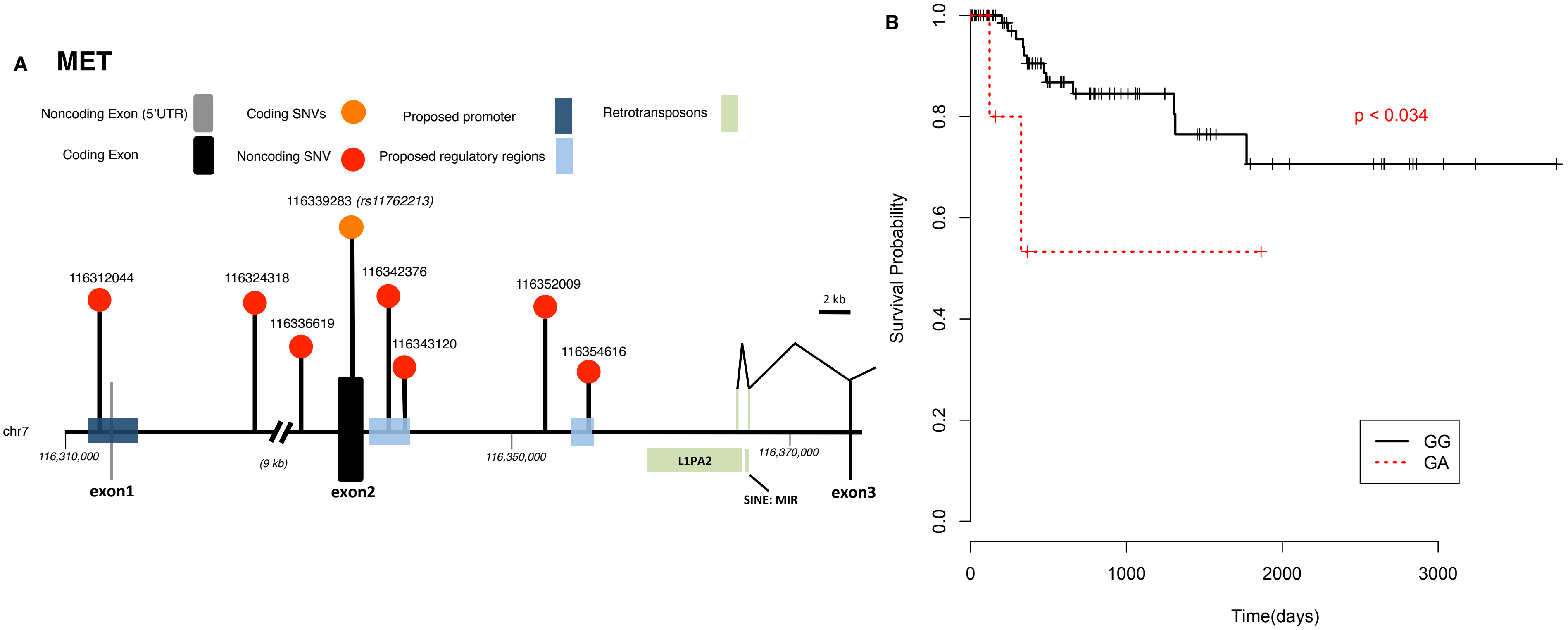
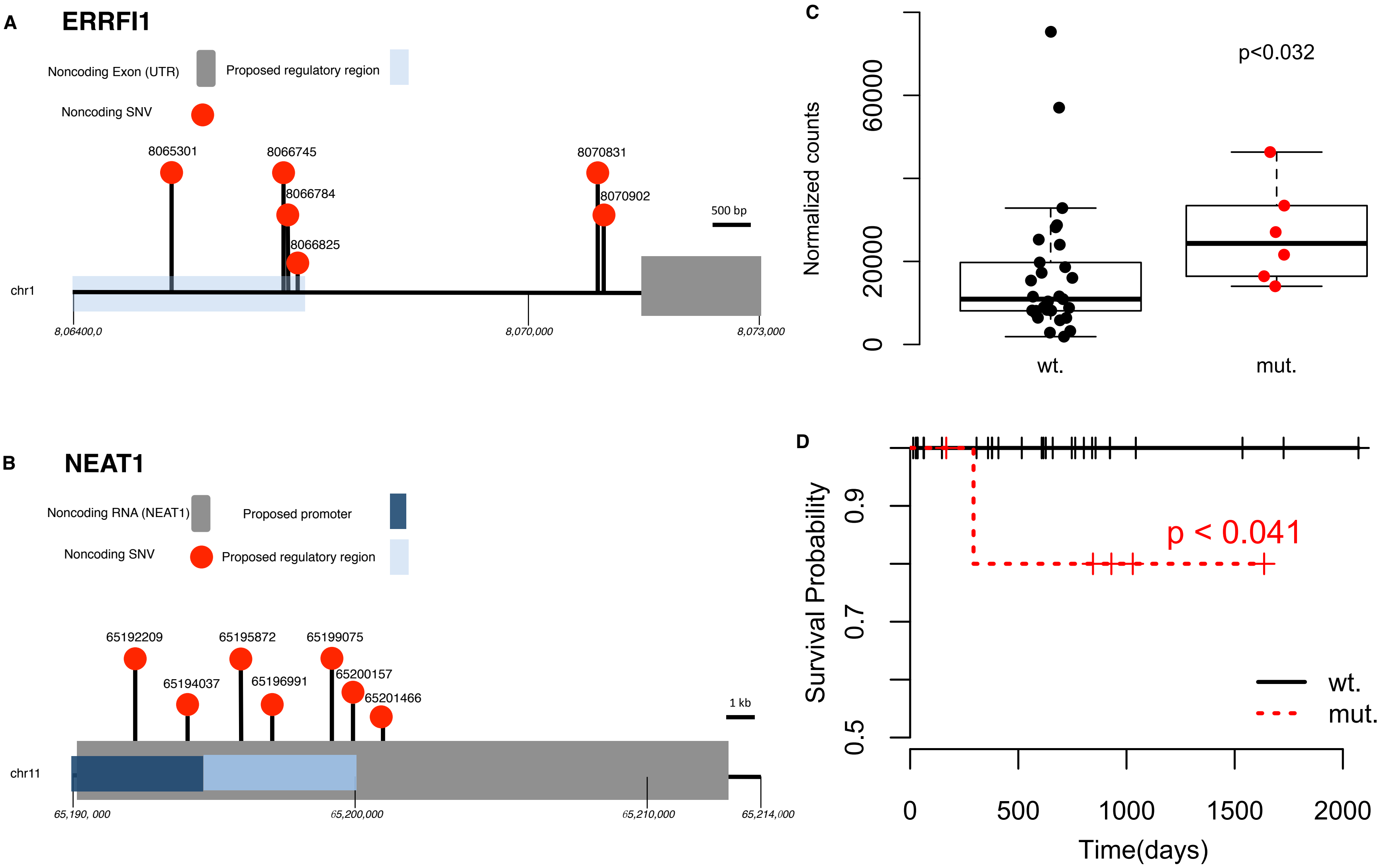
What’s new

