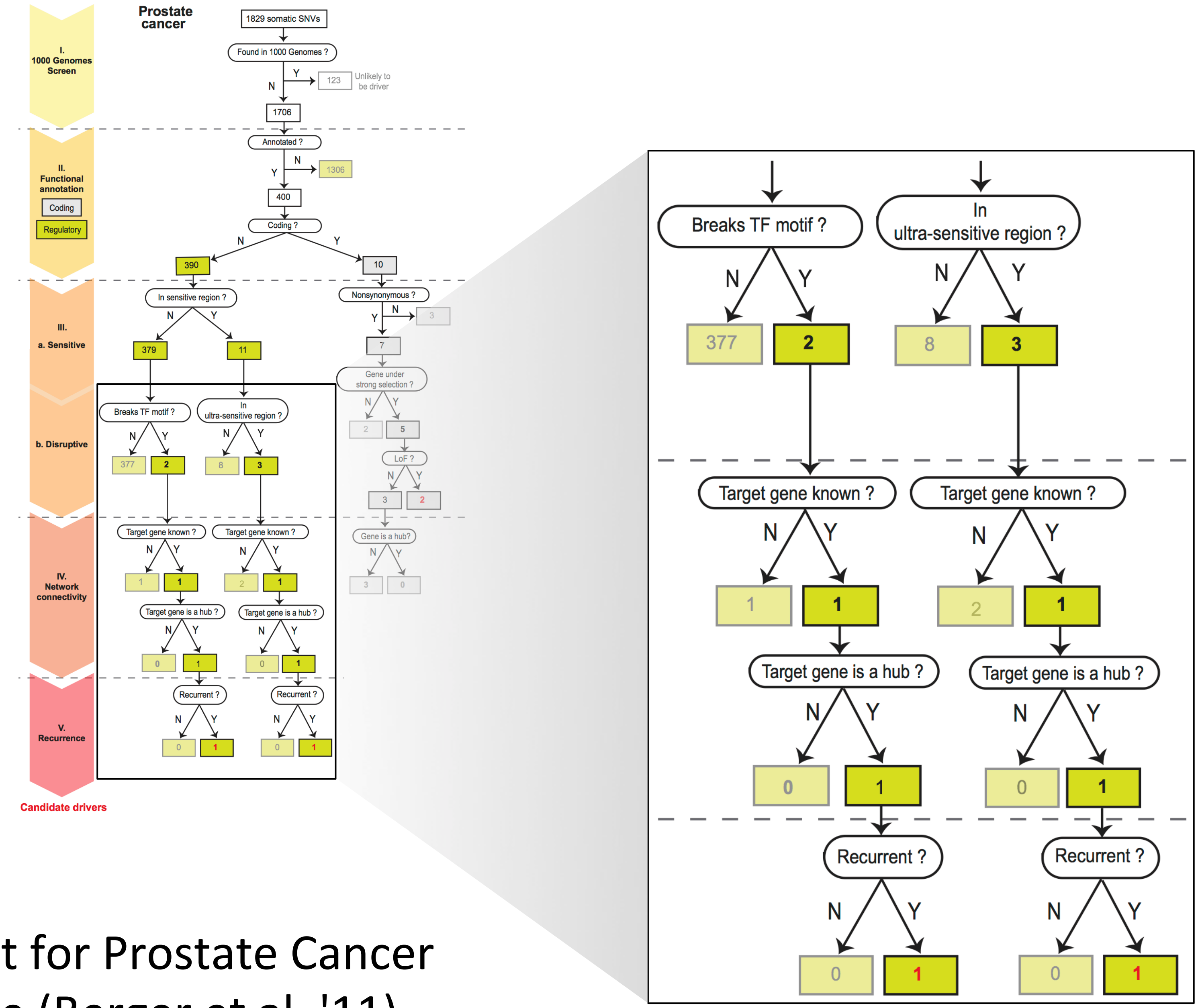
Data sets:

Berger et al, Nature, 2011;

Baca et al, Cell, 2013;

