PCAWG 2,5,9,14

Verona

Verona Agenda for Day 0 (Sun, Feb 15)

Working Group Presentations 10.10 am – 4.30 pm
o Opportunity for scientific update on projects, technical developments, ongoing pancancer analyses
o 20-minute presentations for each working group (4 data slides):
§ (2 minutes) statement of mission and scope of working group
§ (3 minutes) expected outputs
§ (5 minutes) current status

10.10 – 10.30 PAWG-2: Analysis of mutations in regulatory regions Gaddy Getz (Broad Institute), Mark Gerstein (Yale University)
10.30 – 10.50 PAWG-9: Inferring driver mutations and identifying cancer
genes and pathways Michael Lawrence (Broad Institute), Nuria López-Bigas (University Pompeu Fabra)
10.50 – 11.10 PAWG-5: Consequences of somatic mutations on pathway and network activity Ben Raphael (Brown University), Josh Stuart (UCSC)
11.10 – 11.30 PAWG-14: Analysis of mutations in non-coding RNA Daniel Hughes (Baylor College of Medicine) representing David Wheeler (Baylor College of Medicine), Jakob Skou Pedersen (Aarhus University)

So we have 10:10-11:30 including discussion

Outline

1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
   a) Collecting common resources
   b) Pilot datasets
   c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
   a) Datasets pilot-50 (Train 1) Broad calls, public-607
   b) Compare submitted annotations
   c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
   a) TCGA-505, GBM-27 (Nuria)
   b) Simulated data (Gaddy, Inigo)
   c) Compare submitted significance analyses (Nuria)
4) Example of downstream analyses
   a) Analysis of non-coding RNA (Jakob) - annotation compilation, miRNA profiling (slides from Todd Johnson)
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

Outline

1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
    a) Collecting common resources
    b) Pilot datasets
    c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
    a) Datasets pilot-50 (Train 1) Broad calls, public-607
    b) Compare submitted annotations
    c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
    a) TCGA-505, GBM-27 (Nuria)
    b) Simulated data (Gaddy, Inigo)
    c) Compare submitted significance analyses (Nuria)
4) Example of downstream analyses
    a) Analysis of non-coding RNA (Jakob) - annotation compilation, miRNA profiling (slides from Todd Johnson)
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

PAWG-2-5-9-14 - merged group

PAWG-2: Analysis of mutations in regulatory regions
Gaddy Getz (Broad Institute), Mark Gerstein (Yale University)

PAWG-5: Consequences of somatic mutations on pathway and network activity
Ben Raphael (Brown University), Josh Stuart (UCSC)

PAWG-9: Inferring driver mutations and identifying cancer genes and pathways
Michael Lawrence (Broad Institute), Nuria López-Bigas (University Pompeu Fabra)

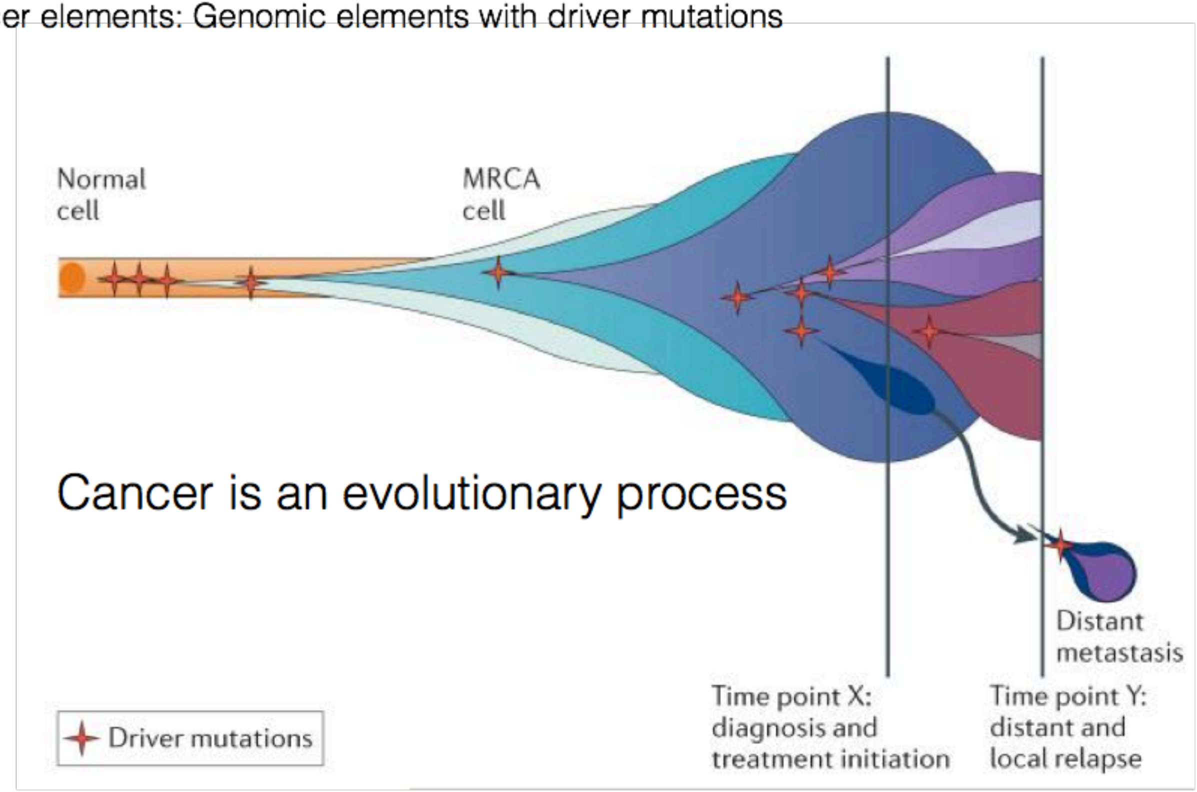
PAWG-14: Analysis of mutations in non-coding RNA
David Wheeler (Baylor College of Medicine), Jakob Skou Pedersen (Aarhus University)

One common objective: Identify driver mutations

Drivers versus Passengers

- Driver mutations: Confer selective advantage to tumour cells
- Passenger mutations: Do not confer selective advantage to tumour cells

- Cancer elements: Genomic elements with driver mutations


Normal cell
MRCA cell
Distant metastasis
Time point X: diagnosis and treatment initiation
Time point Y: distant and local relapse
✦ Driver mutations

Cancer is an evolutionary process

Yates and Campbell et al, Nat Rev Genet 2012

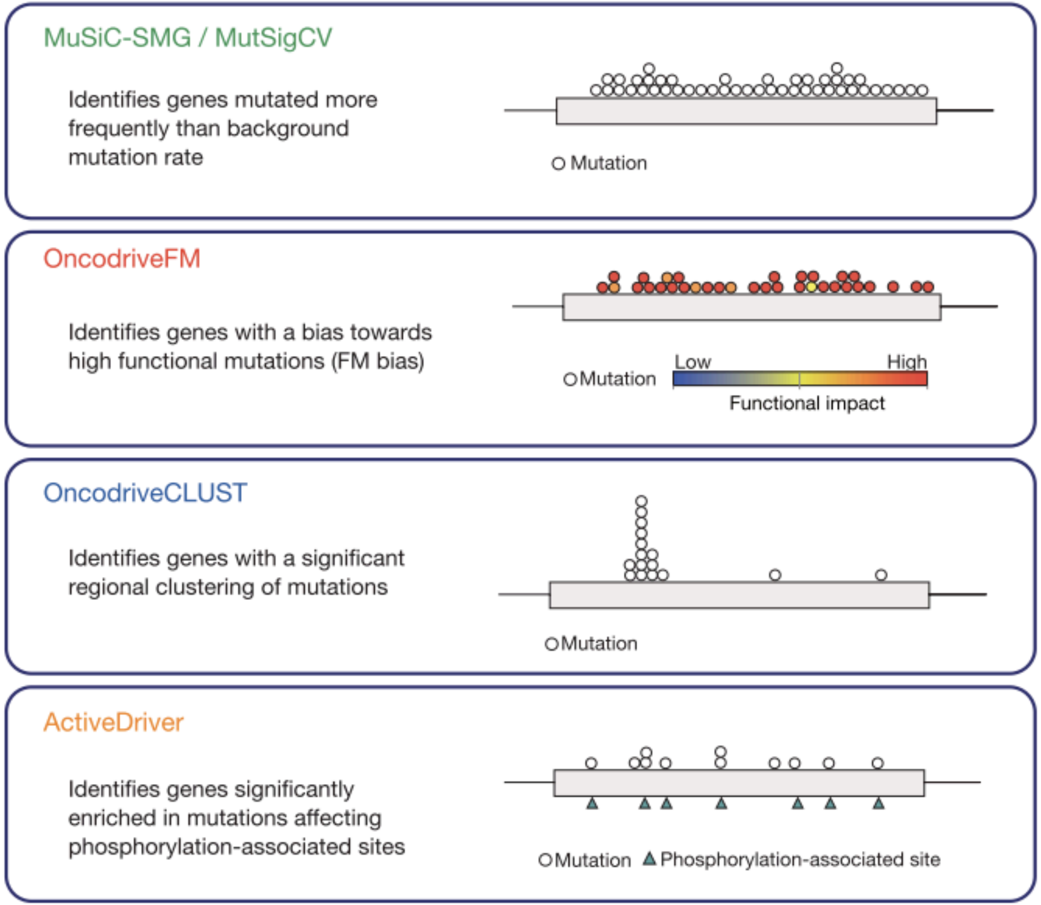
Tasks of our merged group

- **Variant level**: Annotate and score individual variants
- **Element level**: Find elements with signals of positive selection in the pattern of mutations
- **Pathway/Network level**: Identify cancer relevant modules

Expected outputs

- Mutations with extensive annotations
- Catalog of cancer elements with signals of positive selection
- Cancer modules (networks/pathways)

Detect signals of positive selection



MuSiC-SMG / MutSigCV

Identifies genes mutated more frequently than background mutation rate

OncodriveFM

Identifies genes with a bias towards high functional mutations (FM bias)

OncodriveCLUST

Identifies genes with a significant regional clustering of mutations

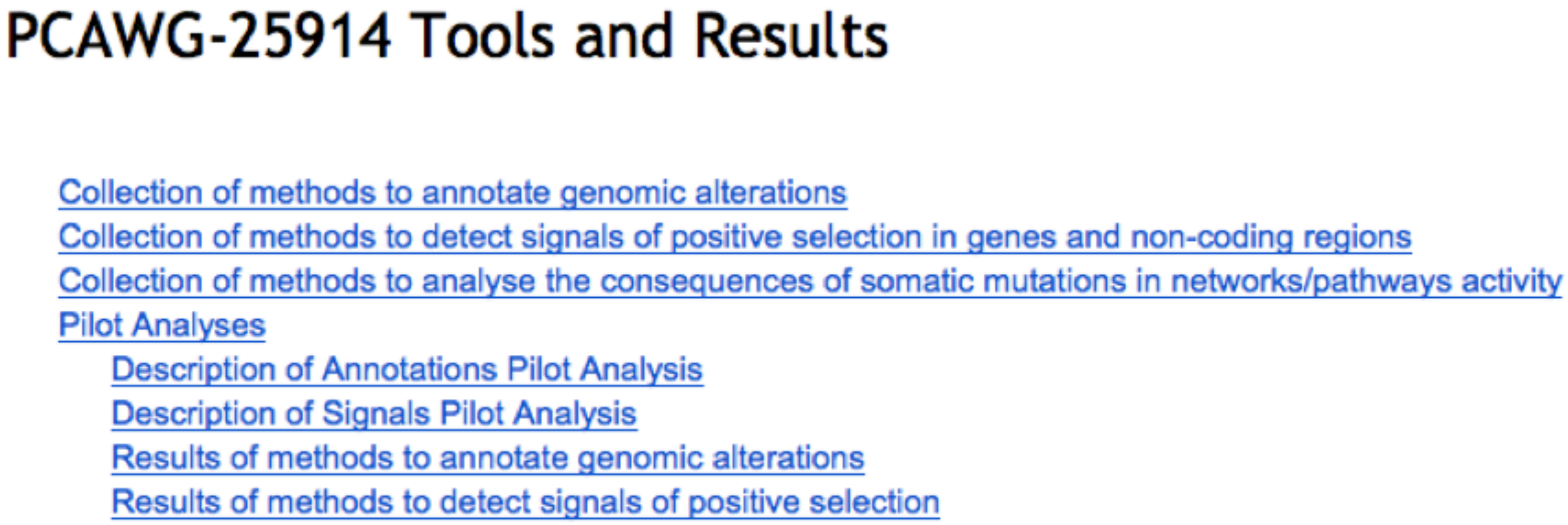
ActiveDriver

Identifies genes significantly enriched in mutations affecting phosphorylation-associated sites