- Figure re-plot
- SV comparison
- New DHS/RepliSEQ
- Evolution tree

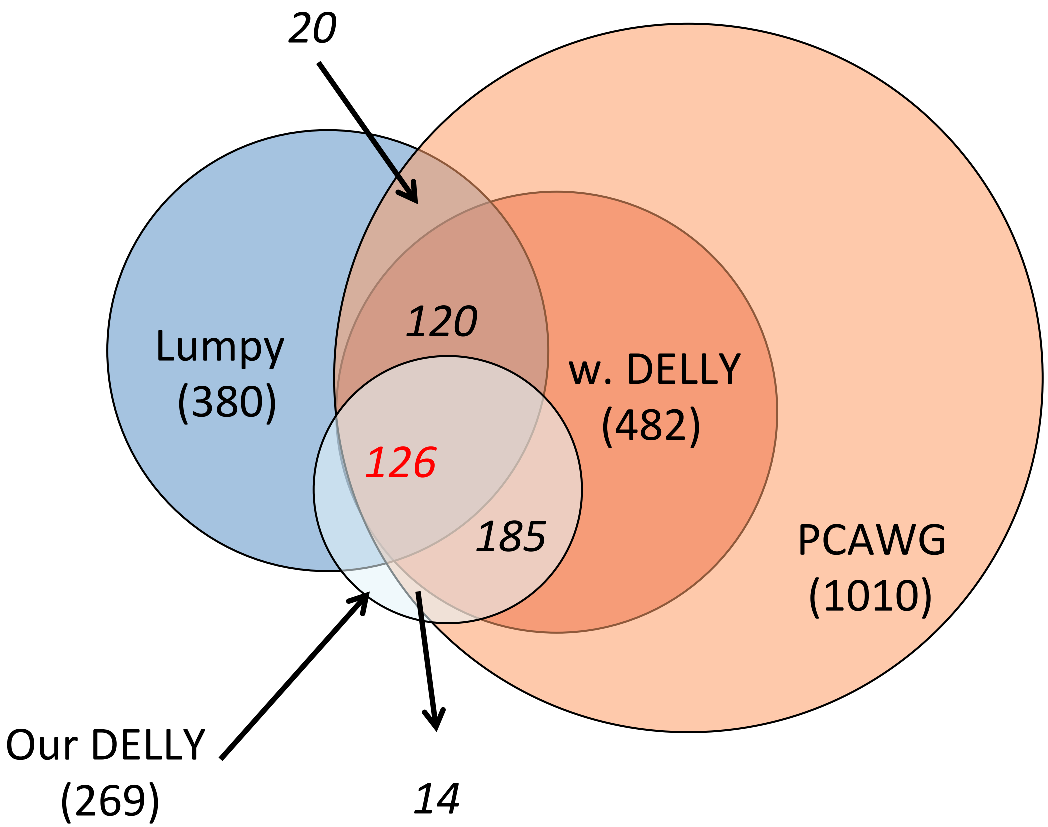

A
MET
Noncoding Exon (5'UTR)
Coding Exon
Coding SNVs
Noncoding SNV
Proposed promoter
Proposed regulatory regions
Retrotransposons
116339283 (rs11762213)
116312044
116324318
116336619
116342376
116343120
116352009
116354616
2 kb
chr7
116,310,000
(9 kb)
116,350,000
116,370,000
L1PA2
SINE: MIR
exon1
exon2
exon3
B
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
p < 0.034
GG
GA
0
1000
2000
3000
Time(days)


A
ERRFI1
Noncoding Exon (UTR)
Proposed regulatory region
Noncoding SNV
8065301
8066745
8066784
8066825
8070831
8070902
500 bp
chr1
8,06400,0
8,070,000
8,073,000
B
NEAT1
Noncoding RNA (NEAT1)
Proposed promoter
Noncoding SNV
Proposed regulatory region
65192209
65194037
65195872
65196991
65199075
65200157
65201466
1 kb
chr11
65,190, 000
65,200,000
65,210,000
65,214,000
C
Normalized counts
0
20000
60000
p<0.032
wt.
mut.
D
Survival Probability
0.5
0.7
0.9
p < 0.041
wt.
mut.
0
500
1000
1500
2000
Time(days)

SV comparison

- First, why Lumpy/SVscore?
  - Our strength at bkpts (as to aCGH)
    - Overlapping bkpts with functional regions
    - What about bkpt labeled as IMPRECISE
  - Lumpy/SVscore gives the right normalization
    - Normalized CADD scores around bkpts
    - Report intervals as well
  - Issues: Quality & SVTyper
    - The genotyper works weirdly

SV comparison



This is on 32 samples
(*we called SV from 35*)

Criteria:

1. *0.5* reciprocal overlap
2. Matched sample
3. Matched SV CLASS

Only get *126* SVs if overlapping DELLY with Lumpy using 32 samples

SV comparison

- If we think PCAWG DELLY is the ground truth
  - – Other methods rely on assembly…different searching space (complicated SV events produce way too many bkpts)
  - – Not a fair comparison

	#SVs	FN	FP	Addl. catch
Lumpy	380	71%	63.2%	20 (5.26%)
Delly (ours)	269	61.6%	31.2%	14 (5.20%)

Bkpts&interval overlapping

- No overlapping with pRCC MutSig genes (~10)
- COSMIC genes

	bkpts
1	AFF3
1	AKAP9
1	ATRX
1	CDK6
1	CDKN2A
1	CLP1
1	CREBBP
1	ERCC2
1	LRP1B
1	MKL1
1	MLLT10
1	POLE
1	POT1
1	SND1
1	SOX2-OT
1	STAT5B
1	STK11
1	THRAP3
2	SPEN

- Three overlapping with pRCC MutSig
  - Extensive SVs in TCGA-B9-4116
  - DEL of STAG2
  - A large INV involves NFE2L2

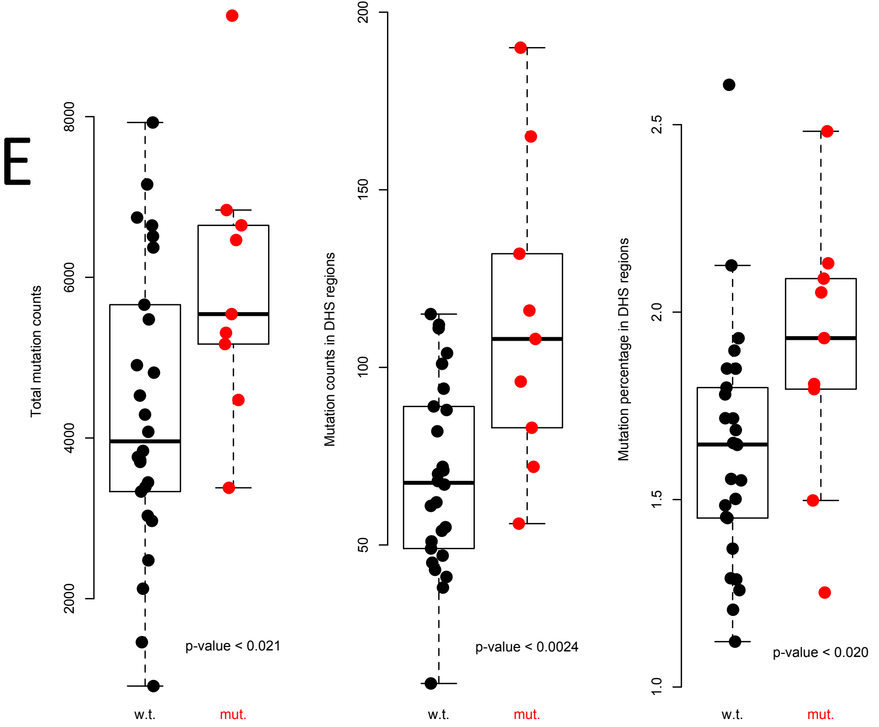
SV overlapping

- Some interesting COSMIC cases
  - 3 CDKN3A (confirmed NEJM 3/5 cases)
  - 2 SDHB deletion (interacts with FH and SDHA)
  - 1 EGFR duplication (pRCC responses to TKIs)
  - 1 HIF1A duplication (inhibited by VHL;)
  - 1 polyE bkpts (same case)
  - 1 DNMT3A deletion (affect methylation)
  - 1 MALAT1 *deletion*
  - 1 HGF *deletion (ligand for MET)*

