FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer

Fu et al, 2014 Genome Biology **15**:480

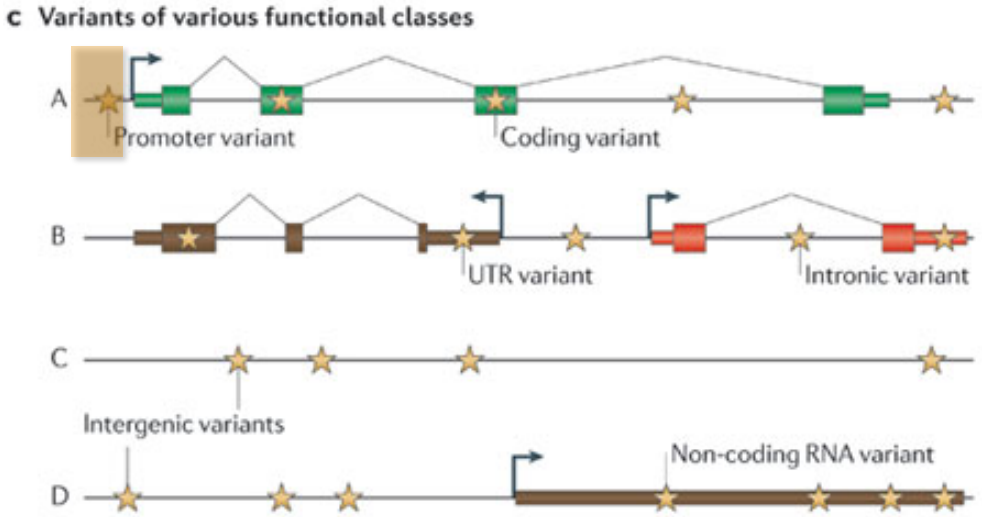
Critical paper review by
Jasleen Grewal

The hunt is on…

- DNA damage is at the root of
  - Hereditary diseases
  - Cancer formation and evolution
- Millions of single nucleotide polymorphisms
- (Hundreds of) thousands of single nucleotide variants per tumour sample
- Average mutation rate per normal cell division = 1.1
- Acquisition of mutations
  - Primary (driver)
  - Passenger
- Looking for causal variants is not an easy job!

But it's a proteomics world!

- Mutations can occur in
  - Protein coding regions
  - Non-protein coding regions
    - **98% of our genome!**
    - Intronic
    - **Regulatory (bind Transcription Factors)**
    - ncRNA
    - Pseudogenes, repeats, telomeres
- ENCODE results: >**80%** of our DNA has biochemical activity



1. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* **12**, 628-640 (September 2011) | doi: 10.1038/nrg3046

NGS data = lots of data

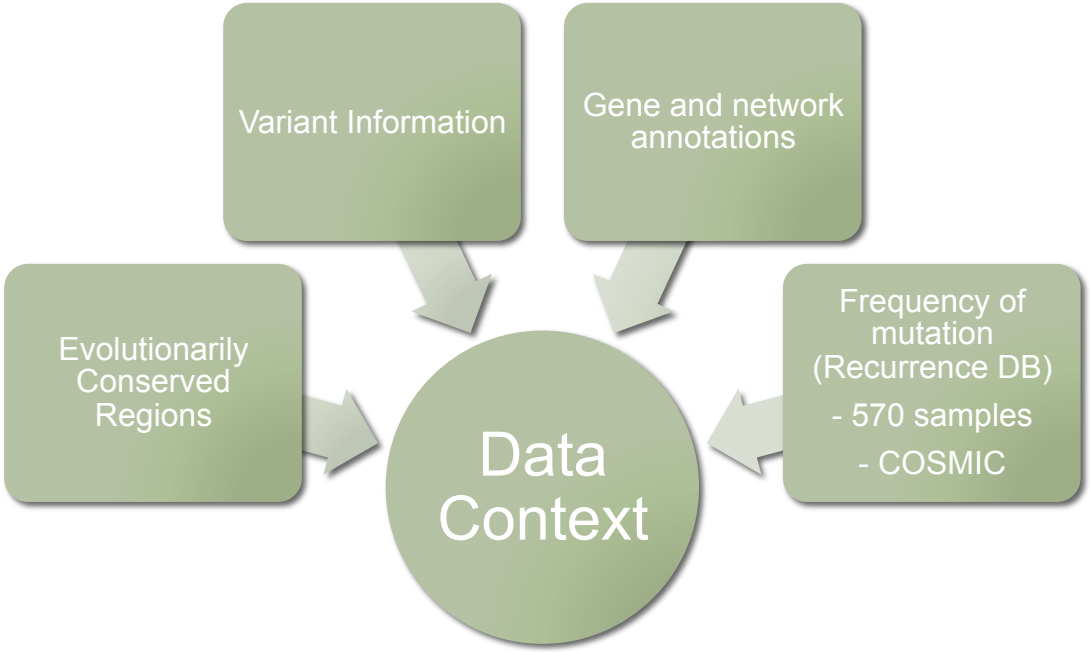
- Prioritizing variants is important and directly relevant to the study of genetic diseases with a high mutational load
- Protein-coding mutations
- ncRNAs
- Mutations in regulatory regions
  - Regulate gene expression levels
  - Immune system (MHC Class II promoters)
  - Largely ignored in GWAS studies
  - Effect prediction is challenging