Tamborero et al., 2013

Collection of methods and results

goo.gl/MmuWmt



PCAWG-25914 Tools and Results

- Collection of methods to annotate genomic alterations
- Collection of methods to detect signals of positive selection in genes and non-coding regions
- Collection of methods to analyse the consequences of somatic mutations in networks/pathways activity
- Pilot Analyses
    - Description of Annotations Pilot Analysis
    - Description of Signals Pilot Analysis
    - Results of methods to annotate genomic alterations
    - Results of methods to detect signals of positive selection

TABLE A: Methods to annotate genomics alterations

Method	Authors	Description	Coding genes	Promoters	Enhancers	UTRs	lncRNAs	microRNAs	tRNA	...
FunSeq	Ekta Khurana and Yao Fu (Mark Gerstein's lab)	Identifies somatic mutations predicted to have high functional impact, specially noncoding ones	XXX	XXX	XXX	XXX	X	X		
3D_SNP	Francisco Martínez-Jiménez (Marti-Renom Lab)	Functional impact of non-synonymous SNPs in modeled 3D structures of proteins from coding-regions of the genome.	X							
wKinMut	Jose MG Izarzugaza (CBS/DTU) and Alfonso Valencia Lab	Analysis and classification of mutations in protein kinases *.	XX							
CanDrA	Ken Chen lab	Identify the driver potential of somatic mutations	XXX							
	Todd A. Johnson (Tsunoda Lab/RIKEN)	Functional classification of germline or somatic variants. Includes annotation of miRNA related elements (genes, predicted promoters, target-sites)	X	X	X	X		X		
IGR (Intra-Genomic replicates)	Sallari & Sinnott-Armstrong (Kellis lab, MIT & Broad)	Prediction of affinity modulation based on ENCODE transcription factor ChIP-seq data.		XXX	XXX					
MutationAssessor	Reva, Antipin, Sheridan, Sander (MSKCC)	Functional impact of AA-changing mutations; somatic or germline; also mapped to 3D in mutation tab of cbioportal.org	XXX							
AGO-CLIP target Atlas **miSNP algorithm**	Hamilton, Coarfa, Wheeler, McGuire (BCM)	List of AGO-CLIP validated miRNA target sites annotated by recurrence. Currently updated to GC19-NC-extended transcriptome. We are generating novel CLIP data in multiple tumor cell lines to compliment ICGC analysis. The miSNP algorithm identifies mutations significantly enriched in CLIP target sites and determines if these correspond to changes in complementary RNA-seq data from the same tumor.				XXX				
DKFZ Pipeline	Jäger, Hutter, Buchhalter, Schlesner, Feuerbach, et al. (DKFZ Heidelberg)	Identifies somatic point mutations and small indels Annotates functional consequences Integrates external databases Filters high-confidence calls	XXX	XXX	XX	XXX	XXX	X		
AncestralAlleles	Javier Herrero	Identifies SNV and small indels that revert to the ancestral state (and are therefore less likely to be driver)								
Oncotator	Alex Ramos, Lee Lichtenstein, Gaddy Getz	Comprehensive annotation of variants	XXX	XXX			XXX	XXX	XX	

TABLE B: Methods to detect signals of positive selection

Method	Authors	Description	PCAWG input	External input (if any)	Coding genes	Promoters	Enhancers	UTRs	lncRNAs	microRNAs	tRNA
OncodriveFM	Lopez-Bigas lab	Identifies genes/elements with a significant bias towards the accumulation of functional variants	List of tumor somatic mutations	-	XXX	XX	XX	XX	XX	XX	
OncodriveCLUST	Lopez-Bigas lab	Identifies genes/elements with mutations significantly clustered in particular regions	List of tumor somatic mutations	-	XXX	X	X	X	X	X	
Two methods inspired on dN/dS	Inigo Martincorena (Peter Campbell's lab)	They identify genes and non-coding elements with significant recurrence, considering the mutation spectrum, the sequence composition and the variation of the mutation rate along the genome, with or without covariates. Ready to run but unpublished for WGS.			XXX	XX	XX	XX	XX	XX	
LARVA	Lucas Lochovsky (Mark Gerstein's lab)	Identifies elements with more recurrent mutations than expected randomly			XX	XX	XX	XX	X	X	
ActiveDriver	Jüri Reimand (Gary Bader's Lab)	Site-specific mutational enrichment analysis of genes and other genomic regions			XXX	XX	XX	XX	XX	XX	
MIMP	Jüri Reimand, Mohamed Helmy, Omar Wagih (Gary Bader Lab)	Predicting mutational rewiring of sequence elements			XXX	X	X	X	X	X	
ExInAtor	Rory Johnson / Andres Lanzos (Roderic Guigo Lab)	Identifies lncRNAs with excess of exonic mutations. First version ready, undergoing testing.							XX		
InterScreener	Lars Feuerbach (Brors Lab)	Integrative screener for functional non-coding SNVs. Integrates SNVs, CNV, SVs, mRNA, miRNA and methylation data				XX	X	XX			
3D_permutation	Akihiro Fujimoto (Riken)	Analysis of mutation clusters in 3D protein structures. Applied to Riken liver cancer data and COSMIC data. Functional analysis will be started.			XX				X		
ncDriver	Henrik Hornshøj (Jakob Skou Pedersen lab)	Multi-step significance evaluation of mutation rates and intensities in non-coding elements. Combines four separate tests on: intensity, cancer type specificity, local conservation, & global conservation.			XX	XX	XX	XX	XX	XX	
rwClust	Jakob Skou Pedersen lab	Significance evaluation of mutation clusters within genomic elements using Random Walk theory.			X	X	X	X	X	X	
Significance evaluation of mutational hot spots	Jakob Skou Pedersen & Asger Hobolth labs	Significance evaluation of mutational hotspots based on probabilistic null model capturing different levels of mutational heterogeneity (between samples, along genome, mutational context).			X	X	X	X	X	X	
Identification of driver mutational hotspots	Ken Chen lab	Identification of driver mutational hotspots in a knowledge based statistical model (cancer type-specific, gene-specific, sequence context, etc)			XX	X	X	X	X	X	
MuSiC2 - Mutation Significance in Cancer:	Ding Lab	A suite of tools equipped to identify genetic loci contributing to cancer on the gene, pathway, and clinical level. Calculations of significance incorporate mutation rates, protein databases, drug databases, and previous literature.			XXX	XX	XX	XX	XX	XX	
Onkomers	Calvin Chan, Carl Herrmann (DKFZ Heidelberg, Germany)	Patterns of significantly altered kmers (either created or disrupted) using background model. Assembly of kmers clusters into longer motifs/PWMs				X	X	X	X	X	
Plexus recurrence test	Sallari & Sinnott-Armstrong (Kellis lab, MIT & Broad)	Identifies recurrently mutated plexi (gene body and interacting regulatory elements).	List of somatic mutations	Hi-C and ChIA-PET	X	XX	XXX	X			
Genomic Recurrence	Lee, Weinhold, Schultz, Sander	Analyzes recurrence in non-protein-coding regions	Somatic mutations	promoters, UTRs, etc.		XXX	XXX	XXX			

TABLE C: Pathway/Network methods

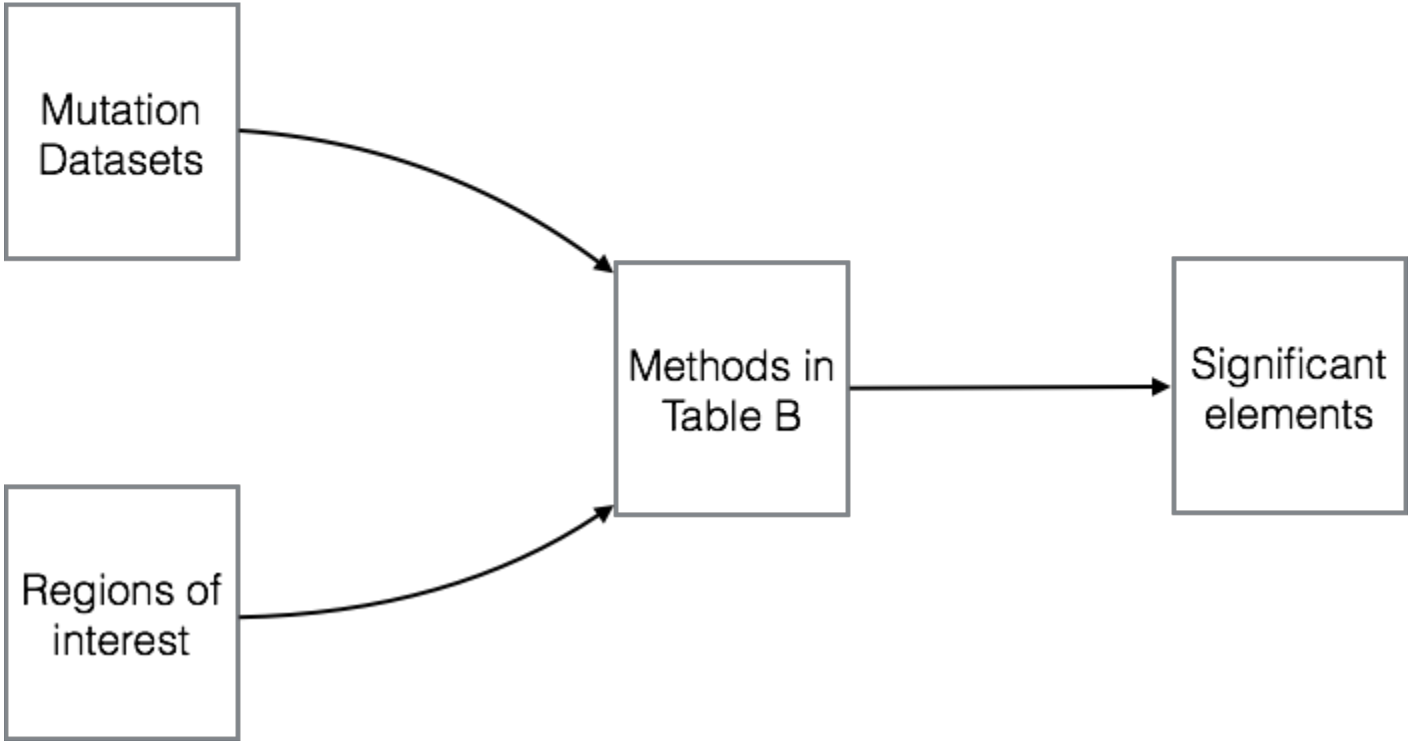
Method	Authors	Description	Coding	nonCoding	fusions	mRNA-GL	mRNA-AS	SCNA	epi	external
HotNet2	Raphael	Subnetworks of mutated genes	XXX	X	X			XXX		iRef, HPRD, MultiNet
Dendrix and CoMEt	Raphael	Mutually exclusive genomic alterations	XXX	X	X	XXX		XXX		
Paradigm-Shift	Stuart	Predicts GOF/LOF of genes using pathway neighborhood	XXX	X	X	XXX		XXX		NCI-PID, Reactome, KEGG
String-based	Christian von-Mering		X							STRING
Firestar	Alfonso Valencia	Predict if mutations affect ligand and drug binding sites	XXX							
Co-evolutionary analysis	Alfonso Valencia	Co-evolutionary networks of mutated genes (ID uncommon cancer genes).	X							
Tumour molecular context delineation using network based data integration	Kathleen Marchal	Prioritized drivers, tumour specific subnetworks, molecular subtypes. Per molecular tumour subtype: a subnetwork enriched for combinations of mutations, connecting genetic aberrations with downstream molecular phenotype	X	X		X		X	X	KEGG, ENCODE, ...
PhaC	Kathleen Marchal	Finds mutual exclusivity patterns by small subnetwork analysis with reinforced learning	X	X		X		X	X	KEGG, NCI-PID, Reactome
Reactome-FI	Lincoln Stein	Integrate multiple data types onto the Reactome FI network, perform network-based clustering, and search for cancer subtype-based network modules	X	X		X		X	X	
FunSeq	Khurana, Fu and Gerstein	Identifies mutations targeting hubs in various networks	XXX	XXX	X			X		Regulatory network from ENCODE; Multinet from Khurana et al, PLoS Comp Bio
g:Profiler, Cytoscape, Enrichment Map	Juri Reimand, Gary Bader	Enrichment of mutations in biological pathways and processes, network visualisation	XXX	XXX		XXX				Gene Ontology, Reactome, KEGG, HPO, miRBase, Transfac.
HyperModules	Juri Reimand, Gary Bader	Network clustering, detection of sub-networks with clinical and survival correlations, linking networks to tumor subtypes	XXX	X		X				molecular interaction networks (PPI, co-expression, TF-DNA interactions)
MIMP	Juri Reimand, Mohamed Helmy, Gary Bader	Impact of mutations on networks, e.g. SNVs in transcription factor binding sites or kinase binding sites, to predict gains and losses of regulatory interactions.	XXX	XX						
HIT'nDRIVE	Raunak Shrestha, Ermin Hodzic, Cenk Sahinalp	Integrates various alterations to its downstream targets (direct/indirect) using network information, prioritizing altered genes as potential drivers.	XXX	X	XX	XXX	X	XXX	X	molecular interaction networks (PPI, TF-DNA interactions)
NetBox	Cerami, Schultz, Liu, Sander	Discovers oncogenically altered pathway modules	any alteration type							Pathway Commons
MutEx	Fredriksson, Larsson-Lekholm	Uncovers associations between somatic regulatory mutations and mRNA level changes (individual genes)	XXX	XXX		XXX	X	XXX	X	Regulatory region annotation, e.g. DNaseI HS sites. No pathway data used.
Oncotator	Alex Ramos, Lee Lichtenstein, Gaddy Getz	Comprehensive annotation of variants	XXX	XXX			XXX	XXX	XX	Uses many external data sources(version numbers are provided in the header of the annotated file)