1	POLE2
1	DNMT3A
1	HIF1A
1	IDH1
1	MALAT1
2	EGFR
2	SDHB
3	CDKN2A
1	HGF

New DHS scheme

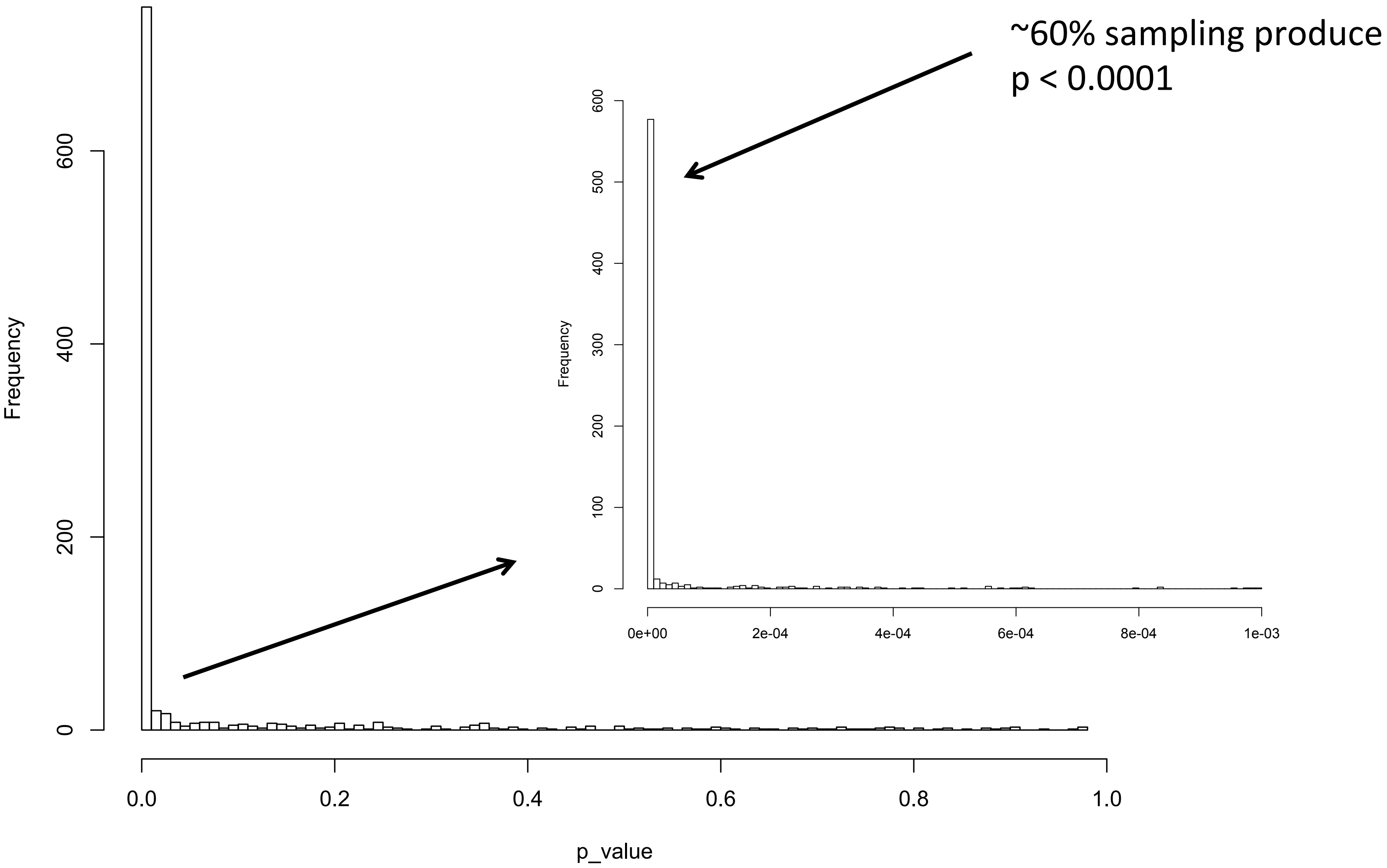
- Pulled fetal kidney cortex DHS from Roadmap
- 11 samples from different fetuses
  - Ultra-conservative: take the overlap of all samples
  - DHS percentage roughly matches ENCODE HEK293 narrowpeak (previously used)



RepliSEQ

- Newly defined CR-genes
  - Overlapping NEJM spreadsheet(pRCC-related) genes with CR&SWI/SNF pathway gene list
- Conservative RT signals (*Ref: NGen.*)
  - Taken median from 11 ENCODE cell lines
- Prudent adapted KS test (subsampling)
  - Q: does the RT dist. of SNVs from CR$^{mut.}$ samples differ from CR$^{wt.}$ samples
  - Randomly shuffle labels (9 v.s. 26) for 1,000 times to generate imperial TS distribution