Table 1 Summary of recurrence database

Cancer type	Samples (n)	Somatic mutations (SNVs) (n)
AML	7	271-1,068
Breast	119	1,043-67,347
CLL	28	522-3,338
Liver	88	1,348-25,131
Lung adeno	24	9,284-297,569
Lymphoma B cell	24	1,502-37,848
Medulloblastoma	100	44-47,440

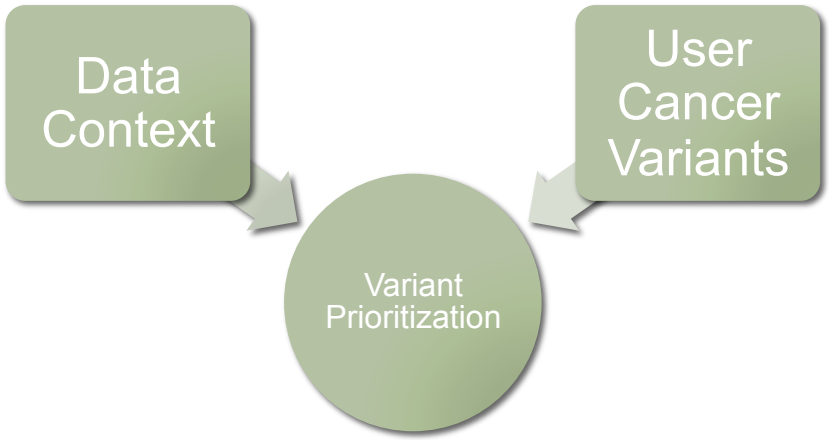
Prioritizing non-coding regulatory variants

- Build a compendium of ranked variants ('Data Context')
  - Regions: ENCODE, 1000 Genomes, HOTS, GERP score
  - Variants: Cancer WGS
  - Genes: TCGA, ICGC, COSMIC
  - Networks: PPI, ChIP-Seq, RNA-Seq



Prioritizing non-coding regulatory variants

- Get user's variant dataset, and prioritize against built data context (use weighted scoring scheme)



Scoring non-coding regulatory variants

1. What are the potential regulatory sequences?
    1. Get annotated transcription factor binding sites, enhancers, promoters, methylation sites
2. **What does the variant do in the sequence?**
    1. Loss of function (motif breaker)
    2. Gain of function (motif gain, new binding site?)
3. **Has the region been conserved traditionally?**
    1. GERP score – conservation over species
    2. 1000 genomes project – evolutionarily constrained regions in human populations
4. **What does the site regulate?**
    1. Neighborhood genes and validated interactions/regulation
    2. Role of gene in the cell (network centrality)

1. Functional Annotation of Putative Regulatory Elements at Cancer Susceptibility Loci. Cancer Inform. 2014; 13(Suppl 2): 5-17.

Scoring non-coding regulatory variants

1. **What are the potential regulatory sequences?**
    1. Get annotated transcription factor binding sites, enhancers, promoters, methylation sites
2. What does the variant do in the sequence?
    1. Loss of function (motif breaker)
    2. Gain of function (motif gain, new binding site?)
3. **Has the region been conserved traditionally?**
    1. GERP score – conservation over species
    2. 1000 genomes project – evolutionarily constrained regions in human populations
4. **What does the site regulate?**
    1. Neighborhood genes and validated interactions/regulation
    2. Role of gene in the cell (network centrality)

1. Functional Annotation of Putative Regulatory Elements at Cancer Susceptibility Loci. Cancer Inform. 2014; 13(Suppl 2): 5-17.

Scoring non-coding regulatory variants

1. **What are the potential regulatory sequences?**
    1. Get annotated transcription factor binding sites, enhancers, promoters, methylation sites
2. **What does the variant do in the sequence?**
    1. Loss of function (motif breaker)
    2. Gain of function (motif gain, new binding site?)
3. Has the region been conserved traditionally?
    1. GERP score – conservation over species
    2. 1000 genomes project – evolutionarily constrained regions in human populations
4. **What does the site regulate?**
    1. Neighborhood genes and validated interactions/regulation
    2. Role of gene in the cell (network centrality)

1. Functional Annotation of Putative Regulatory Elements at Cancer Susceptibility Loci. Cancer Inform. 2014; 13(Suppl 2): 5-17.

Scoring non-coding regulatory variants

1. **What are the potential regulatory sequences?**
    1. Get annotated transcription factor binding sites, enhancers, promoters, methylation sites
2. **What does the variant do in the sequence?**
    1. Loss of function (motif breaker)
    2. Gain of function (motif gain, new binding site?)
3. **Has the region been conserved traditionally?**
    1. GERP score – conservation over species
    2. 1000 genomes project – evolutionarily constrained regions in human populations
4. What does the site regulate?
    1. Neighborhood genes and validated interactions/regulation
    2. Role of gene in the cell (network centrality)

1. Functional Annotation of Putative Regulatory Elements at Cancer Susceptibility Loci. Cancer Inform. 2014; 13(Suppl 2): 5-17.

Resultant scoring method

Coding region variants (cis regulation)	**Noncoding region variants (trans regulation)**
Non-synonymous mutation?	Motif breaking/making score
Selection pressure on gene?	Region (ultra) sensitive/conserved?
Mutation is recurrent in cohort? GERP score > 2 Gene is a network hub?	