Pilot Analyses

Annotation Pilot
coordinated by
Ekta Khurana

Signals Pilot
coordinated by
Nuria Lopez-Bigas

Signals Pilot



Mutation Datasets

Regions of interest

Methods in Table B

Significant elements

Signals Pilot - Mutation Datasets

- TCGA-505 (pan)
- GBM-27
- Simulated TCGA-505
- Simulated GBM-27

Fredriksson et al., Nature Genetics 2014

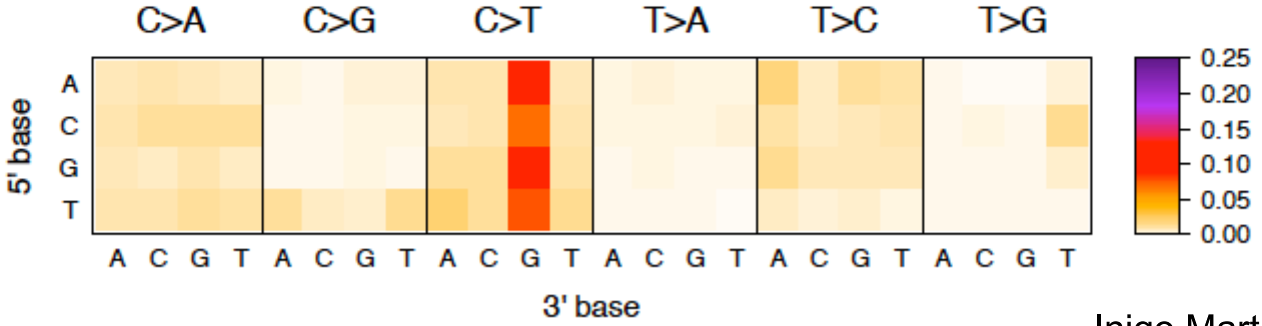
Simulated data prepare by Inigo Martincorena (Sanger)

Simulated data

Simulated data of public-607
Retained:

- Same nucleotide rates (base context)
- Same regional variation of mutation rate (per Mb)
- Same distribution of mutations across samples and tissue types

Not as challenging as true data but significant genes indicate inadequate background model



Inigo Martincorena, Sanger Institute

Signals Pilot - Regions of interest

- Promoters
- Coding regions

Prepared by Jakob Skou Pedersen

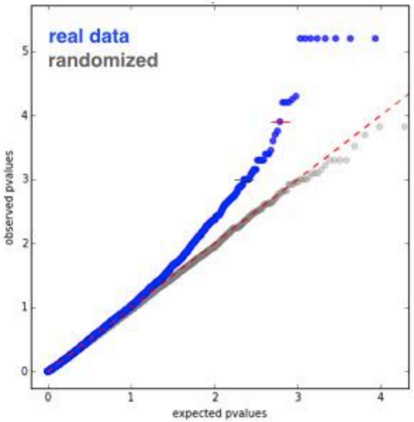
Signals Pilot - Results

goo.gl/MmuWmt

Method	Author (email)	Description	In which types of elements has been run?	Synapse ID
ncDriver	Henrik Hornshøj (hhj@clin.au.dk)	ncDriver CDS and promoter drivers detected by analysis of pilot TCGA50	Protein-coding genes, Promoter	syn3163011
NBR - Sanger	Inigo Martincorena (im3@sanger.ac.uk)	Recurrence by negative binomial regression with covariates. Applied on the pilot CDS and promoter databases on TCGA505, GBM27 and the randomised control datasets.	Protein-coding genes, Promoter	syn3163124
OncodriveFM2	Loris Mularoni (loris.mularoni@upf.edu)	Functional impact bias. Run on TCGA505, GBM27 and the randomised control datasets.	Protein-coding genes, Promoter	syn3163827
MSK-Hotspots	William Lee (leew1@mskcc.org)	Recurrently mutated genomic hotspots calculated as described in Weinhold et al. 201	Protein-coding genes, Promoter	TCGA505: syn3163614 GBM: syn3163617 Randomised: syn3163620
MSK-Regions	Anders Jacobsen Skanderup (jacobsen@cbio.mskcc.org)	Recurrently mutated genomic regions calculated as described in Weinhold et al. 201	Protein-coding genes, Promoter	syn3163754
PhaC	Sergio Pulido-Tamayo (spulido99@gmail.com)	Mutual exclusivity patterns by small subnetwork analysis	CDS	syn3163695
OncoMotifs	c.herrmann@dkfz.de	Patterns of PWM creation/disruption using a local randomized background model	non-coding regions	syn3165097
3D permutation	Akihiro Fujimoto, afujircb@src.riken.j	Mutation cluster in 3D protein structure detected by analysis of pilot TCGA50	CDS	syn3168511
MutSig2CV	lawrence@broadinstitute.org	Analysis of mutation significance based on deviations from background model	Protein-coding genes, Promoters	syn3193626

Signals Pilot - Results - TCGA-505 coding

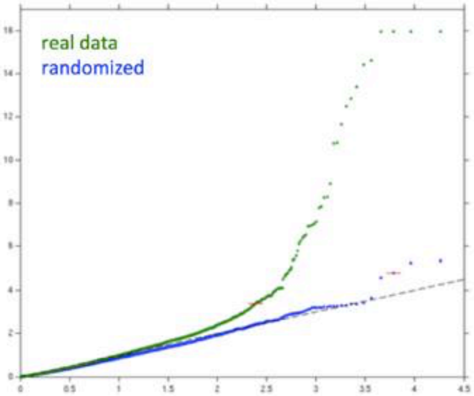
OncodriveFM2



CDKN2A
KRAS
ARID1A
PIK3CA
PTEN
KMT2D
TP53
APC
ATM
NF1
RBM10
DACT1
FBXW7
BCL9
SIRPB1
ARMCX3
PIK3R1
ADAMTS20
PAPD4
PUS7
FFAR1
CHDC2
AP003062.1
NIM1K
MGA

FOXA2
ZNF781
TRIM6
AMER1
OR10H2
ARPP21
C2CD5
MKL2
ZC3H13
DCLRE1A
RB1
VHL
NOTCH1
DNPEP

MutSig2CV

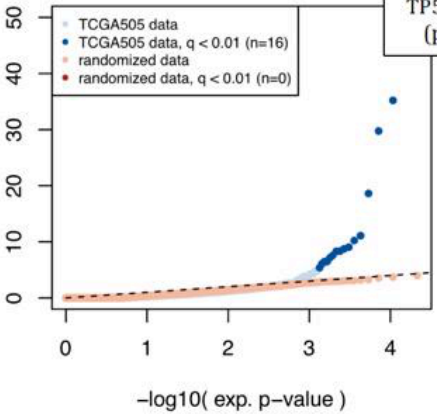


TP53
PIK3CA
BRAF
IDH1
CDKN2A
PTEN
NFE2L2
VHL
APC
KRAS
STK11
NRAS
MUC4
HLA-DRB1
ARID1A
RNF43
ARHGAP35
HLA-B
CBFB
B2M
PRSS3
FBXW7
POLDIP2
LARP4B
CHD1

MPHOSPH9
PITRM1
MUC17
KRTAP4-5
ACVR1B
MKL2
LCT
SLC38A1
USP6
ZFP36L2
BCLAF1
ATXN1
IL23R
C15orf23
THEMIS
TBXA2R
HRC
CTCF
INPPL1
IL7R
PCDHGA1
EYA4
TGFBR2
MYH7B
SVIL

EGFR
DNAH12
DCAF4L1
AVPR1B
HRNR
HRAS
PABPC1
NOTCH1
HLA-A
OSMR
ATP1A4
MRPL32
FGFR3
TTN
IL10RB
PZP
TACC3
WNT16
CDH23
APOL2
ACOT4
CTNNB1

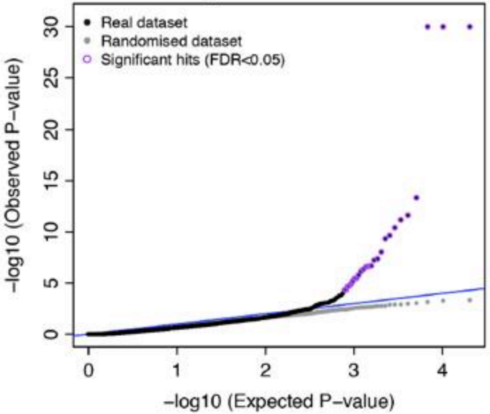
MSK-regions



TP53 not shown
(p = 3e-195)

TP53
PIK3CA
KRAS
PTEN
AL390778.1
FAM230A
BRAF
IDH1
TBP
GOLGA6L2
CDKN2A
AC093323.1
NRAS
VHL
RPL22
MUC4

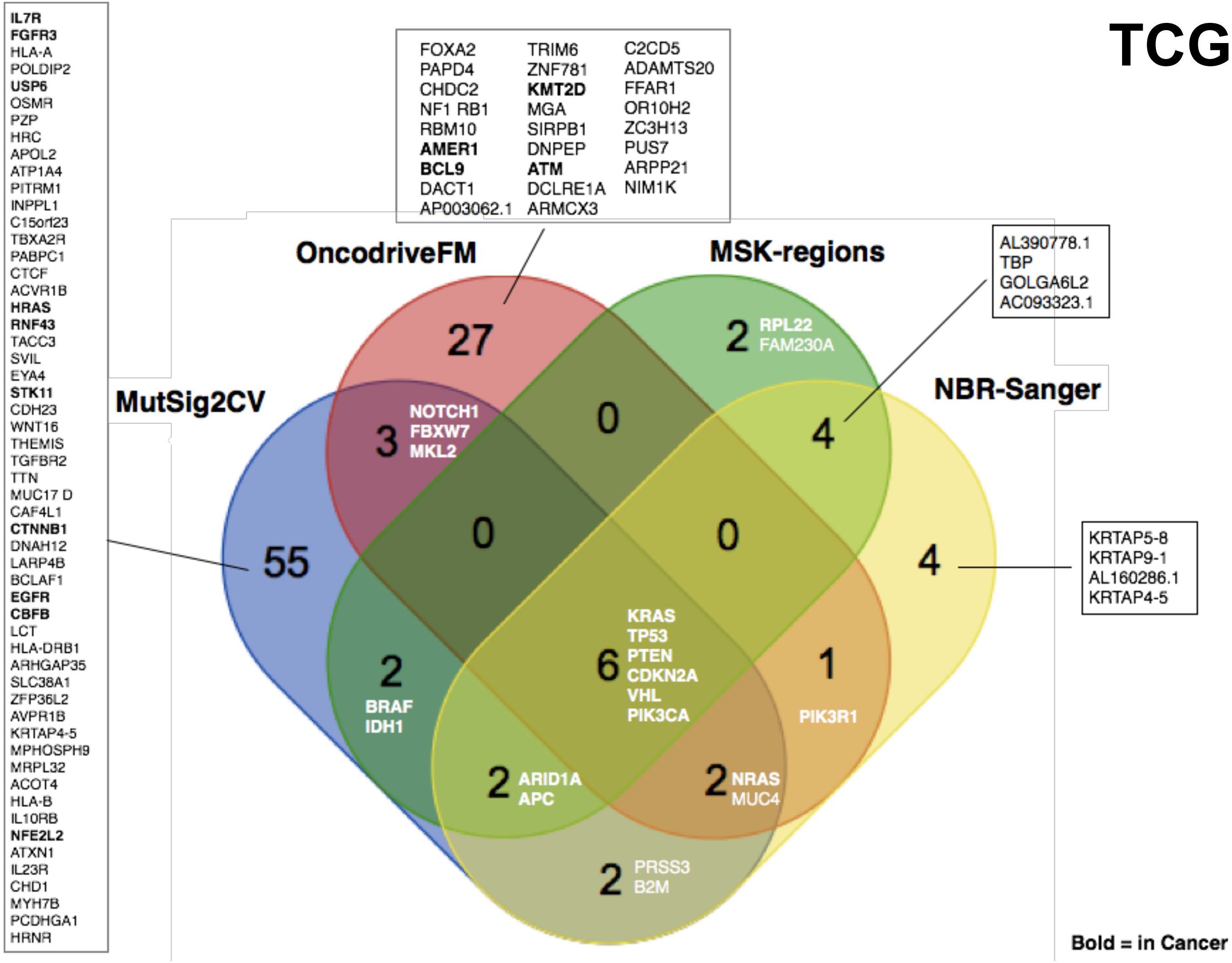
NBR-Sanger



PTEN
KRAS
TP53
PIK3CA
AC093323.1
KRTAP9-1
PRSS3
VHL
KRTAP4-5
TBP
CDKN2A
AL390778.1
KRTAP5-8
PIK3R1
BRAF
APC
AL160286.1
GOLGA6L2
IDH1
B2M
ARID1A