Why we need subsampling

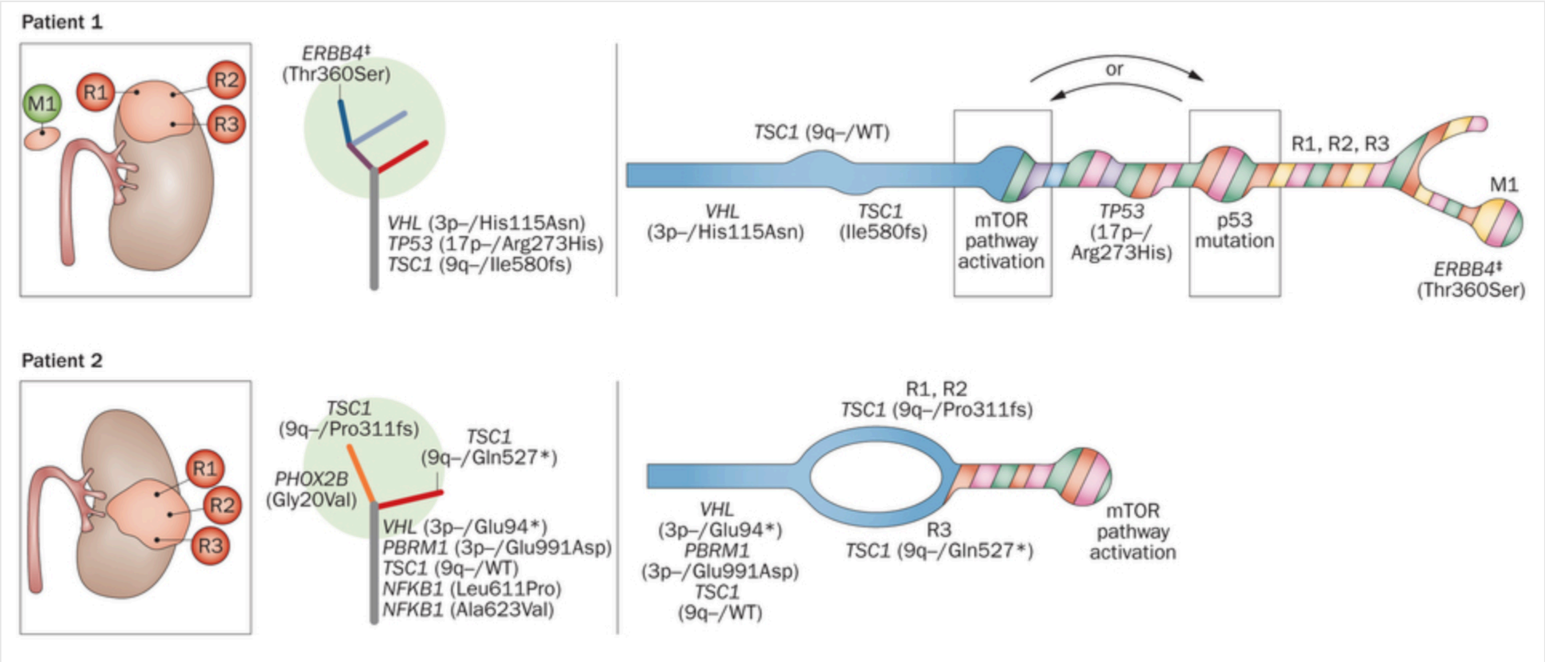


~60% sampling produce
$p < 0.0001$

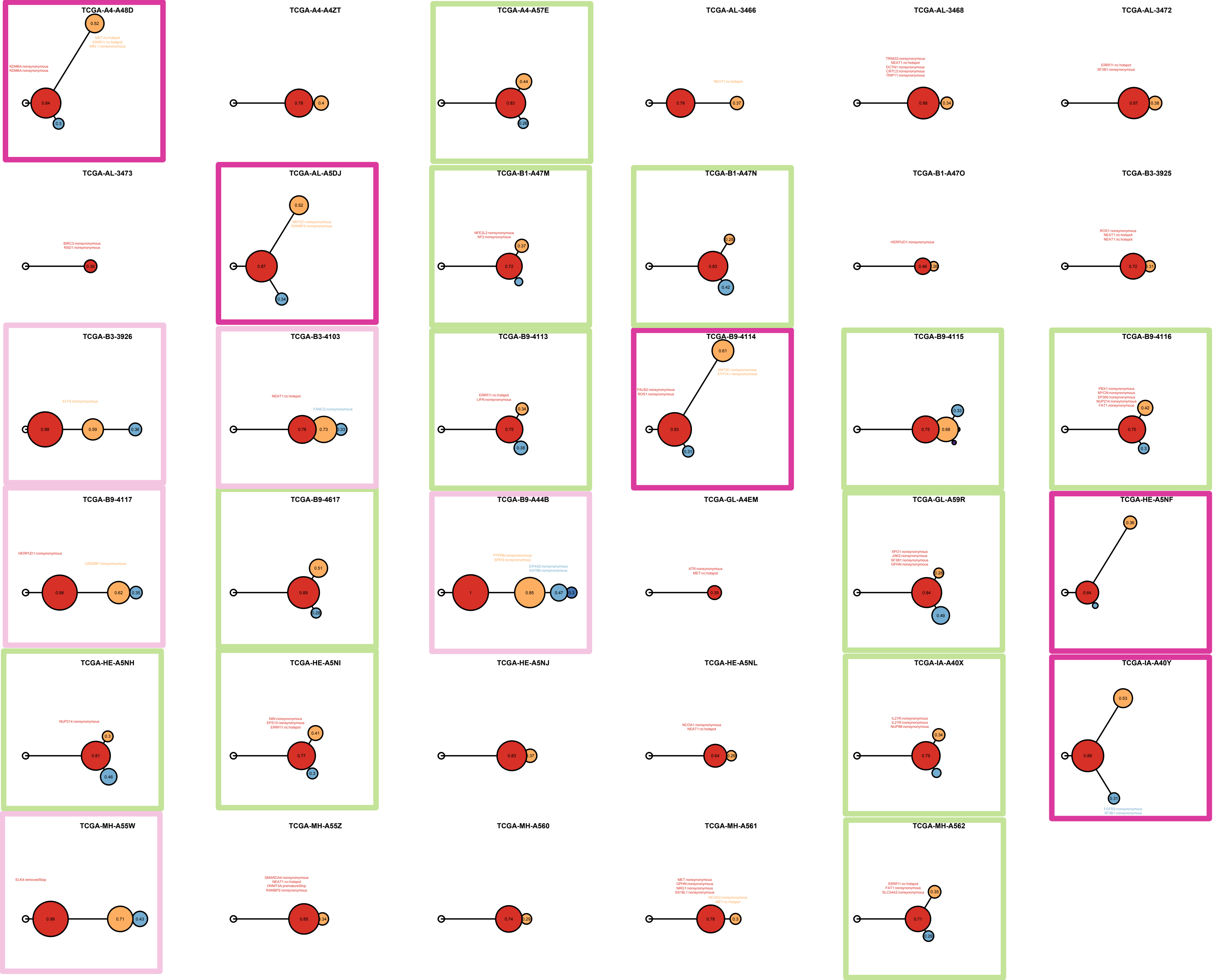
KS tests w/ subsampling: Issues

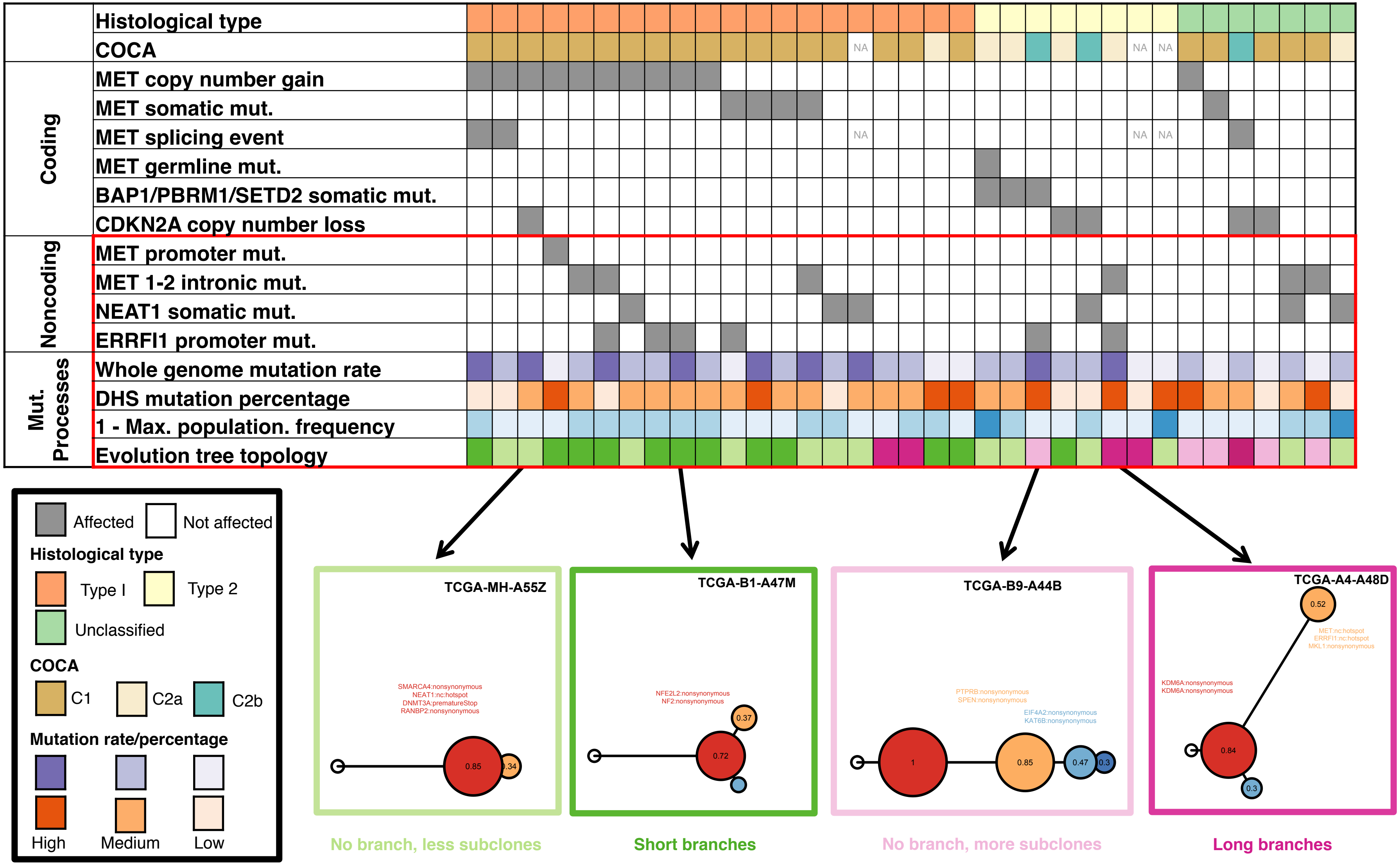
- A. p-value below limits
  - Our test produces p < 2.2e-16 (which becomes 0)
  - However, we observed 26% such cases
- B. test power not uniform
  - SNVs in each sample varies
  - Makes test statistic (D) not directly comparable
  - Worse: power is not uniform!