Next step

- Have an annotated, scored set of variants
- Have a recurrence database
- Compare each of the input variants to the dataset, and prioritize variants

Validation

- Comparative tools
  - CADD – Combined Annotation Dependent Depletion. SVM model based on features from previous genome wide annotations.
  - GWAVA – Genome Wide Annotation of Variants.
- **Regulatory cancer variants (somatic)**
  - **TCGA + ICGC cancer cohorts**
  - COSMIC regulatory somatic variant annotations (can discuss after presentation)
- Germline pathogenic variants
  - Human Gene Mutation Database (HGMD)
  - Human inherited diseases – 5700 genes

Results – cancer variants

- Take 1

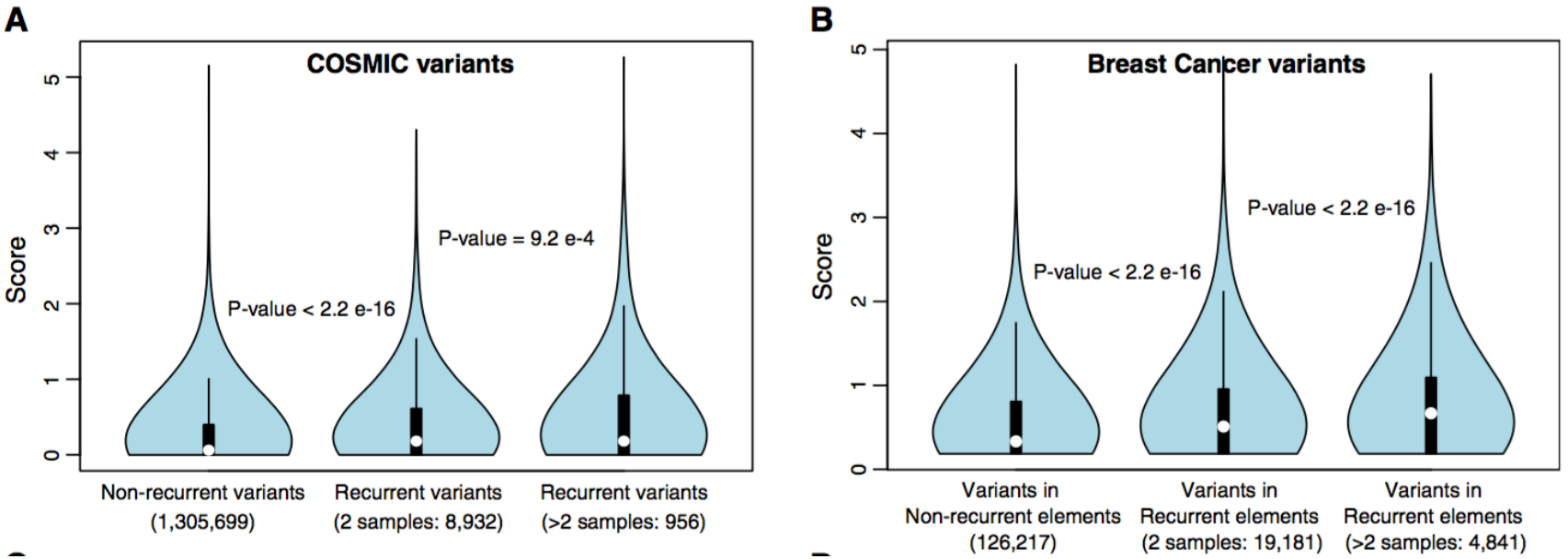
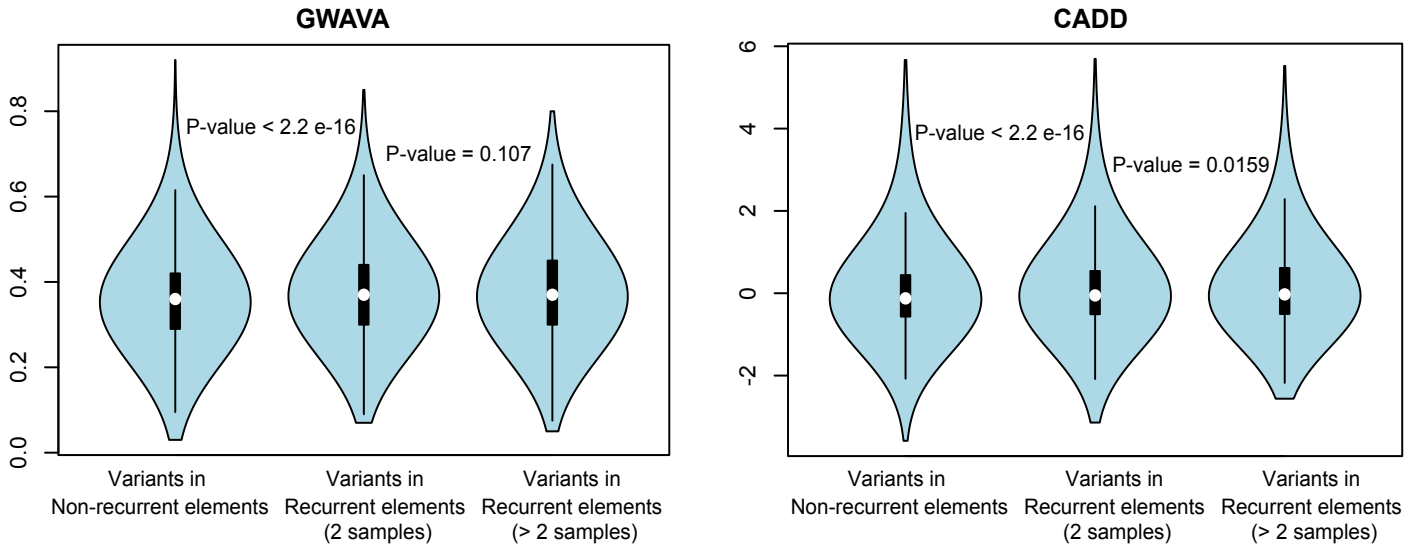


Figure 4. Score distribution of variants based on their recurrence in COSMIC and 119 breast cancer samples.