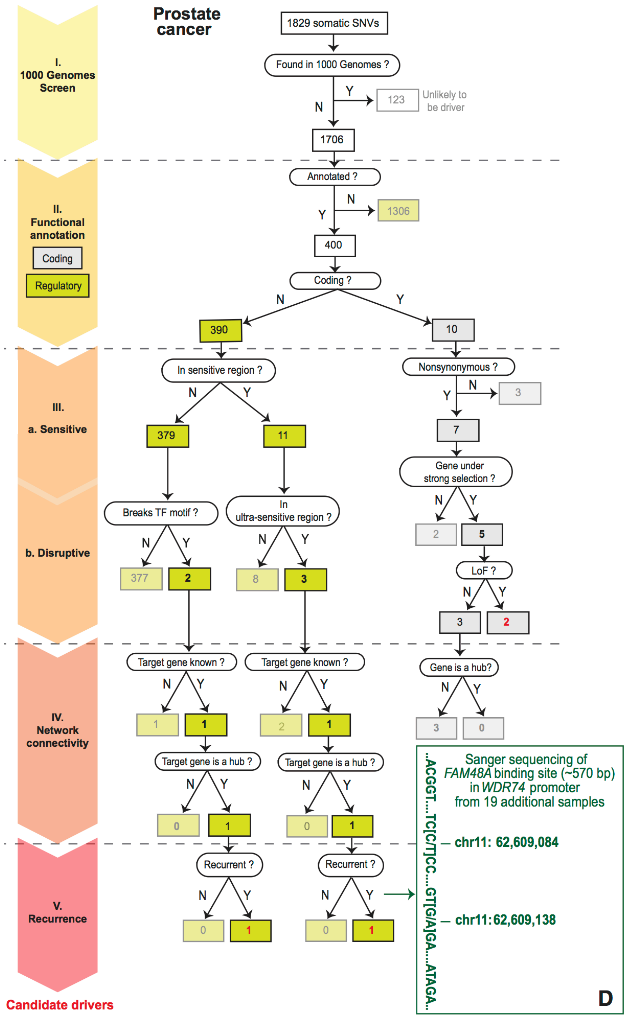
Rausch et al, Cell, 2012;

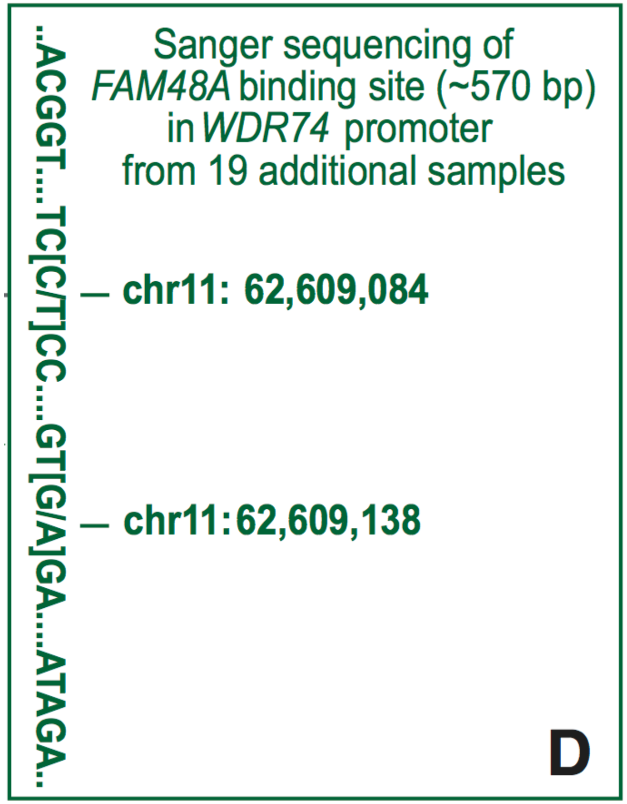
Nik-Zainal et al, Cell, 2012

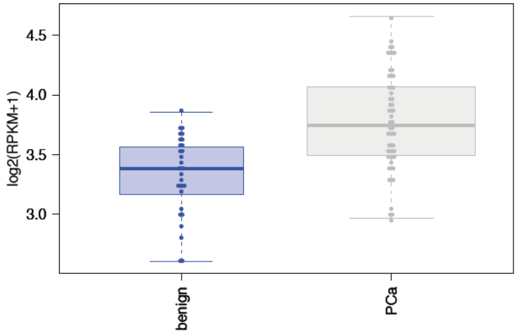


Flowchart for Prostate Cancer Genome (Berger et al. '11)

Validation of a candidate




...ACGGT...TC[C/T]CC...GT[G/A]GA...ATAGA...
Sanger sequencing of
FAM48A binding site (~570 bp)
in WDR74 promoter
from 19 additional samples
— chr11: 62,609,084
— chr11:62,609,138
D



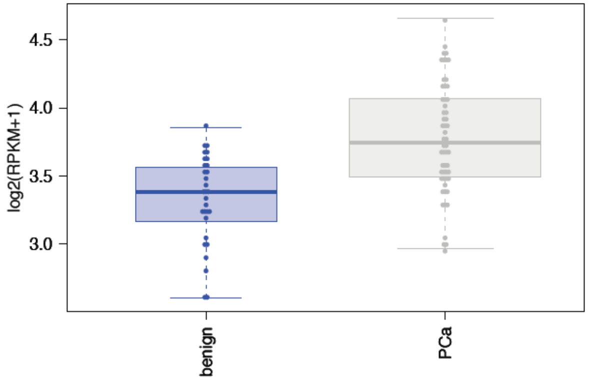
WDR74 shows increased expression in tumor samples