Evolution tree



- Less interesting because:
  - Single punch from each patient
  - Not ultra-deep sequence
  - Masked all CNV + MAF>0.6


TCGA-A4-A48D
0.52
MET:nc:hotspot
ERRFI1:nc:hotspot
MKL1:nonsynonymous
KDM6A:nonsynonymous
KDM6A:nonsynonymous
0.84
0.3
TCGA-A4-A4ZT
0.78
0.4
TCGA-A4-A57E
0.44
0.83
0.28
TCGA-AL-3466
NEAT1:nc:hotspot
0.79
0.37
TCGA-AL-3468
TRIM33:nonsynonymous
NEAT1:nc:hotspot
DCTN1:nonsynonymous
CRTC3:nonsynonymous
TRIP11:nonsynonymous
0.88
0.34
TCGA-AL-3472
ERRFI1:nc:hotspot
SF3B1:nonsynonymous
0.87
0.38
TCGA-AL-3473
BIRC3:nonsynonymous
NSD1:nonsynonymous
0.36
TCGA-AL-A5DJ
0.52
KMT2C:nonsynonymous
RANBP2:nonsynonymous
0.87
0.34
TCGA-B1-A47M
NFE2L2:nonsynonymous
NF2:nonsynonymous
0.37
0.72
TCGA-B1-A47N
0.28
0.83
0.42
TCGA-B1-A47O
HERPUD1:nonsynonymous
0.46
.26
TCGA-B3-3925
ROS1:nonsynonymous
NEAT1:nc:hotspot
NEAT1:nc:hotspot
0.72
0.31
TCGA-B3-3926
ELF4:nonsynonymous
0.98
0.59
0.36
TCGA-B3-4103
NEAT1:nc:hotspot
FANCG:nonsynonymous
0.78
0.73
0.33
TCGA-B9-4113
ERRFI1:nc:hotspot
LIFR:nonsynonymous
0.34
0.75
0.38
TCGA-B9-4114
0.61
KMT2C:nonsynonymous
ATP1A1:nonsynonymous
PALB2:nonsynonymous
ROS1:nonsynonymous
0.93
0.31
TCGA-B9-4115
0.33
0.75
0.68
TCGA-B9-4116
PBX1:nonsynonymous
MYCN:nonsynonymous
EP300:nonsynonymous
NUP214:nonsynonymous
FAT1:nonsynonymous
0.42
0.75
0.3
TCGA-B9-4117
HERPUD1:nonsynonymous
CREBBP:nonsynonymous
0.98
0.62
0.35
TCGA-B9-4617
0.51
0.89
0.28
TCGA-B9-A44B
PTPRB:nonsynonymous
SPEN:nonsynonymous
EIF4A2:nonsynonymous
KAT6B:nonsynonymous
1
0.85
0.47
0.3
TCGA-GL-A4EM
ATR:nonsynonymous
MET:nc:hotspot
0.39
TCGA-GL-A59R
XPO1:nonsynonymous
JAK2:nonsynonymous
SF3B1:nonsynonymous
GPHN:nonsynonymous
0.28
0.84
0.49
TCGA-HE-A5NF
0.36
0.64
TCGA-HE-A5NH
NUP214:nonsynonymous
0.3
0.81
0.46
TCGA-HE-A5NI
NIN:nonsynonymous
EPS15:nonsynonymous
ERRFI1:nc:hotspot
0.41
0.77
0.3
TCGA-HE-A5NJ
0.83
.37
TCGA-HE-A5NL
NCOA1:nonsynonymous
NEAT1:nc:hotspot
0.64
0.29
TCGA-IA-A40X
IL21R:nonsynonymous
IL21R:nonsynonymous
NUP98:nonsynonymous
0.34
0.79
TCGA-IA-A40Y
0.53
0.89
0.31
FGFR2:nonsynonymous
SF3B1:nonsynonymous
TCGA-MH-A55W
ELK4:removedStop
0.99
0.71
0.43
TCGA-MH-A55Z
SMARCA4:nonsynonymous
NEAT1:nc:hotspot
DNMT3A:prematureStop
RANBP2:nonsynonymous
0.85
.34
TCGA-MH-A560
0.74
0.29
TCGA-MH-A561
MET:nonsynonymous
GPHN:nonsynonymous
NRG1:nonsynonymous
SS18L1:nonsynonymous
NCOA2:nonsynonymous
MET:nc:hotspot
0.78
0.3
TCGA-MH-A562
ERRFI1:nc:hotspot
FAT1:nonsynonymous
SLC34A2:nonsynonymous
0.35
0.71
0.29


Histological type
COCA
Coding
MET copy number gain
MET somatic mut.
MET splicing event
MET germline mut.
BAP1/PBRM1/SETD2 somatic mut.
CDKN2A copy number loss
Noncoding
MET promoter mut.
MET 1-2 intronic mut.
NEAT1 somatic mut.
ERRFI1 promoter mut.
Mut. Processes
Whole genome mutation rate
DHS mutation percentage
1 - Max. population. frequency
Evolution tree topology
NA
Affected
Not affected
Histological type
Type I
Type 2
Unclassified
COCA
C1
C2a
C2b
Mutation rate/percentage
High
Medium
Low
TCGA-MH-A55Z
SMARCA4:nonsynonymous
NEAT1:nc:hotspot
DNMT3A:prematureStop
RANBP2:nonsynonymous
0.85
0.34
No branch, less subclones
TCGA-B1-A47M
NFE2L2:nonsynonymous
NF2:nonsynonymous
0.37
0.72
Short branches
TCGA-B9-A44B
PTPRB:nonsynonymous
SPEN:nonsynonymous
EIF4A2:nonsynonymous
KAT6B:nonsynonymous
1
0.85
0.47
0.3
No branch, more subclones
TCGA-A4-A48D
0.52
MET:nc:hotspot
ERRFI1:nc:hotspot
MKL1:nonsynonymous
KDM6A:nonsynonymous
KDM6A:nonsynonymous
0.84
0.3
Long branches

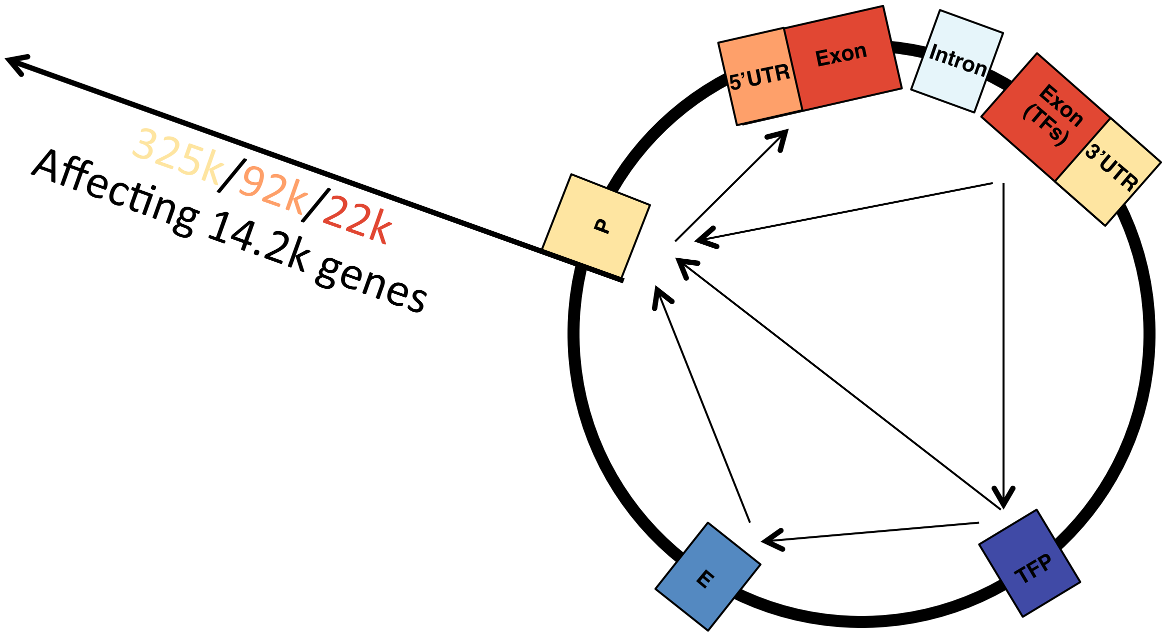
Paper E thoughts

- What we want to sell
  - A complete “tumor portrait”
  - Completeness on several levels
    - Unprecedentedly large WGS cohort
    - Coding/noncoding/SVs…