Results – cancer variants



Comparison	GWAVA (AUC)	CADD (AUC)	FunSeq2 (AUC)
Recurrent vs. non-recurrent variants	0.53	0.52	0.59
>2 samples recurrent vs. non-recurrent variants	0.54	0.53	0.62

Figure S6. Comparison of GWAVA and CADD using Breast Cancer set

Validation

- Regulatory cancer variants (somatic)
  - TCGA + ICGC cancer cohorts
  - COSMIC regulatory somatic variant annotations (can discuss after presentation)
- **Germline pathogenic variants**
  - **Human Gene Mutation Database (HGMD)**
  - **Human inherited diseases – 5700 genes**
- Comparative tools
  - CADD – Combined Annotation Dependent Depletion. SVM model based on features from previous genome wide annotations.
  - GWAVA – Genome Wide Annotation of Variants.

Results – germline variants

- Ranked HGMD variants higher than the controls
- Hence, the prioritization works

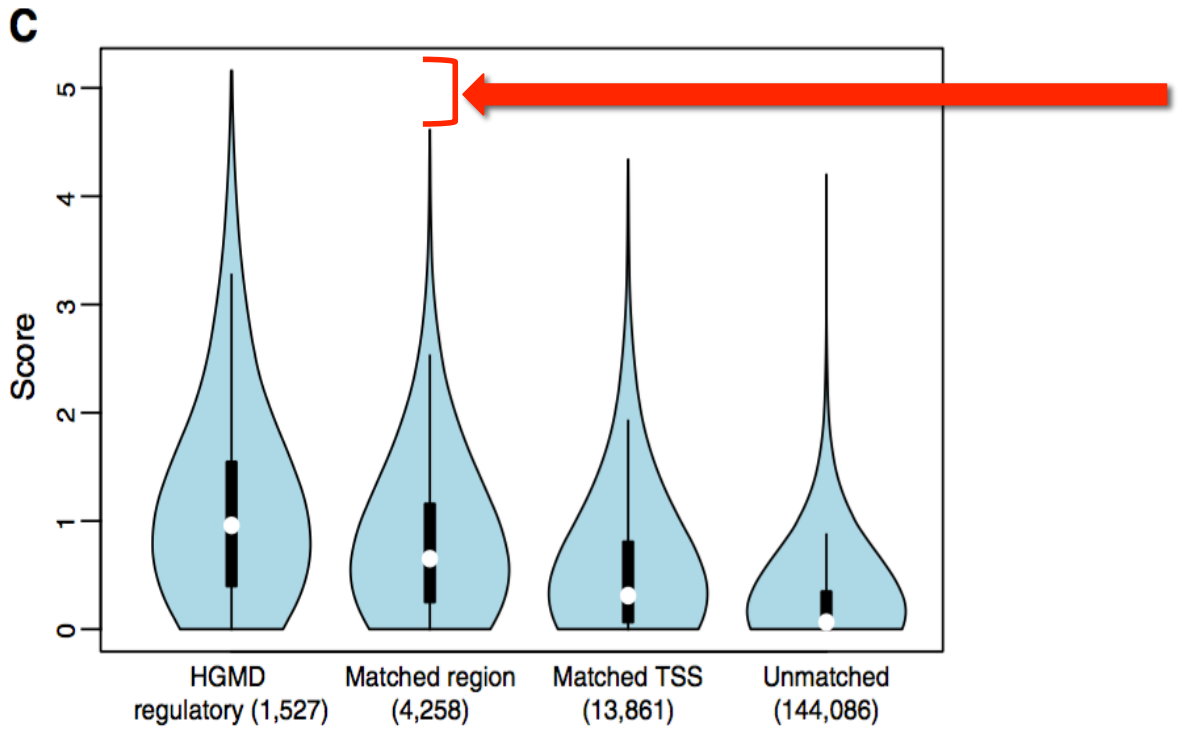


Figure 4. Prediction scores for regulatory elements in HGMD and controls

Drawbacks

- Prioritization o– what? Oh look, score distributions!
- Comparison with other tools is hand wavy
  - using only 2 TERT variants to prove ‘we are better’
- Only show results from large cancer datasets (with relatively lower average mutational load)
- Unsatisfactory treatment of results for germline pathogenic variant ranking
  - Role of regulatory elements in hereditary diseases is more well studied
  - β-thalassemia, hemophilia and atherosclerosis[1]
  - ‘We called variants, variants’

1. Cis regulatory mutations in human disease. Brief Funct Genomic Proteomics 2009 Jul; 8(4): 310–316

Important considerations

- Driver mutations may vary across tumours
  - Such drivers will not show recurrence across samples
  - Option to not add extra weight for cohort-wide recurrence (of mutation, or regulatory element)
- Recurrence based method may also not be as successful in small sample sizes
  - Generated a ‘recurrence database’ to support annotations

Questions?


WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS

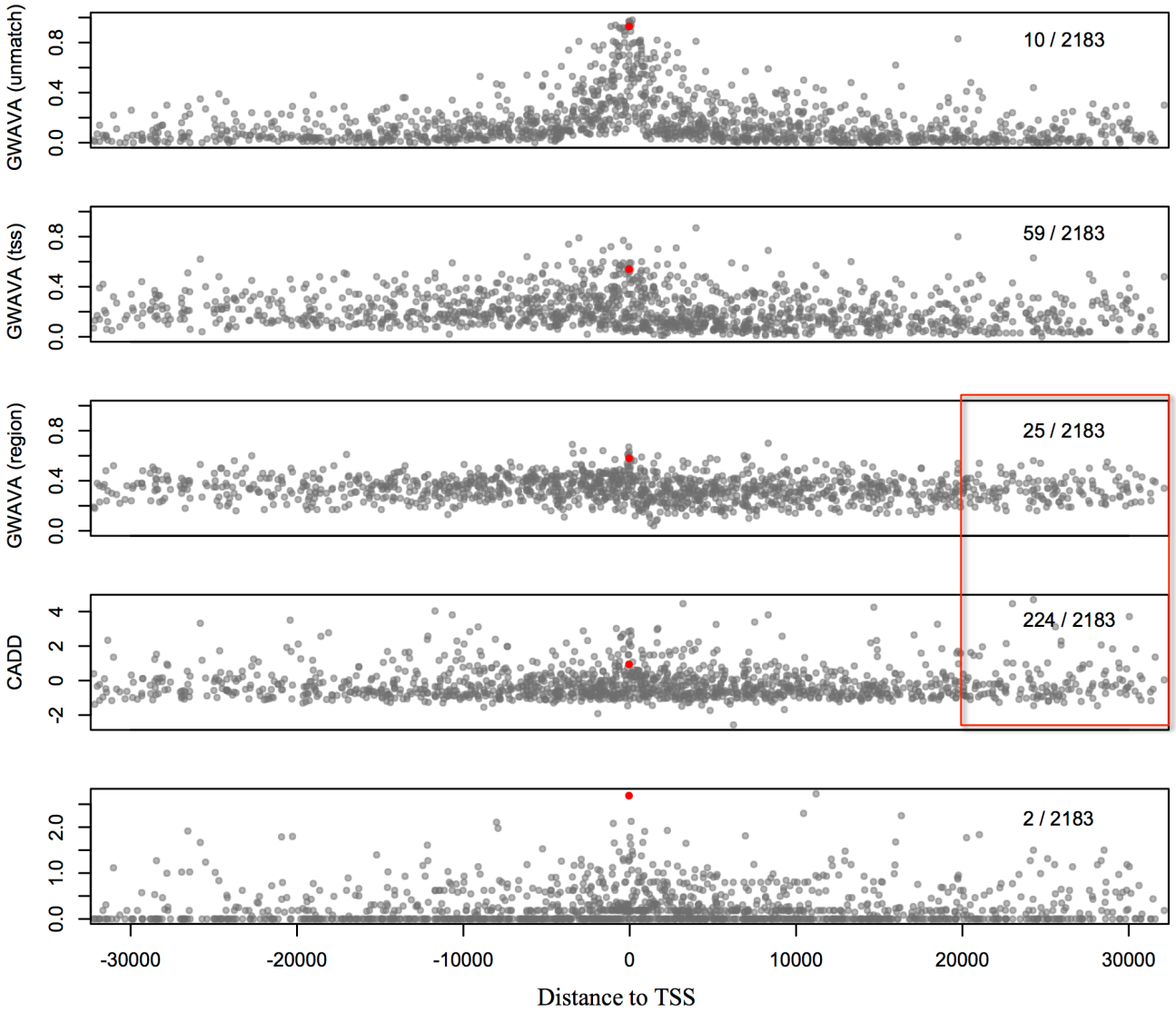
Future directions

- Validation with new target regulatory mutations
  - **TERT**, PLEKHS1, WDR74, SDHD[1]
- Other tools for comparison
  - Segway implements a dynamic Bayesian network method using the aforementioned biofeatures (ChIP-seq and DHS signals) to identify patterns associated with transcription start sites, gene boundaries, enhancers, and other transcription regulators in an unsupervised approach.
- Studying kataegis (regions of hypermutation)
  - Associated with somatic rearrangement events
  - APOBEC family

1. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics* **46**, 1160–1165 (2014)

Results – cancer variants

- Take 2 (using recurrence database)
  - 1 Medulloblastoma sample (only 2/100 samples had TERT mutations)
  - 2183 somatic SNVs
  - Ranked TERT promoter mutation 2nd
  - Other methods were just oh so bad….
- Comparative tools
  - CADD – Combined Annotation Dependent Depletion. SVM model based on features from previous genome wide annotations.
  - GWAVA – Genome Wide Annotation of Variants. Random Forest approach



- **Figure S10. Relationship between distance to TSS and prediction scores (using variants from one Medulloblastoma sample - MB59).** Red dot is the *TERT* promoter mutation. Authors reported 'matched region' model of GWAVA for all analysis, as the model is less prone to bias.