Identification of non-coding candidate drivers amongst somatic variants: Examples

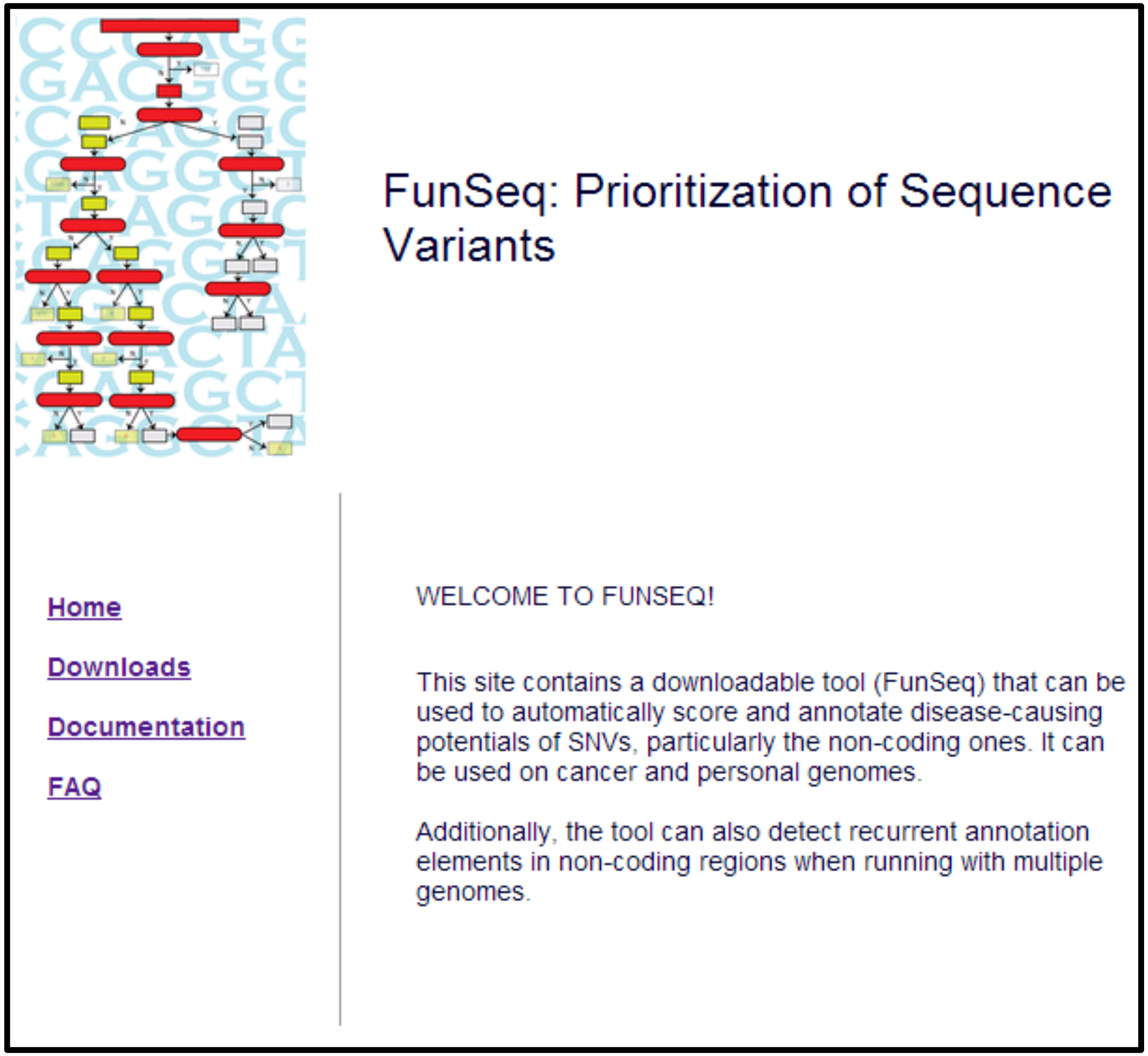
Validation of a candidate driver identified in prostate cancer sample in *WDR74* gene promoter