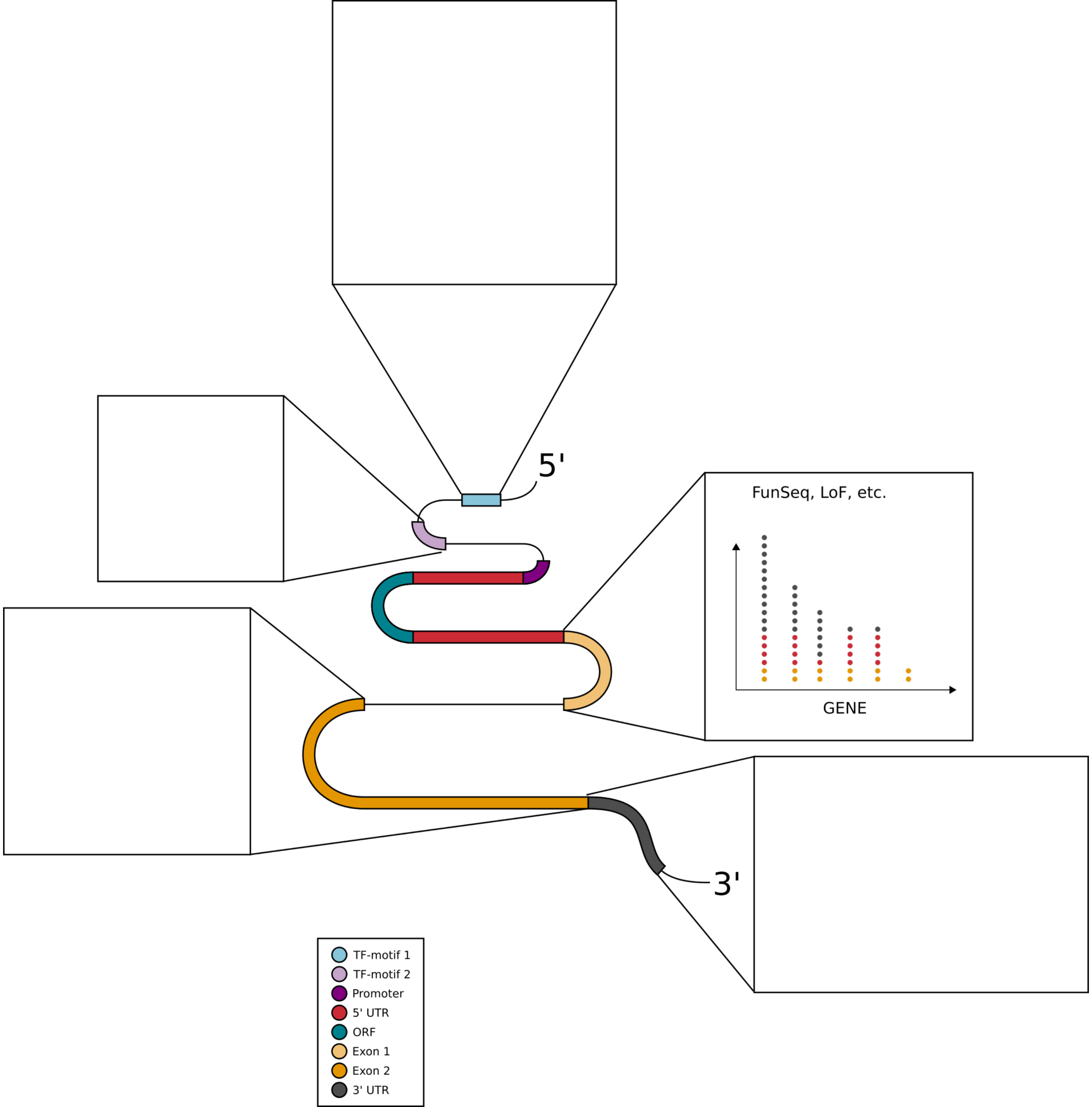
- Fig1: a fancy figure to visualize the deluge of the data
  - Show a gene (“genecreek” or “-circle”)

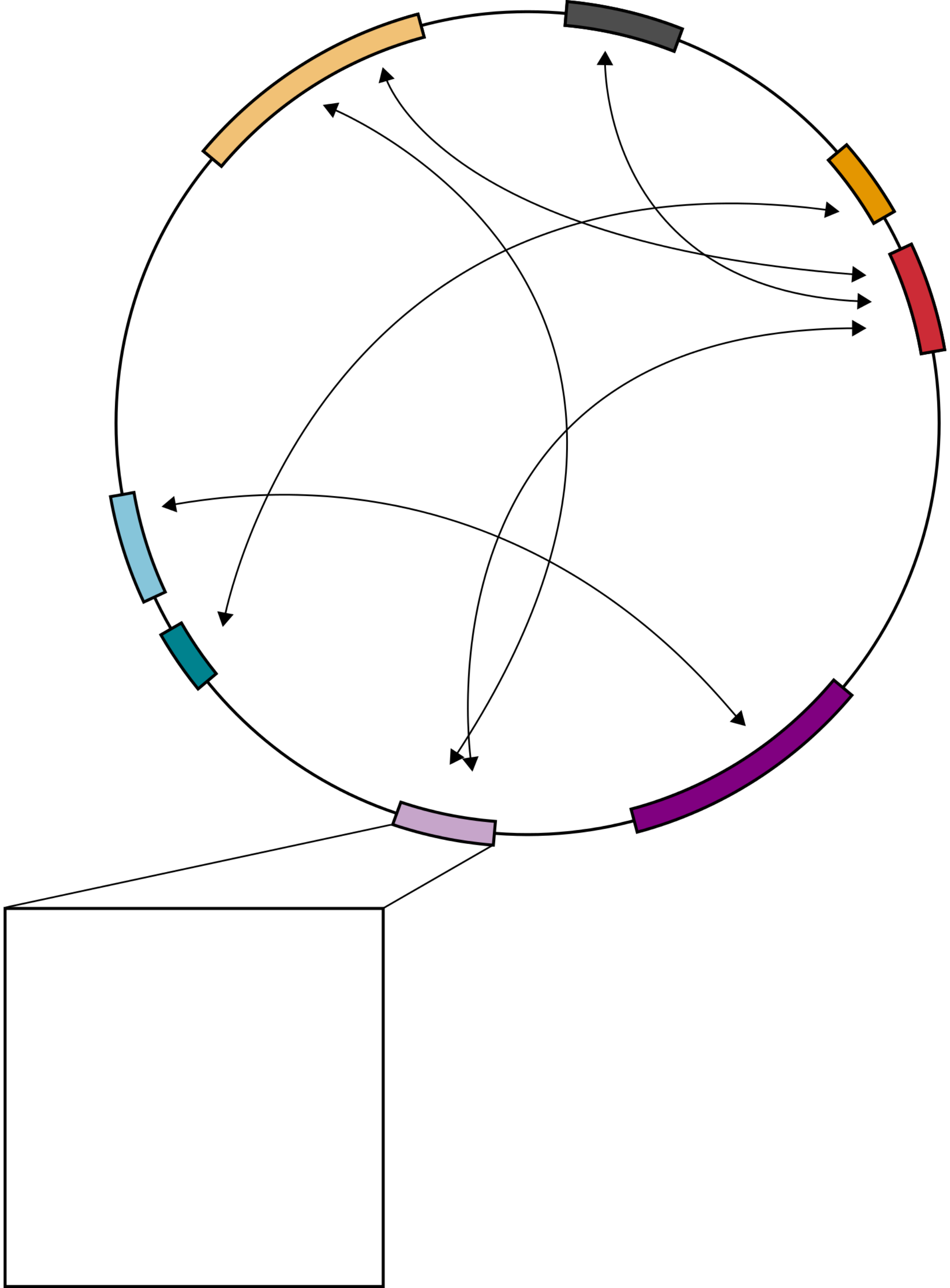
Issues

- One sample with 23 EPB41L2 promoter mutation?
  - And 9 has the identical FunSeq score!
  - Annotation issue? This promoter spans >170kb!
  - However, this is an excluded sample…
  - Still, another included sample has 8…