Non-coding regulatory variants

- Cis-
  - Located on same DNA strand as the gene
  - Promoters, enhancers, silencers
- Trans-
  - DNA sequences that encode transcription factors

Methylation sites

- There are three histone modification marks that are particularly informative for the identification of most active enhancer and promoter regulatory regions, namely H3k4me1, H3k27ac, and H3k4me3. The H3k4me1 histone mark is associated with enhancers downstream of transcription start sites, and the H3k27ac signal is similarly thought to enhance transcription. Alternatively, the H3k4me3 mark is associated with active promoters.[1]

1. Functional Annotation of Putative Regulatory Elements at Cancer Susceptibility Loci. Cancer Inform. 2014; 13(Suppl 2): 5-17.

GERP score

- Genomic Evolutionary Rate Profiling
- Identify slowly evolving regions in an MSA
- Nucleotide constraint score
  - Provides enrichment for deleterious causal mutations
  - Allows quantitative ranking of candidates
  - Usually places known causal genes at/near top of candidate lists

GWAVA

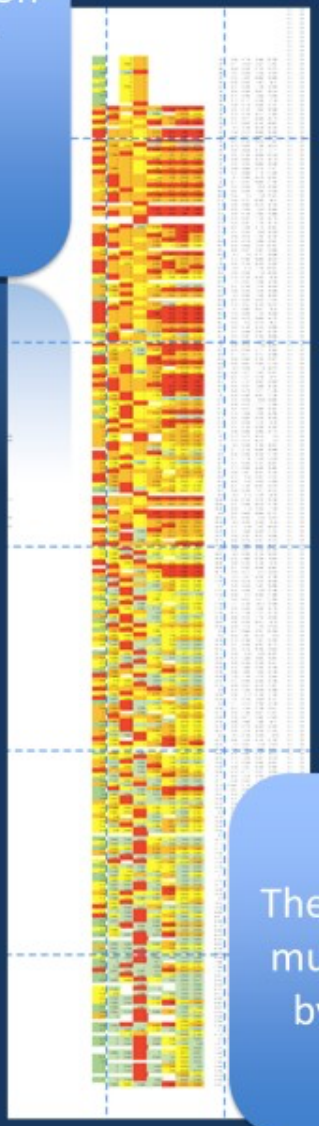
- Predicts functional impact of non-coding genetic variants
- Uses ENCODE/GENCODE annotation of non-coding elements, alongwith genome wide evolutionary conservation and GC content data

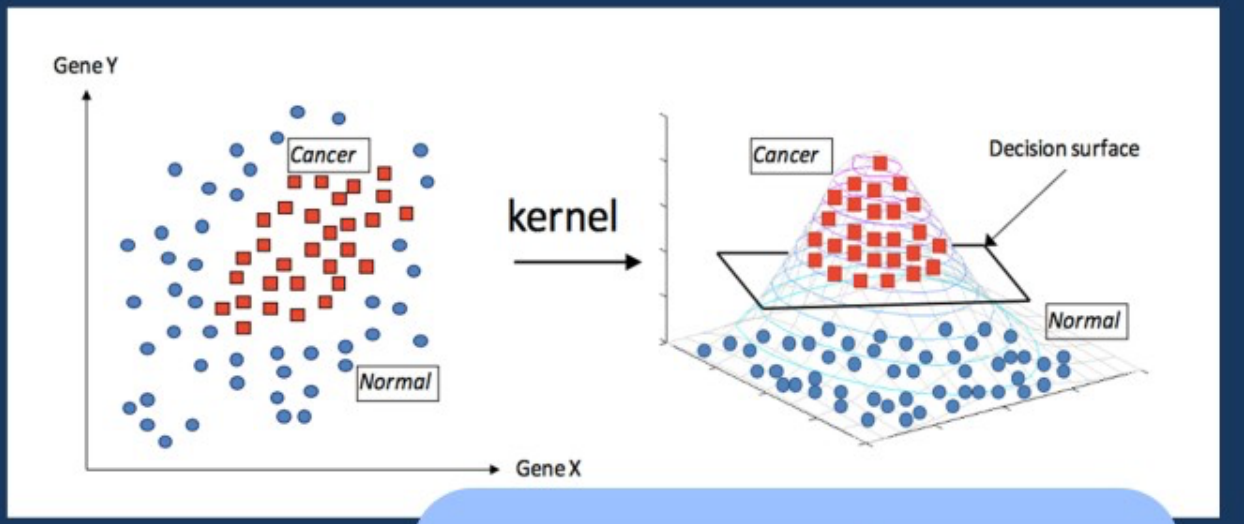
CADD

- Ranking of mutation (tolerated or deleterious) by >60 tools, using CADD score
- Train SVM based on the ranking
- *simulated data to represent possible de novo mutations in humans, that may've occurred but were not fixed in the human population.