Bold = in Cancer Gene Census

TCGA-505 coding



IL7R
FGFR3
HLA-A
POLDIP2
USP6
OSMR
PZP
HRC
APOL2
ATP1A4
PITRM1
INPPL1
C15orf23
TBXA2R
PABPC1
CTCF
ACVR1B
HRAS
RNF43
TACC3
SVIL
EYA4
STK11
CDH23
WNT16
THEMIS
TGFBR2
TTN
MUC17 D
CAF4L1
CTNNB1
DNAH12
LARP4B
BCLAF1
EGFR
CBFB
LCT
HLA-DRB1
ARHGAP35
SLC38A1
ZFP36L2
AVPR1B
KRTAP4-5
MPHOSPH9
MRPL32
ACOT4
HLA-B
IL10RB
NFE2L2
ATXN1
IL23R
CHD1
MYH7B
PCDHGA1
HRNR

FOXA2
PAPD4
CHDC2
NF1 RB1
RBM10
AMER1
BCL9
DACT1
AP003062.1
TRIM6
ZNF781
KMT2D
MGA
SIRPB1
DNPEP
ATM
DCLRE1A
ARMCX3
C2CD5
ADAMTS20
FFAR1
OR10H2
ZC3H13
PUS7
ARPP21
NIM1K

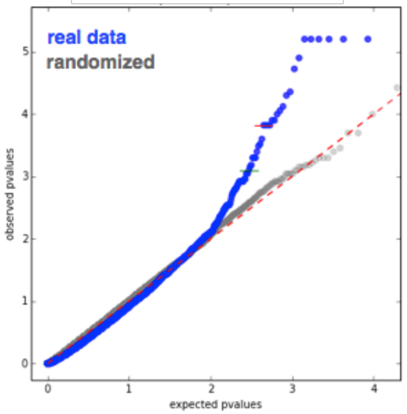
AL390778.1
TBP
GOLGA6L2
AC093323.1

KRTAP5-8
KRTAP9-1
AL160286.1
KRTAP4-5

Bold = in Cancer Gene Census

Signals Pilot - Results - TCGA-505 promoter

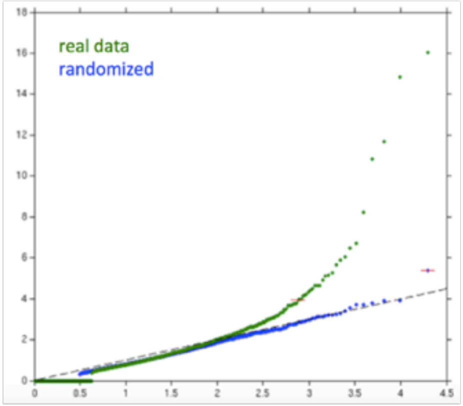
OncodriveFM



TERT
ALYREF
ARHGEF18
MRPS31
POLR2D
SYF2
ZNF324
FBXO18
ZNF343
ASXL2
PPP4C
C16orf13
DNAJC28
MSH5-SAPCD1
CDC20
CDC37
SYT15
TRMT10C
SPN
IGSF5
TOR1B
C16orf91
ZNF160

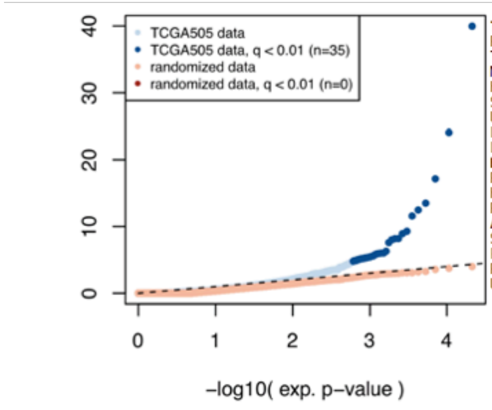
TRIM3
CORO6
KLHDC1
ILF2
NFKB2
YIPF1

MutSig2CV



NEXN
TERT
SNX32
ZNF717
CSF2RA
SMG1
TTC40
MUC12
AKAP17A
TRIOBP
RP11-1220
K2.2
BCL9
MROH2A
RNF219
CCDC66
RIC8A
AP2A1
C19orf112
RNF17
TDRD15
KNTC1
SEMA6B
STAG3
ATP10D
EIF2AK3
GOLGA6L2

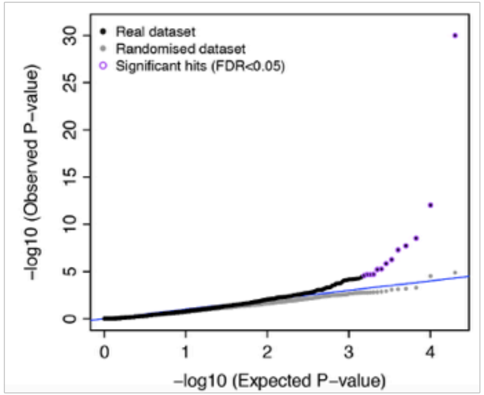
MSK-regions



TERT
RP11-190A12.7
TMEM9B
MUC3A
HIST1H2AM
SNX16
UBE2K
LCN8
LY6K
BCL9
RNF185
KRTAP1-3
RP11-597K23.2
ACBD7
SSBP1
PRAMEF4
RNF219
URB1

STX6
S1PR4
OR1E1
C11orf30
OR7G3
CCL26
RP11-481A20.11
EN1
TM4SF18
IFIT5
TMX2-CTNND1
TMX2
CD300A
C6orf62
ZNF805
CASC4
FUCA1

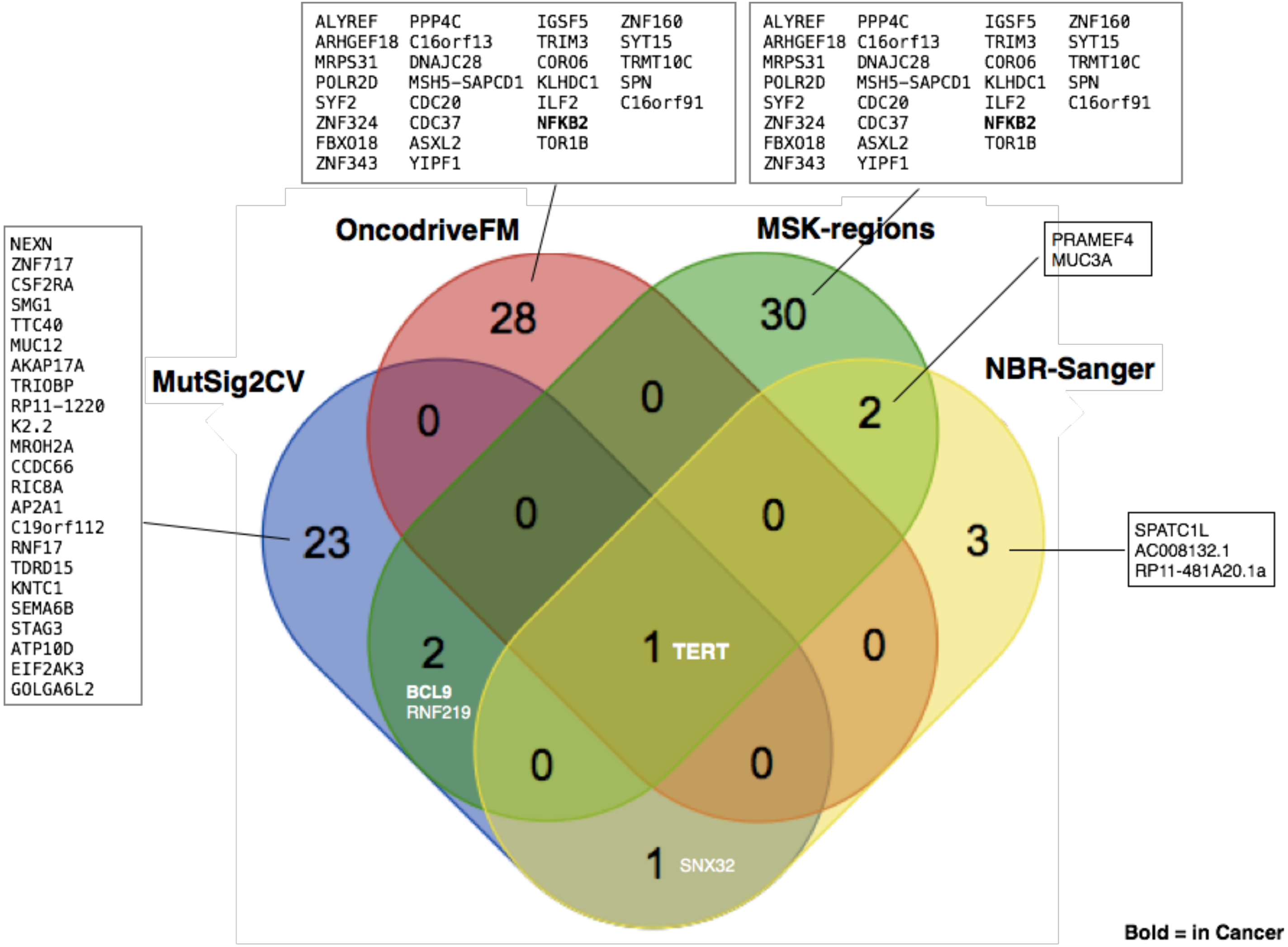
NBR-Sanger



TERT
MUC3A
RP11-481A20.1a
AC008132.1
SNX32
SPATC1L
PRAMEF4

Bold = in Cancer Gene Census

TCGA-505 promoter



ALYREF
ARHGEF18
MRPS31
POLR2D
SYF2
ZNF324
FBXO18
ZNF343
PPP4C
C16orf13
DNAJC28
MSH5-SAPCD1
CDC20
CDC37
ASXL2
YIPF1
IGSF5
TRIM3
CORO6
KLHDC1
ILF2
NFKB2
TOR1B
ZNF160
SYT15
TRMT10C
SPN
C16orf91

ALYREF
ARHGEF18
MRPS31
POLR2D
SYF2
ZNF324
FBXO18
ZNF343
PPP4C
C16orf13
DNAJC28
MSH5-SAPCD1
CDC20
CDC37
ASXL2
YIPF1
IGSF5
TRIM3
CORO6
KLHDC1
ILF2
NFKB2
TOR1B
ZNF160
SYT15
TRMT10C
SPN
C16orf91

NEXN
ZNF717
CSF2RA
SMG1
TTC40
MUC12
AKAP17A
TRIOBP
RP11-1220
K2.2
MROH2A
CCDC66
RIC8A
AP2A1
C19orf112
RNF17
TDRD15
KNTC1
SEMA6B
STAG3
ATP10D
EIF2AK3
GOLGA6L2