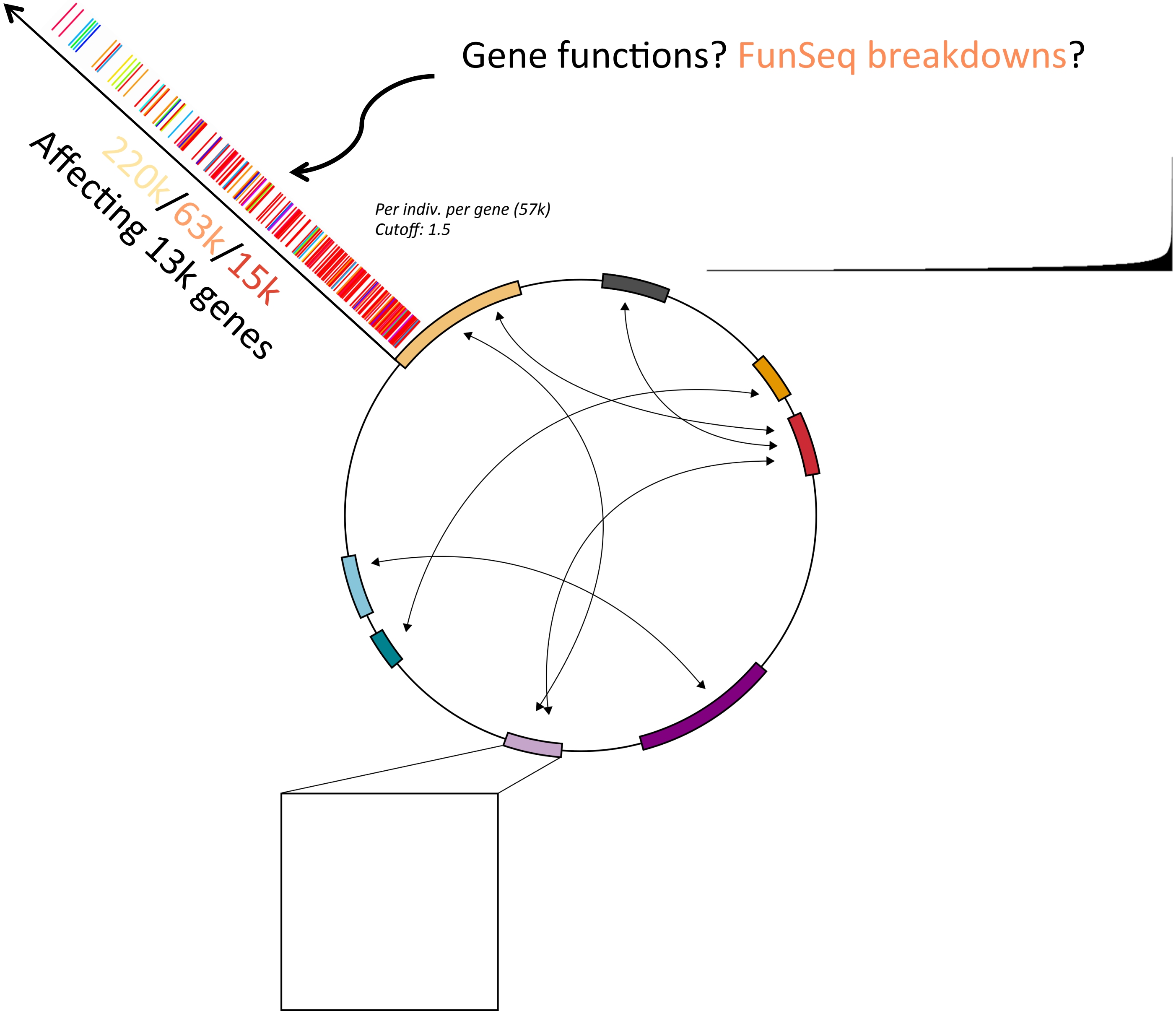


325k/92k/22k
Affecting 14.2k genes


5'
FunSeq, LoF, etc.
GENE
3'
TF-motif 1
TF-motif 2
Promoter
5' UTR
ORF
Exon 1
Exon 2
3' UTR

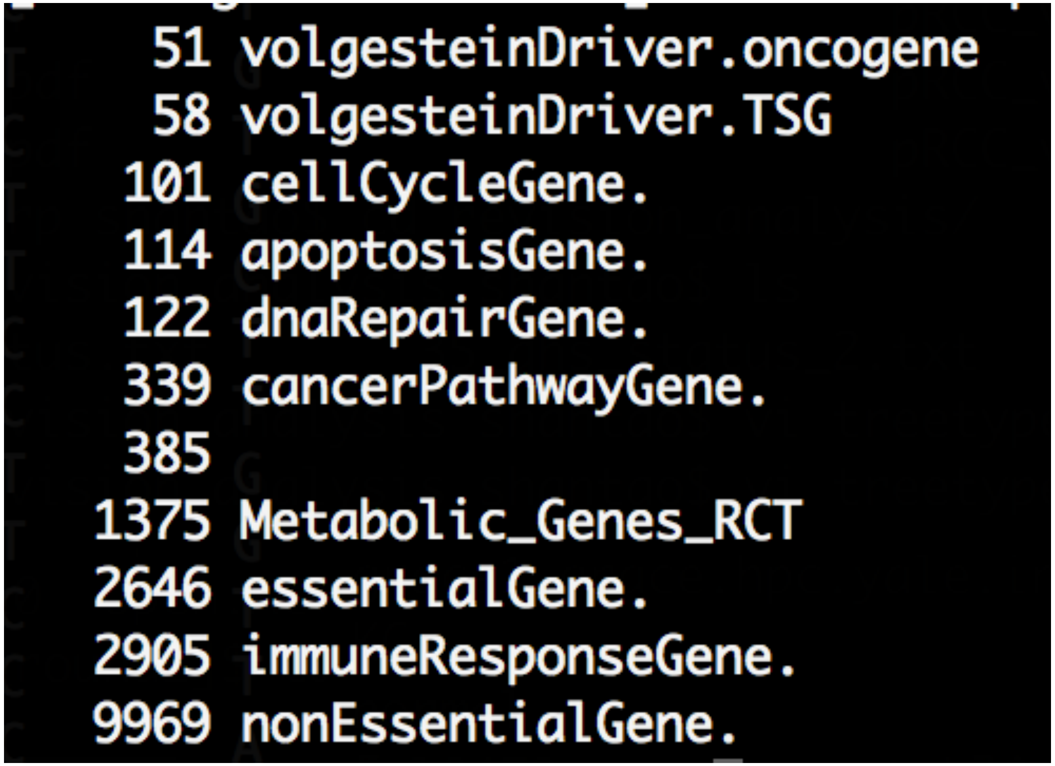


Gene functions? FunSeq breakdowns?