❑ Sanger sequencing in 19 additional samples confirms the recurrence

❑ *WDR74* shows increased expression in tumor samples



FunSeq.GersteinLab.org : webserver & code download



FunSeq: Prioritization of Sequence Variants

Home

Downloads

Documentation

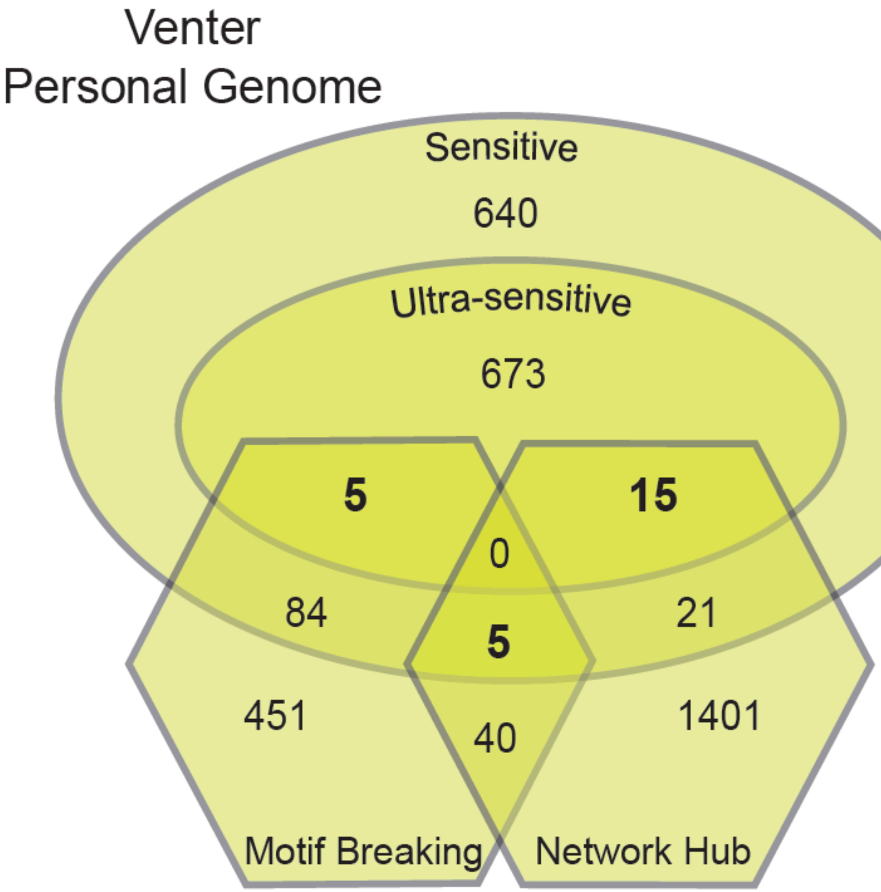
FAQ

WELCOME TO FUNSEQ!

This site contains a downloadable tool (FunSeq) that can be used to automatically score and annotate disease-causing potentials of SNVs, particularly the non-coding ones. It can be used on cancer and personal genomes.

Additionally, the tool can also detect recurrent annotation elements in non-coding regions when running with multiple genomes.

Can also use FunSeq to prioritize non-coding variants in personal genomes



Out of a total of ~3 million non-coding variants, 25 highly likely to be deleterious

Outline

Our Approach : Use 1000G & ENCODE to characterize natural patterns of inherited variants in functional elements. Identify drivers as somatic variants breaking these patterns.

- Finding ultra-sensitive non-coding regions & disruptive mutations (eg motif breakers)
- Prioritizing based on network connectivity
- Building a workflow & software tool for cancer genomes

Acknowledgements

Yale University

Ekta Khurana, Yao Fu, Xinmeng Mu,
Jieming Chen, Lucas Lochovsky, Arif Harmanci,
Alexej Abyzov, Suganthi Balasubramanian, Cristina
Sisu, Declan Clarke

Wellcome Trust Sanger Institute

Vincenza Colonna, Yali Xue, Chris Tyler-Smith

Weill Cornell Medical College

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

Cornell University, Ithaca

Jishnu Das, Robert Fragoza, Xiaomu Wei, Haiyuan
Yu

US, UK, Switzerland

Hyun Min Kang, Tuuli Lappalainen, Kathryn
Beal, Daniel Challis, Yuan Chen, Laura
Clarke, Fiona Cunningham, Emmanouil T.
Dermitzakis, Uday Evani, Paul Flicek, Erik
Garrison, Richard Gibbs, Javier Herrero,
Yong Kong, Kasper Lage, Daniel G.
MacArthur, Gabor Marth, Donna Muzny,
Tune H. Pers, Graham R. S. Ritchie, Jeffrey
A. Rosenfeld, Michael Wilson, Fuli Yu