PRAMEF4
MUC3A

SPATC1L
AC008132.1
RP11-481A20.1a

BCL9
RNF219

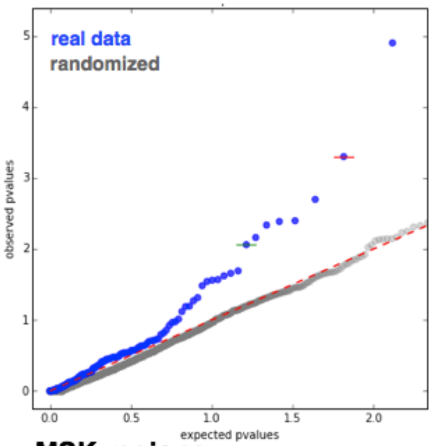
TERT

SNX32

Bold = in Cancer Gene Census

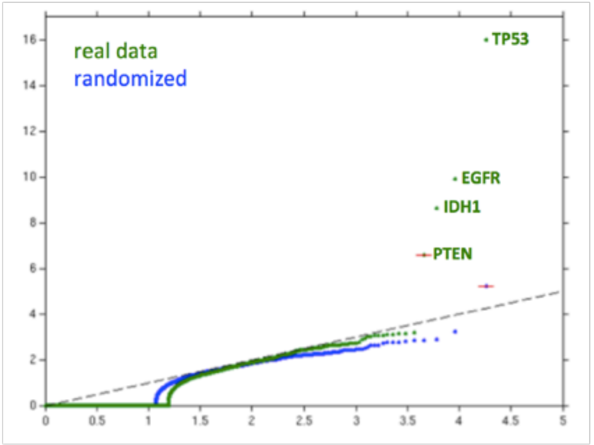
Signals Pilot - Results - GBM-27 coding

OncodriveFM

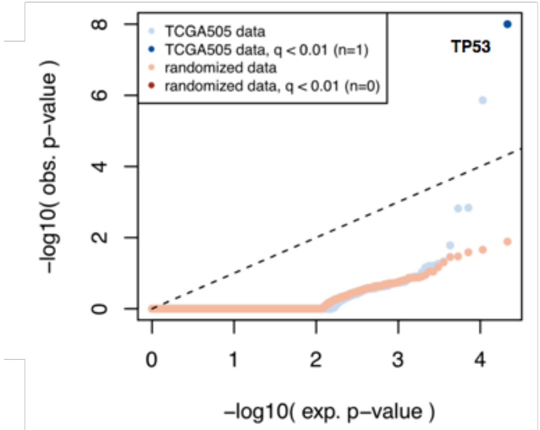


TP53
RIMS2
UGGT1
IDH1
CTSG
EGFR
ZNF208
TLL1

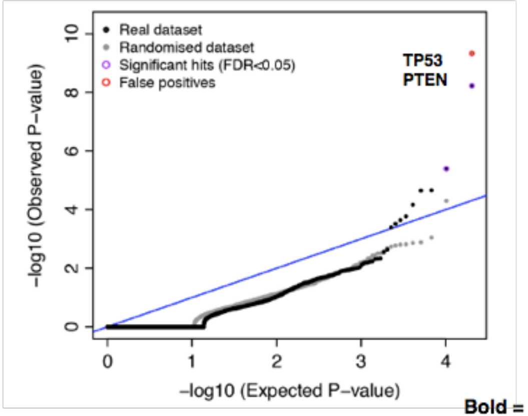
MutSig2CV



MSK-regions



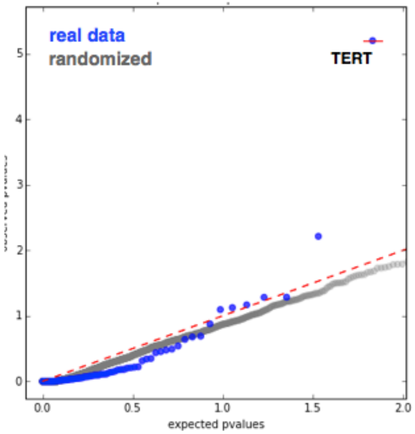
NBR-Sanger



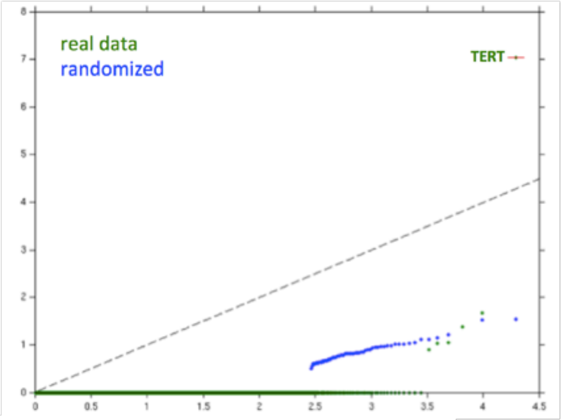
Bold = in Cancer Gene Census

Signals Pilot - Results - GBM-27 promoter

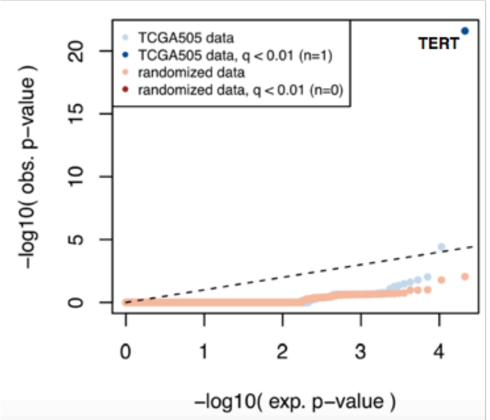
OncodriveFM



MutSig2CV

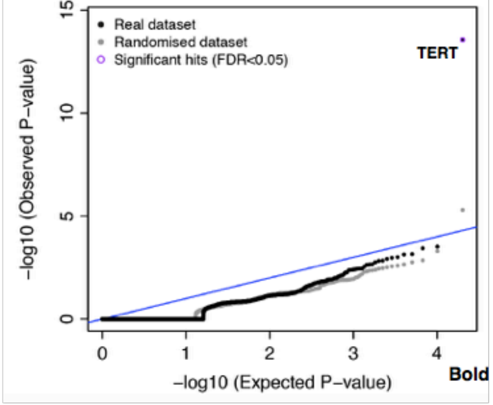


MSK-regions



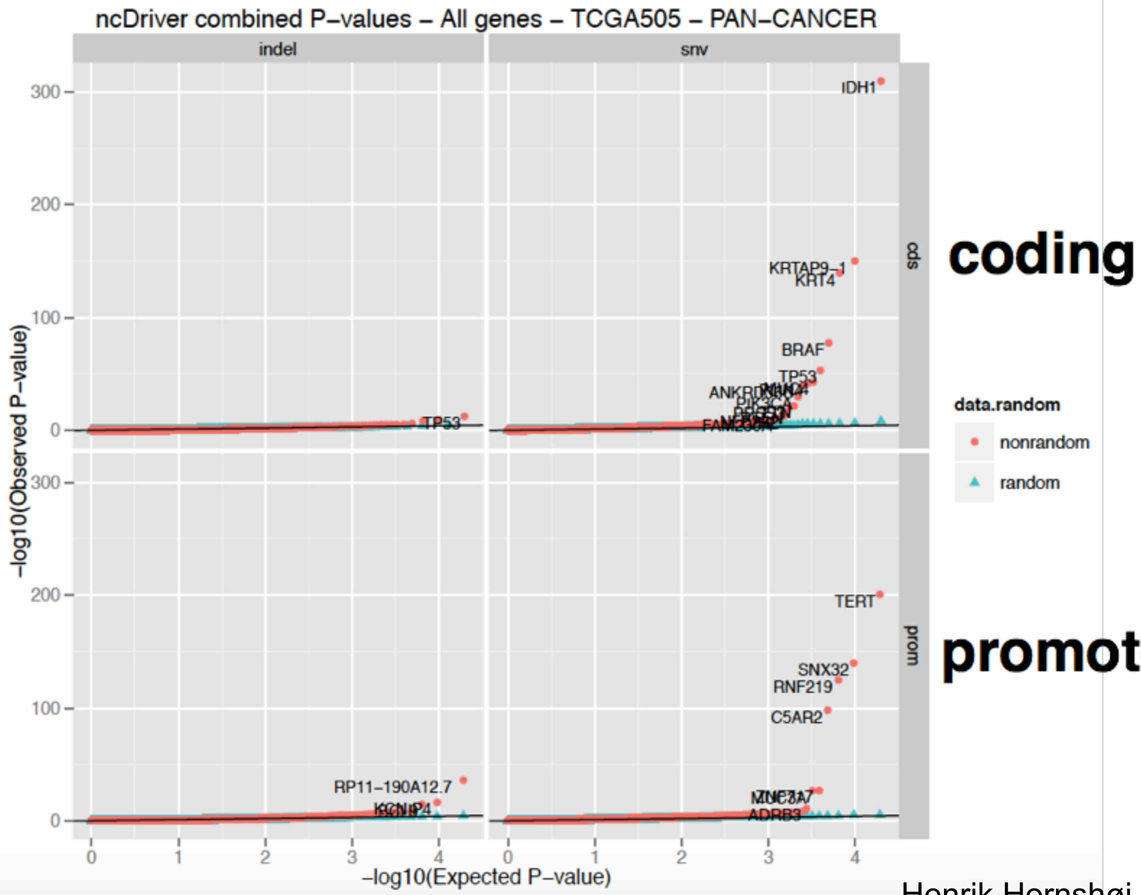
NBR-Sanger

NB - Inigo



Bold = in Cancer Gene Census

Signals Pilot - Results - TCGA-505 - ncDriver

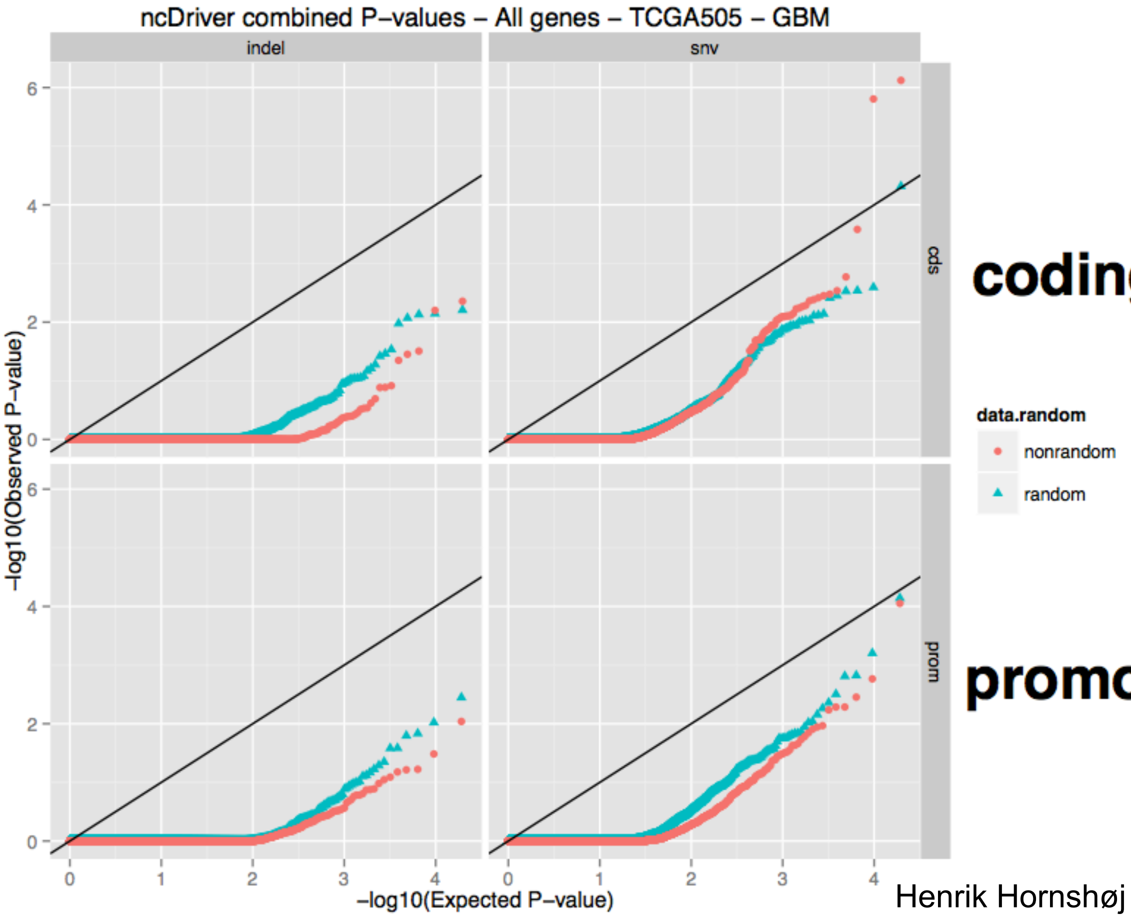


coding

promoter

Henrik Hornshøj and Jakob Skou Pedersen (Aarhus University)

Signals Pilot - Results - GBM-27 - ncDriver



coding

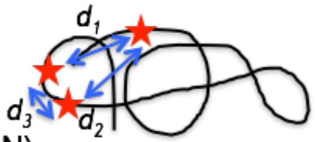
promoter

Henrik Hornshøj and Jakob Skou Pedersen (Aarhus University)