108	TCF4
109	DUSP22
124	TERT
125	SEPT9
154	ZNF595

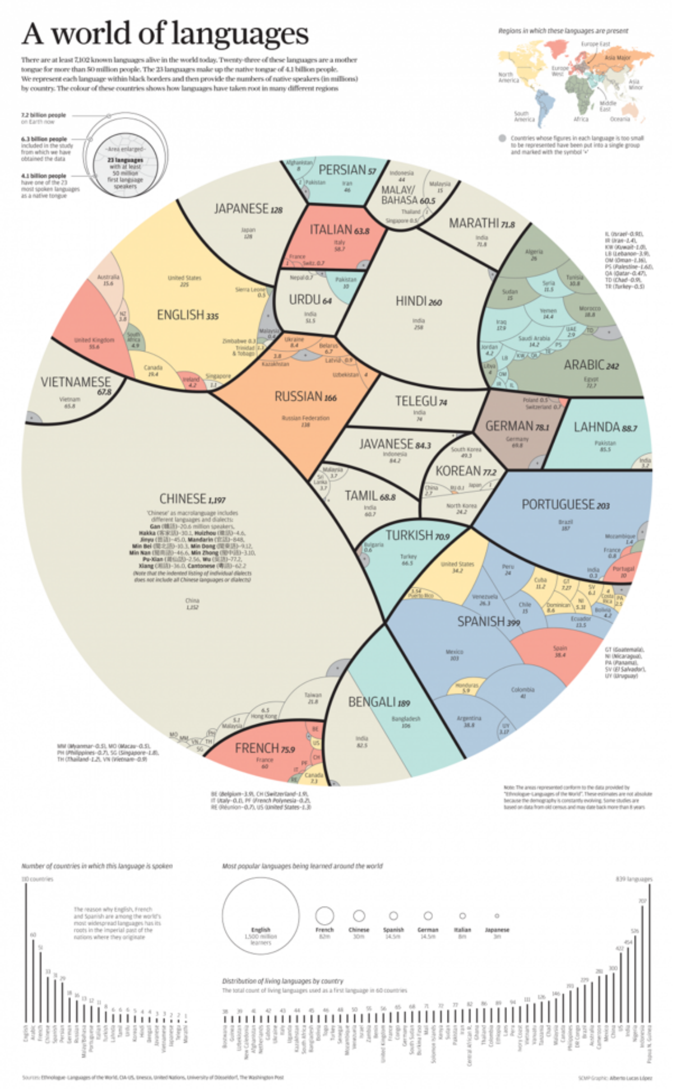
Annotation

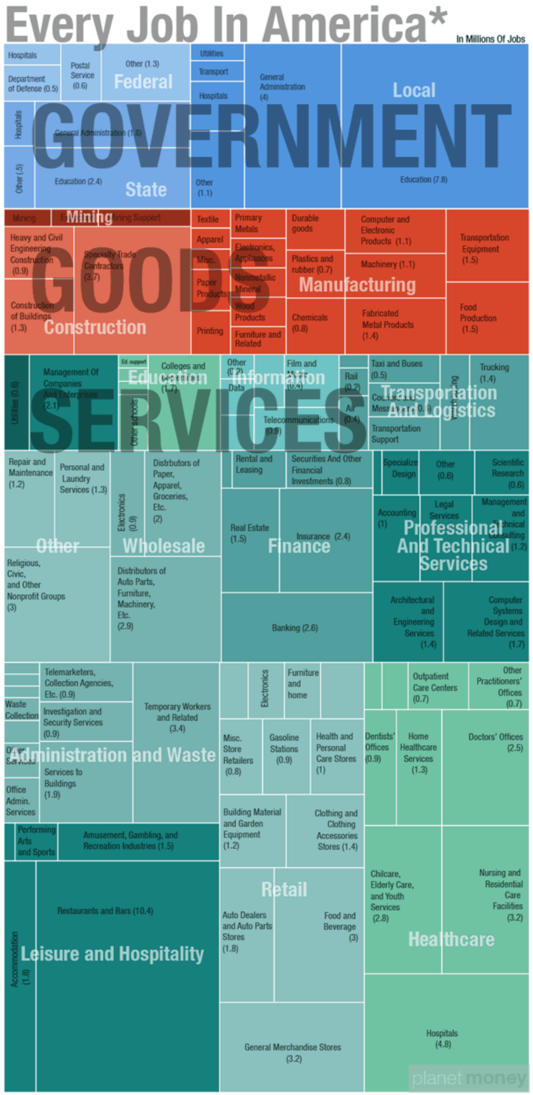


```
  51 volgesteinDriver.oncogene
  58 volgesteinDriver.TSG
 101 cellCycleGene.
 114 apoptosisGene.
 122 dnaRepairGene.
 339 cancerPathwayGene.
 385
1375 Metabolic_Genes_RCT
2646 essentialGene.
2905 immuneResponseGene.
9969 nonEssentialGene.
```

1. Normalize by the #gene
2. Normalize by the length of elements
3. Correct for RT/Trinucleotide context

Some ideas


A world of languages
There are at least 7,102 known languages alive in the world today. Twenty-three of these languages are a mother tongue for more than 50 million people. The 23 languages make up the native tongue of 4.1 billion people. We represent each language within black borders and then provide the numbers of native speakers (in millions) by country. The colour of these countries shows how languages have taken root in many different regions
7.2 billion people on Earth now
6.3 billion people included in the study from which we have obtained the data
4.1 billion people have one of the 23 most spoken languages as a native tongue
Area enlarged
23 languages with at least 50 million first language speakers
Regions in which these languages are present
PERSIAN 57
MALAY/BAHASA 60.5
JAPANESE 128
ITALIAN 63.8
MARATHI 71.8
ENGLISH 335
URDU 64
HINDI 260
ARABIC 242
VIETNAMESE 67.8
RUSSIAN 166
TELEGU 74
GERMAN 78.1
LAHNDA 88.7
JAVANESE 84.3
KOREAN 77.2
CHINESE 1,197
TAMIL 68.8
PORTUGUESE 203
TURKISH 70.9
SPANISH 399
BENGALI 189
FRENCH 75.9
Number of countries in which this language is spoken
Most popular languages being learned around the world
English 1,500 million learners
French
Chinese
Spanish
German
Italian
Japanese
Distribution of living languages by country


Every Job In America*
In Millions Of Jobs
GOVERNMENT
Federal
Hospitals
Department of Defense (0.5)
Postal Service (0.6)
Other (1.3)
State
Hospitals
General Administration (1.8)
Other (.5)
Education (2.4)
Local
Utilities
Transport
Hospitals
General Administration (4)
Other (1.1)
Education (7.8)
GOODS
Mining
Mining
Extraction
Mining Support
Construction
Heavy and Civil Engineering Construction (0.9)
Specialty Trade Contractors (3.7)
Construction of Buildings (1.3)
Manufacturing
Textile
Apparel
Misc.
Paper Products
Printing
Primary Metals
Electronics, Appliances
Nonmetallic Mineral
Wood Products
Furniture and Related
Durable goods
Plastics and rubber (0.7)
Chemicals (0.8)
Computer and Electronic Products (1.1)
Machinery (1.1)
Fabricated Metal Products (1.4)
Transportation Equipment (1.5)
Food Production (1.5)
SERVICES
Utilities (0.6)
Management Of Companies And Enterprises (2.1)
Education
Ed. support
Colleges and Universities (1.7)
Other schools
Information
Other (0.2)
Data
Film and Music (0.4)
Telecommunications (0.9)
Transportation And Logistics
Rail (0.2)
Air (0.4)
Taxi and Buses (0.5)
Couriers and Messengers (0.5)
Transportation Support
Warehousing
Trucking (1.4)
Other
Repair and Maintenance (1.2)
Personal and Laundry Services (1.3)
Religious, Civic, and Other Nonprofit Groups (3)
Wholesale
Electronics (0.9)
Distrbutors of Paper, Apparel, Groceries, Etc. (2)
Distributors of Auto Parts, Furniture, Machinery, Etc. (2.9)
Finance
Rental and Leasing
Securities And Other Financial Investments (0.8)
Real Estate (1.5)
Insurance (2.4)
Banking (2.6)
Professional And Technical Services
Specialize Design
Other (0.6)
Scientific Research (0.6)
Accounting (1)
Legal Services
Management and Technical Consulting (1.2)
Architectural and Engineering Services (1.4)
Computer Systems Design and Related Services (1.7)
Administration and Waste
Waste Collection
Telemarketers, Collection Agencies, Etc. (0.9)
Investigation and Security Services (0.9)
Temporary Workers and Related (3.4)
Services to Buildings (1.9)
Office Admin. Services
Leisure and Hospitality
Performing Arts and Sports
Amusement, Gambling, and Recreation Industries (1.5)
Restaurants and Bars (10.4)
Accommodation (1.8)
Retail
Electronics
Furniture and home
Misc. Store Retailers (0.8)
Gasoline Stations (0.9)
Health and Personal Care Stores (1)
Building Material and Garden Equipment (1.2)
Clothing and Clothing Accessories Stores (1.4)
Auto Dealers and Auto Parts Stores (1.8)
Food and Beverage (3)
General Merchandise Stores (3.2)
Healthcare
Outpatient Care Centers (0.7)
Other Practitioners' Offices (0.7)
Dentists' Offices (0.9)
Home Healthcare Services (1.3)
Doctors' Offices (2.5)
Childcare, Elderly Care, and Youth Services (2.8)
Nursing and Residential Care Facilities (3.2)
Hospitals (4.8)
planet money