The CADD score

Various prediction tools in color coding from damaging to tolerated





Support vector machines use a transformation of “messy data” into a higher dimensional structure where data points in both groups can be separated easily by a “hyperplane”

The Epi4K *de novo* mutations sorted by CADD score

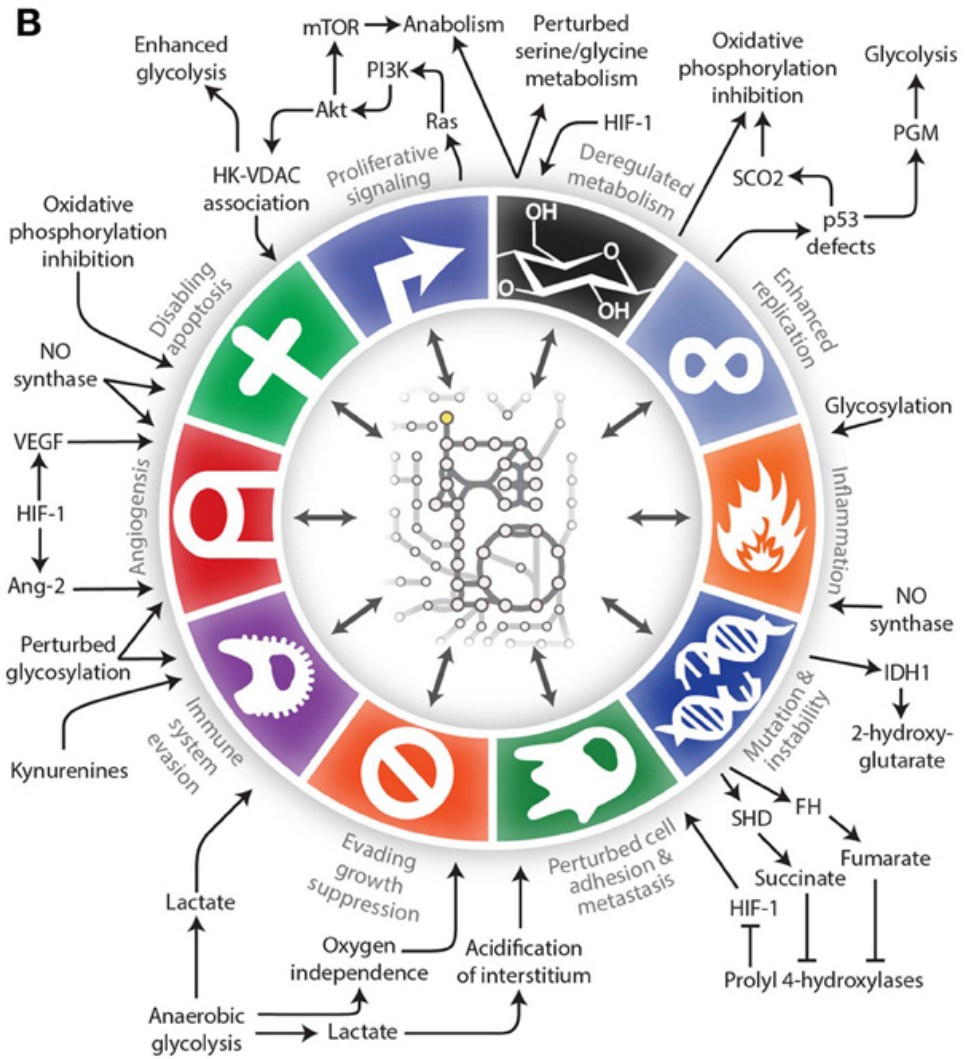
ENCODE

- Encyclopedia of DNA Elements
- Goal – determine role of non-protein coding regions of the human genome
- 90% of disease-associated SNPs are in non-coding regions
- Identified novel DNA regulatory elements