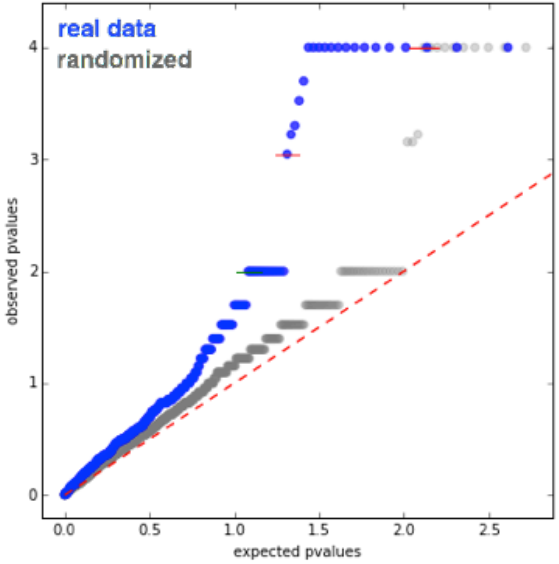
Signals Pilot - Results - TCGA-505 coding

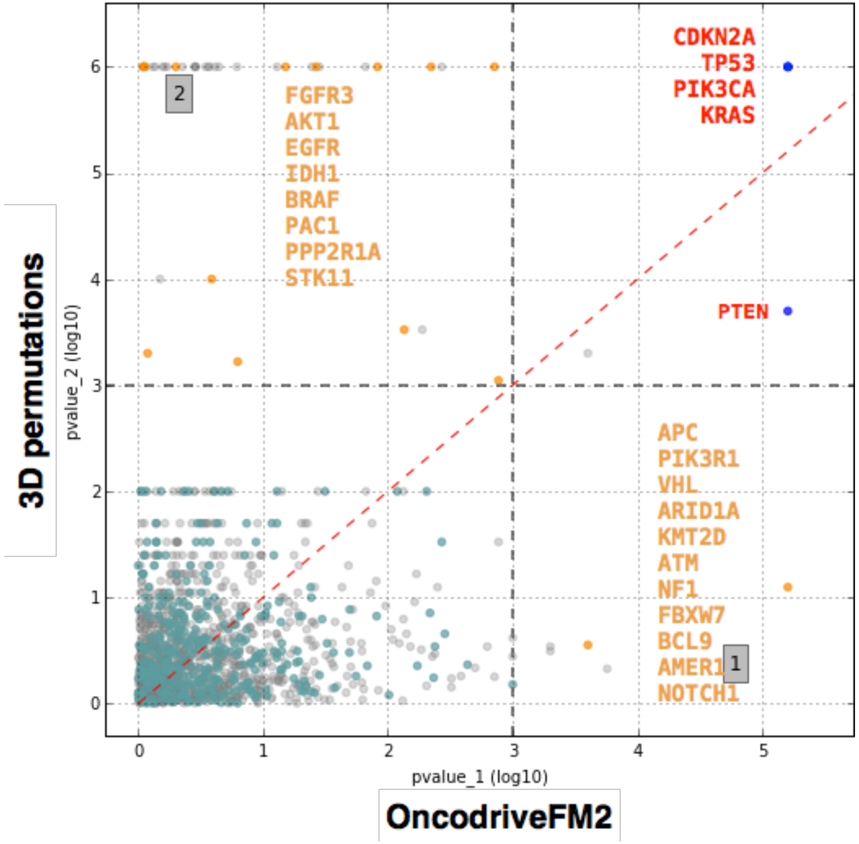
3D permutation

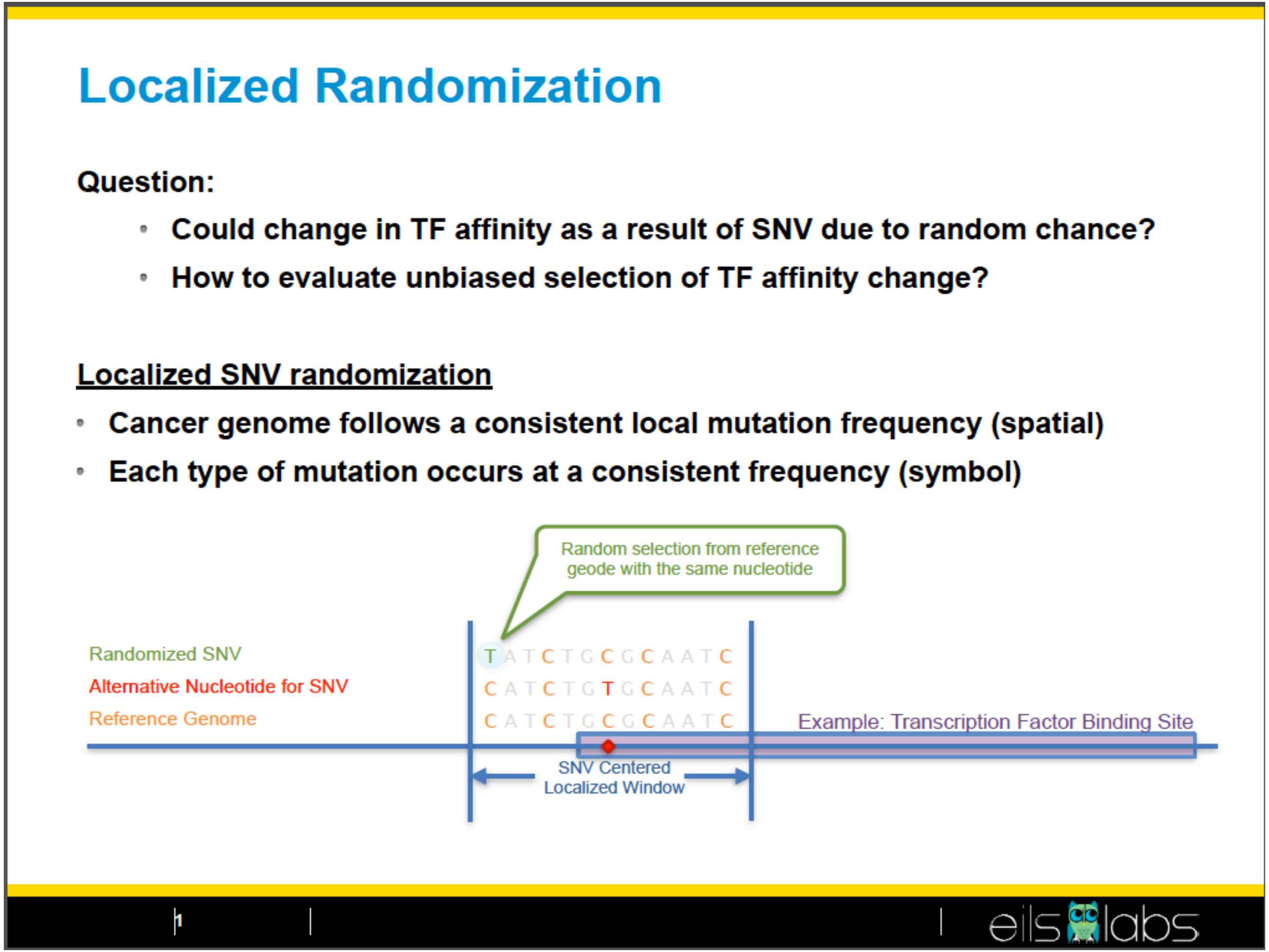
- Examine distribution of mutations in the 3D structure by a permutation test.
- Calculate average distances between mutations in the 3D structure. *P-values* are obtained by a permutation method.



Akihito Fujimoto (RIKEN)







Localized Randomization

Question:

- **Could change in TF affinity as a result of SNV due to random chance?**
- **How to evaluate unbiased selection of TF affinity change?**

<u>Localized SNV randomization</u>

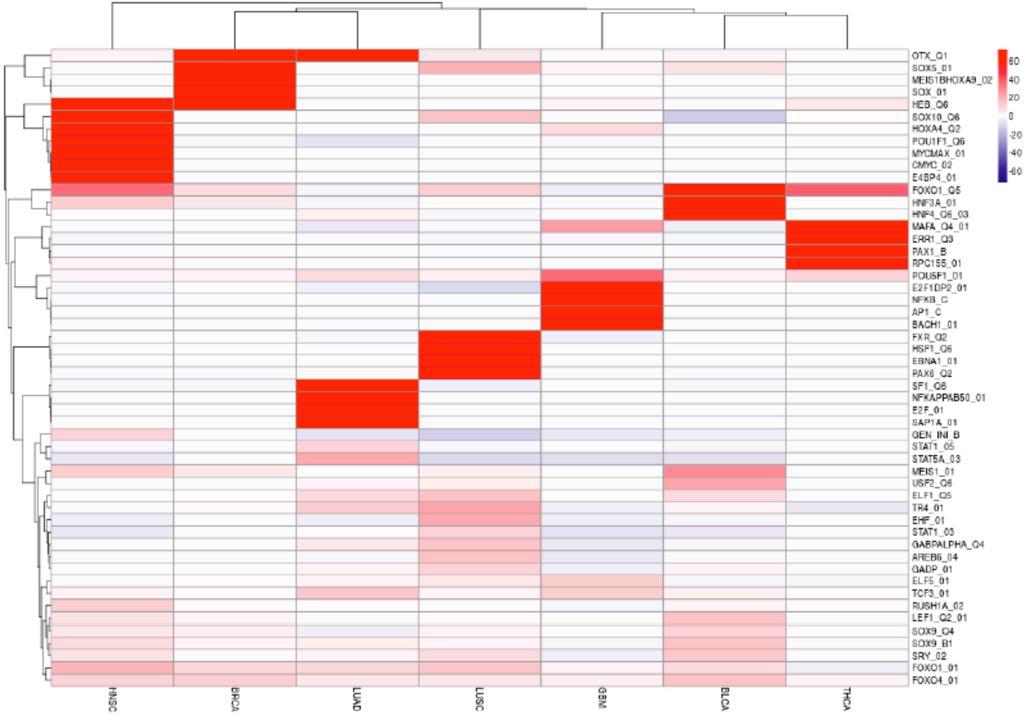
- **Cancer genome follows a consistent local mutation frequency (spatial)**
- **Each type of mutation occurs at a consistent frequency (symbol)**

Random selection from reference geode with the same nucleotide

Randomized SNV
T A T C T G C G C A A T C
Alternative Nucleotide for SNV
C A T C T G T G C A A T C
Reference Genome
C A T C T G C G C A A T C
Example: Transcription Factor Binding Site
SNV Centered
Localized Window

PanCan Pre-Train 1 Data SNV Affected TFBS

TFBS Creation: TF ϵ {z_{max}(cancer type) > 10}



Outline

1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
   a) Collecting common resources
   b) Pilot datasets
   c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
   a) Datasets pilot-50 (Train 1) Broad calls, public-607
   b) Compare submitted annotations
   c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
   a) TCGA-505, GBM-27 (Nuria)
   b) Simulated data (Gaddy, Inigo)
   c) Compare submitted significance analyses (Nuria)
4) Example of downstream analyses
   a) Analysis of non-coding RNA (Jakob) - annotation compilation, miRNA profiling (slides from Todd Johnson)
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

Reference Annotations sub-group

Annotation	**WG**	**Contact persons**	**Source**
Noncoding RNAs	WG-14	Jakob Skou Pederson	
Cis-regulatory regions	WG-2	Ekta Khurana, Esther Rheinbay, Manolis Kellis, Mark Gerstein, Gaddy Getz, Paz Polak	ENCODE & Epigenome Roadmap
PPI network	WG-5,9	Josh Stuart, Ben Raphael, Juri Reimand	STRING, iRef, HPRD, BioGRID
Fragile sites	WG-6	Nicola Roberts	
High-resolution CpG islands	WG-2	Lars Feuerbach	MPI-INF
Expression levels (generic tissue-agnostic values)	WG-6, 3	Nicola Roberts, Angela Brooks	Cancer cell line Encyclopedia, GTEx

And many more …

https://docs.google.com/document/d/1eNjR4vBFltujENA1pYdfFs-DWZYKxyNfOYZ5yToA3Kk/edit

Ekta Khurana

Pilot 1: Annotate variants and score individual variants

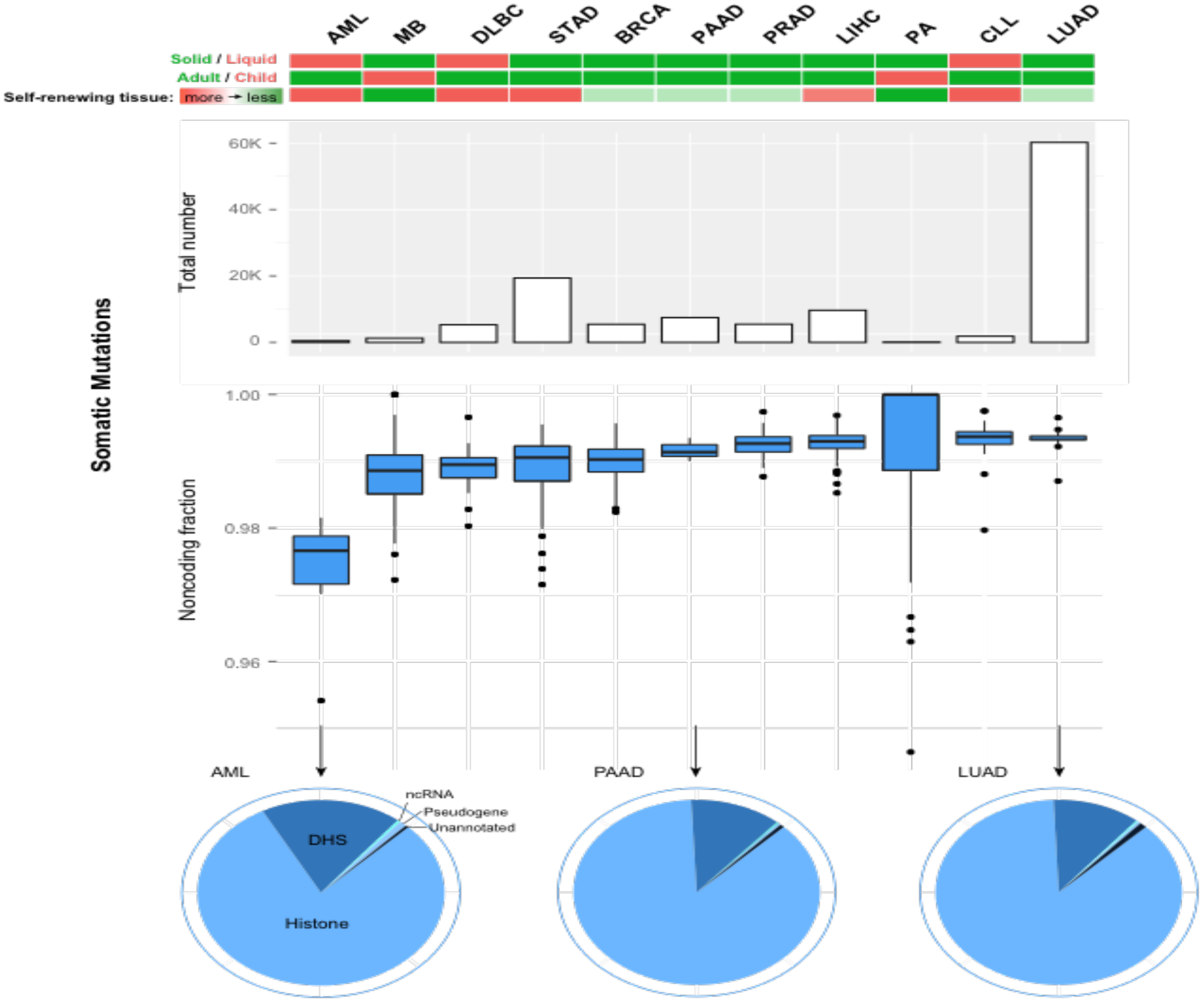
- Datasets used
  - Pilot-50 from Train 1 (Broad calls)
  - PCAWG-607 (Alexandrov et al, Nature, 2013 + STAD, http://bg.upf.edu/projects/pcawg/)

Pilot 1 results submitted

Method	Institute	Coding (GENCODE 19)	Noncoding	Scores/ Drivers
CanDrA & HotDriver	MDACC	Y	N	Y
Oncotator	Broad	Y	Y (intron, ncRNA, UTR)	N
FunSeq2	Yale/WCMC	Y	Y (intron, ncRNA, UTR, promoter, enhancer, DHS, motif)	Y
Johnson et al	RIKEN	Y	Y	Y
Feuerbach et al	DFKZ	Y	Y (intron, ncRNA, UTR)	N
Herrmann et al	DFKZ	N	Y (motif, DHS)	Y
wKinMut	DTU, Denmark	Y (kinases)	N	Y
Herrero et al	UCL	Y (ancestral allele)	Y (ancestral allele)	N

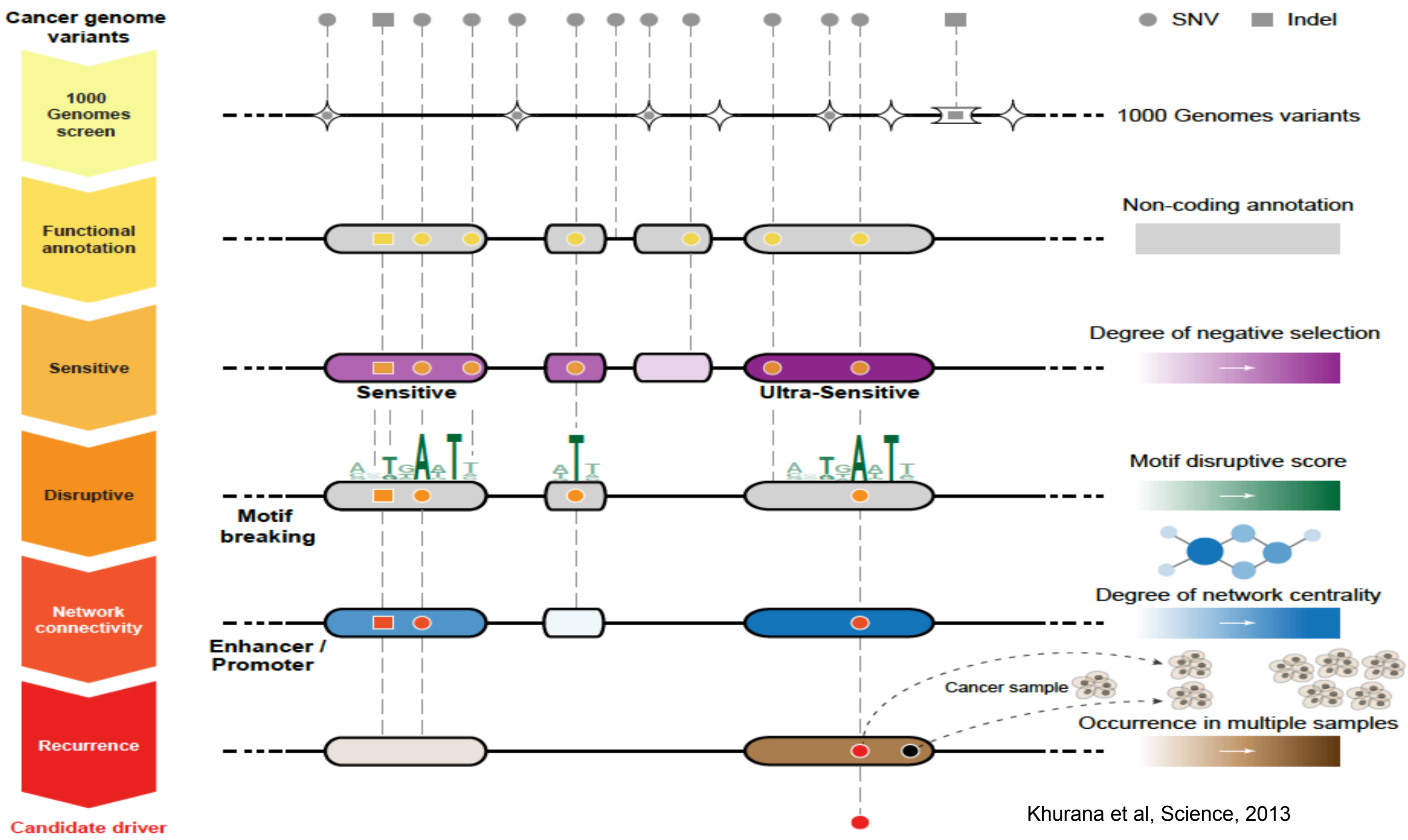
Ekta Khurana

Mutations in noncoding regulatory regions



Yao Fu
Ekta Khurana
Mark Gerstein

Identification of noncoding candidate drivers: FunSeq



Khurana et al, Science, 2013

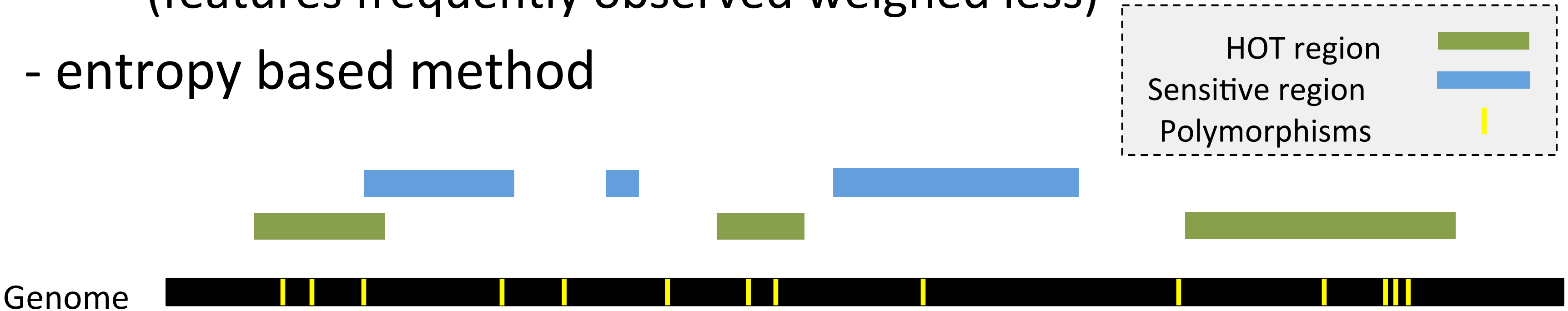
FunSeq2

- Feature weight

  - Weighted with mutation patterns in natural polymorphisms

    (features frequently observed weighed less)

  - entropy based method



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

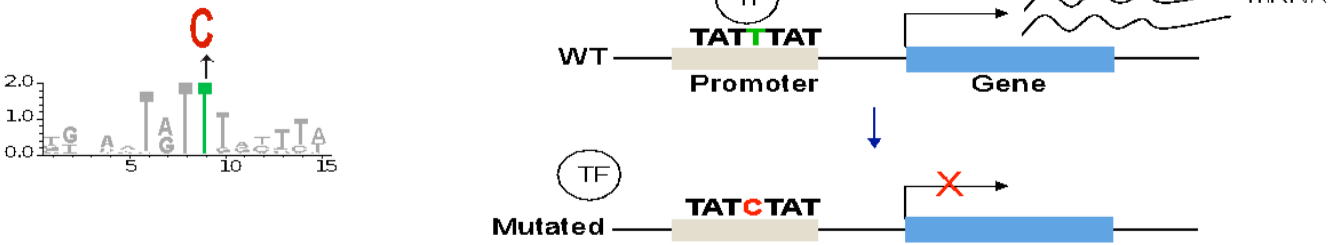
p ↑ w_d ↓ *p = probability of the feature overlapping natural polymorphisms*