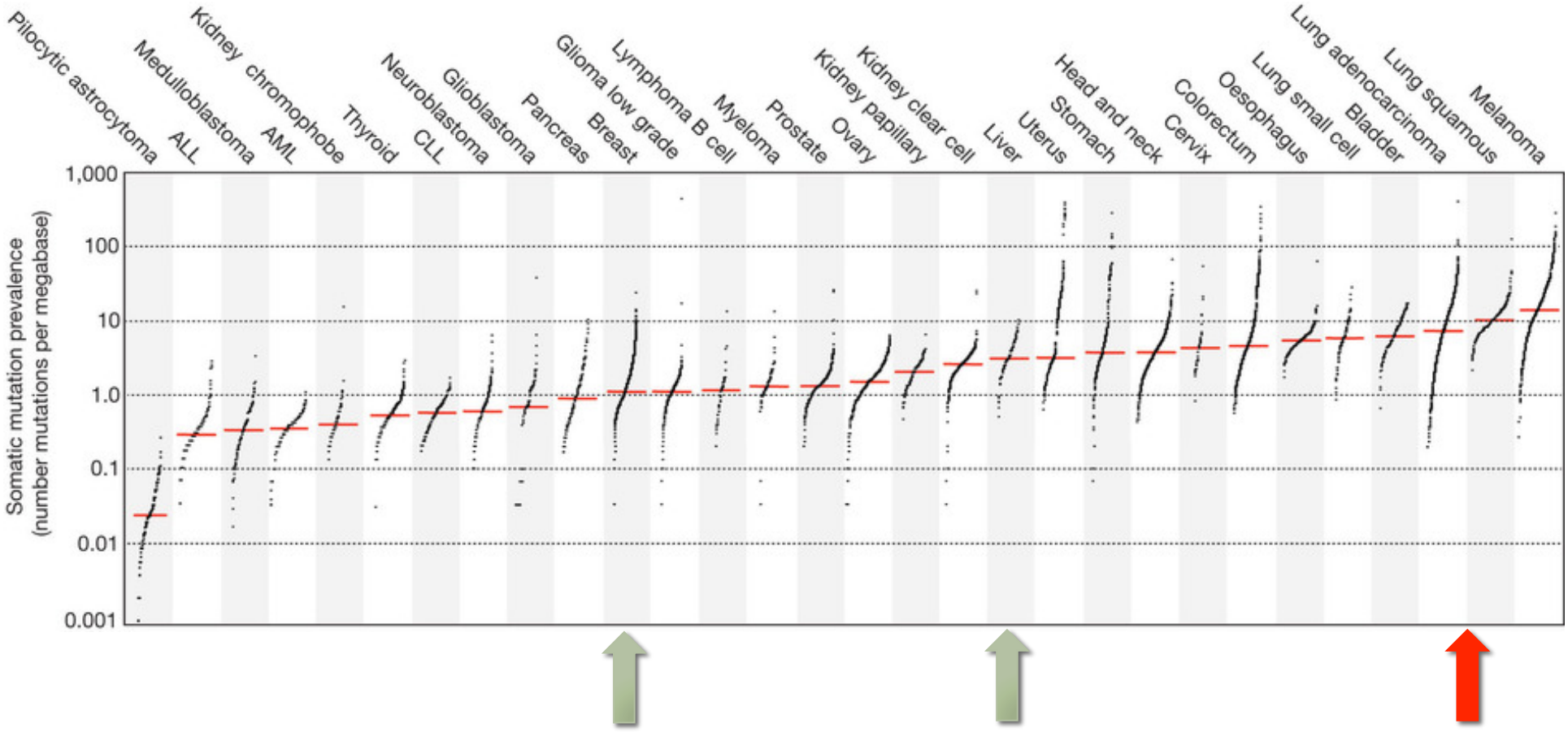
Roadmap epigenomics project

- This study uses it to link prioritized variants with target genes
- Establish an expansive resource of epigenomic maps of normal cells and tissue phenotypes
- Annotate regulatory non-coding elements

More than one way to skin a cat

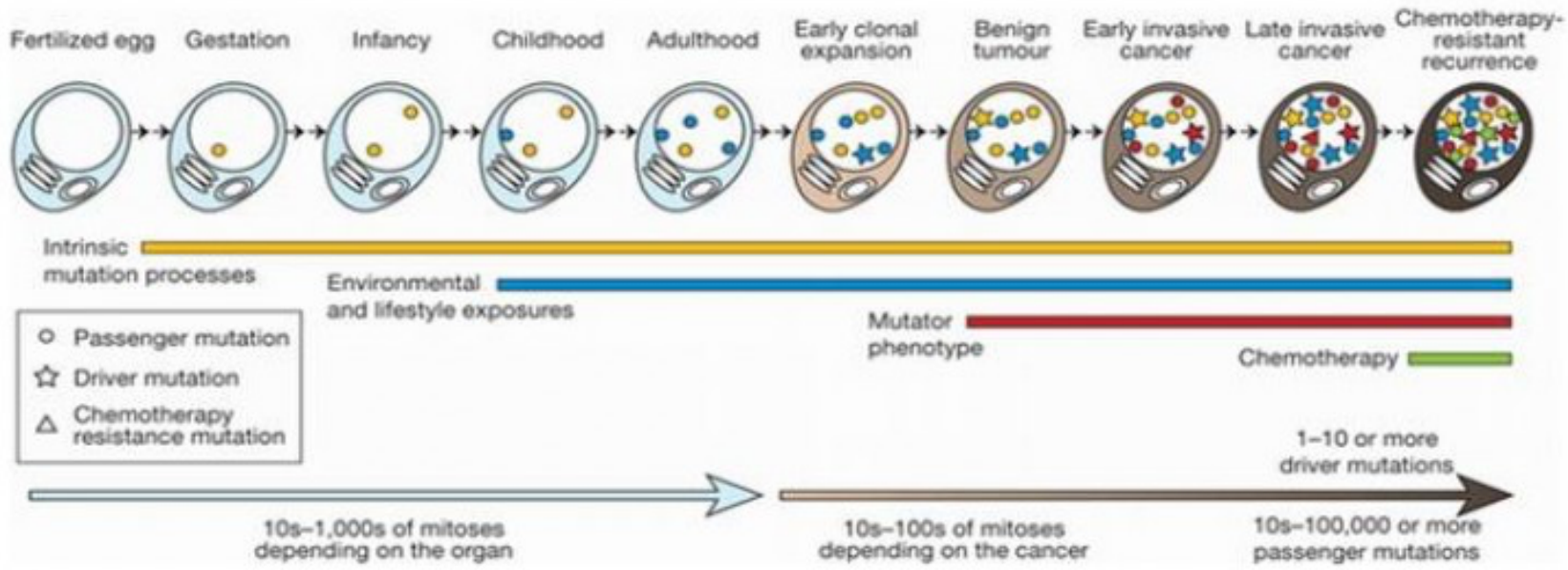


1. The evolution of genome-scale models of cancer metabolism. Front. Physiol., 03 September 2013.



Signatures of mutational processes in cancer. *Nature* **500**, 415–421 (22 August 2013)

What is causal, what is a carrier



Point mutations (single nucleotide)

- SNPs: Most common variants (in >= 1% of population)
  - One every 300 nucleotides
- SNVs:
  - Rarer
- Somatic SNVs
  - Unique to the tumour
  - Spontaneous point mutations