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

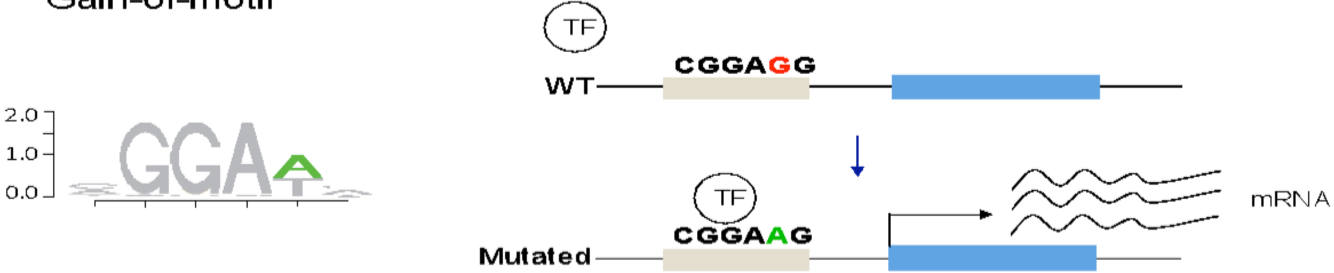
Fu et al, Genome Biology, 2014

Loss- and gain-of TF motif mutations

Loss-of-motif



Gain-of-motif



Ekta Khurana
Yao Fu
Mark Gerstein

Mutations with high FunSeq score

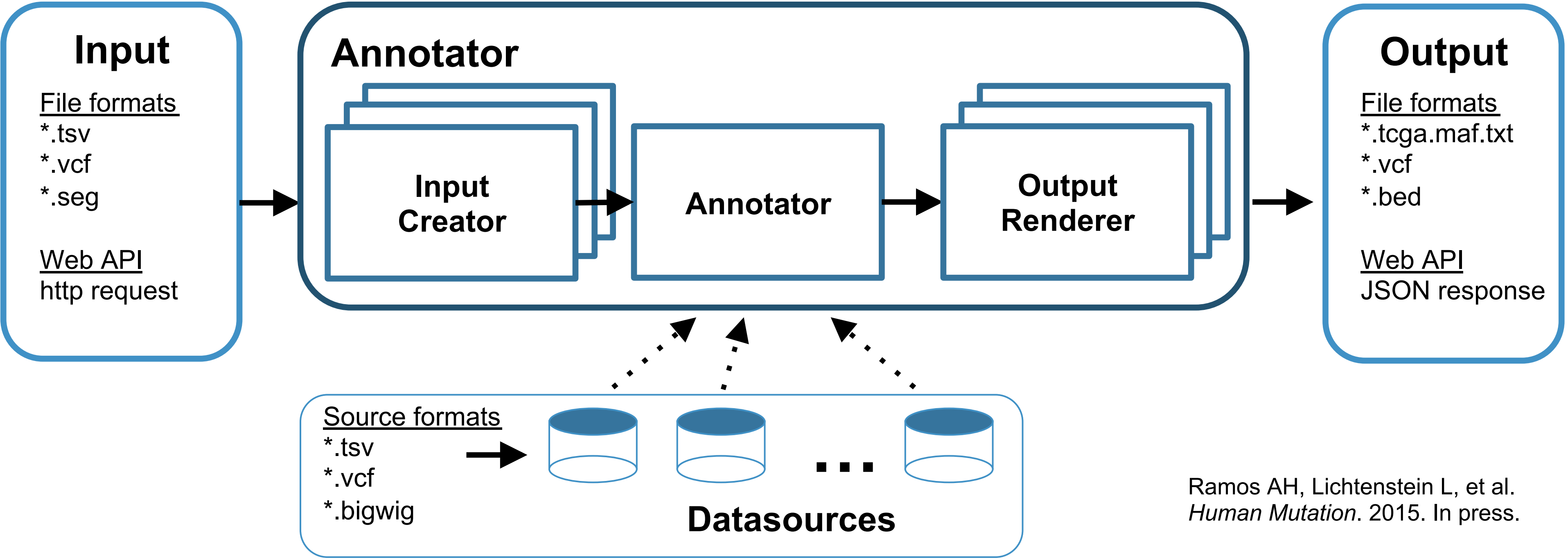
- Can be further prioritized, e.g. gene expression…

Type	Sample	Coding (all samples)	Noncoding (all samples)
ALL	1	17/87	3/7653
AML	7	13/87	10/3325
Breast	119	1430/6495	3128/638835
CLL	28	41/338	124/51406
Liver	88	1464/6257	3188/843489
Lung_Adeno	24	2255/9479	5291/1428263
Lymphoma_B	24	241/1212	529/126186
Medulloblastoma	100	282/1461	226/123387
Pancreas	15	129/965	179/111044
Pilocytic_A	101	13/103	15/10453
Stomach	100	5739/20374	10829/1891465
PCAWG_50	41	667/3745	1195/353489

Oncotator: variant annotation tool

Python tool for annotating variants with variant- and gene-centric data relevant to cancer researchers

Web app: broadinstitute.org/oncotator_beta/
Github: github.com/broadinstitute/oncotator



Input

File formats
*.tsv
*.vcf
*.seg

Web API
http request

Annotator

Output

File formats
*.tcga.maf.txt
*.vcf
*.bed

Web API
JSON response

Source formats
*.tsv
*.vcf
*.bigwig

Ramos AH, Lichtenstein L, et al.
Human Mutation. 2015. In press.

Alex Ramos, Getz Lab, MGH/Broad Institute

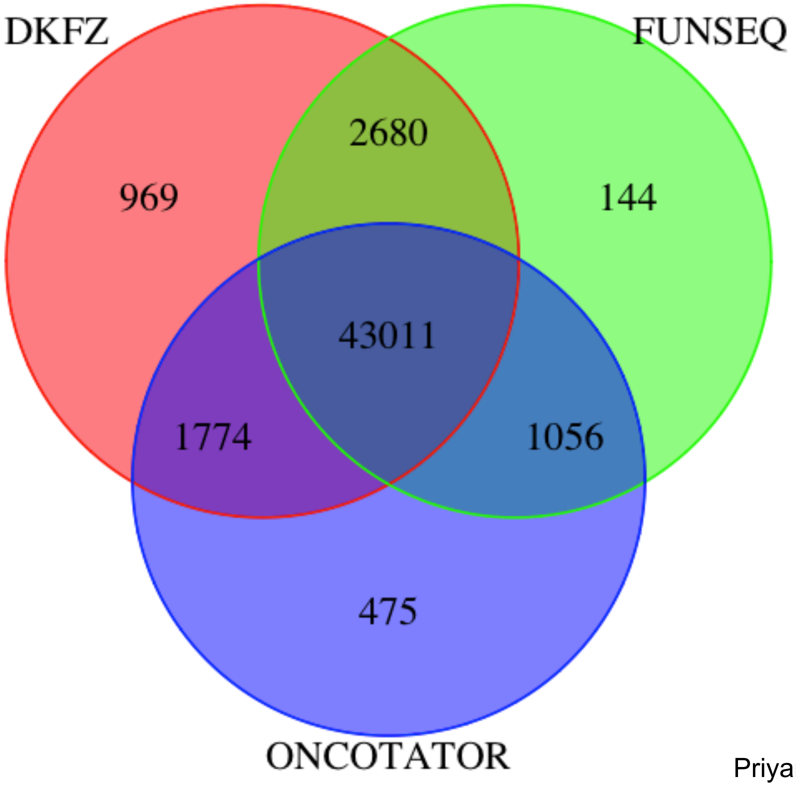
Oncotator default datasources

Annotation Category	Resource	URL	Comments
Genomic	GENCODE	http://www.gencodegenes.org/	GENCODE/ENSEMBL transcripts and annotations for hg19
	ref_context		Can be used for artifact inference
	gc_content		Can be used for artifact inference
	Human DNA Repair Genes	http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html	Alteration in such genes can help explain higher overall mutation rates in specific samples
Protein	UniProt	http://www.uniprot.org/	Includes Drugbank & GO annotations
	dbNSFP	https://sites.google.com/site/jpopgen/dbNSFP	Contains pre-computed conservation scores, prediction classifications, and other information
Cancer Variant	COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/	
	Cancer Gene Census	http://www.sanger.ac.uk/genetics/CGP/Census/	
	CCLE	http://www.broadinstitute.org/ccle/home	Cancer cell line annotations. Can be used to identify cell line models containg variants of interest
	Familial Cancer Database	http://www.familialcancerdatabase.nl/	
	ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/	
Non-Cancer Variant	dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/	b142 release for human (9606)
	1000 Genomes	http://www.1000genomes.org/data	Phase 3 variant set
	NHLBI GO Exome Sequencing Project (ESP)	https://esp.gs.washington.edu/drupal/	

*CLI tool includes framework for adding additional datasources.

Alex Ramos, Getz Lab, MGH/Broad Institute

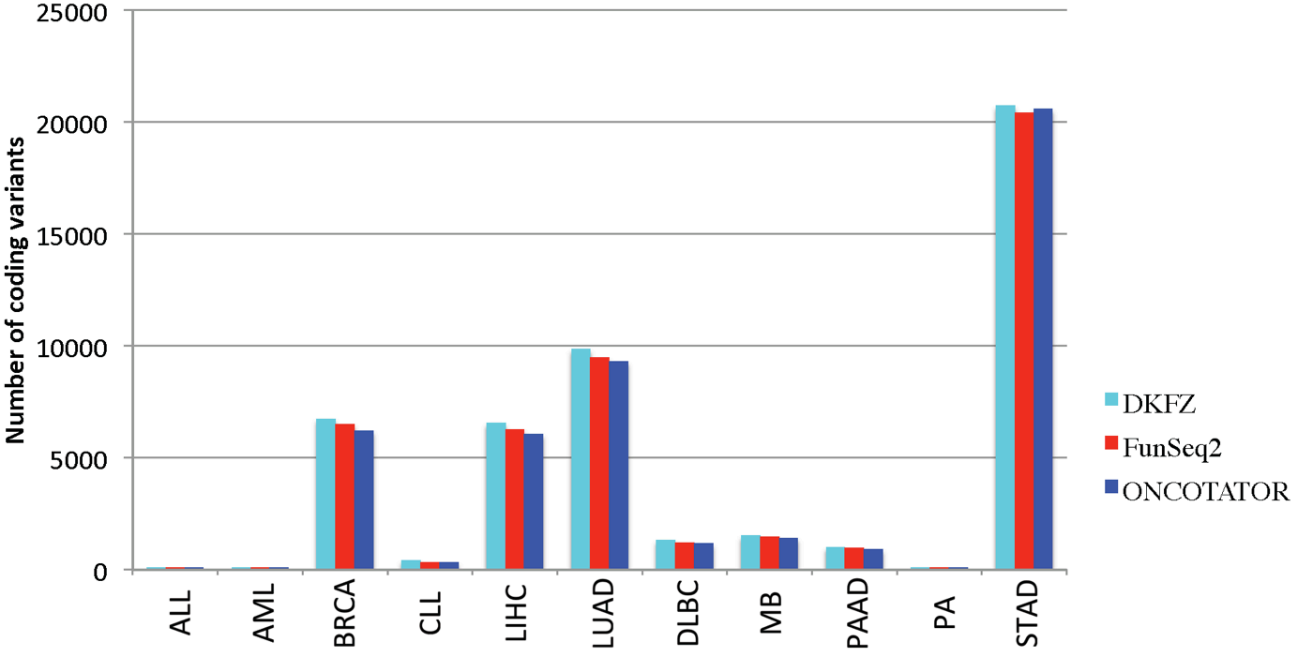
Predicted coding variants for all cancer types using 3 methods



DKFZ	48434
ONCOTATOR	46316
FUNSEQ2	46891

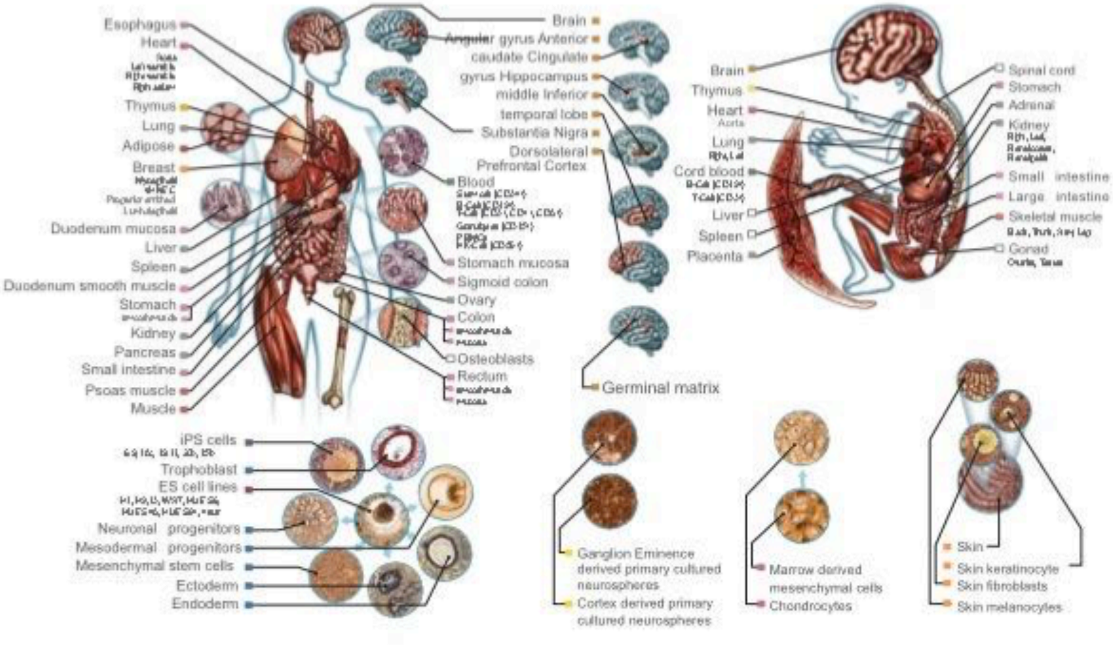
Priyanka Dhingra, Ekta Khurana (WCMC)

Pilot 1 comparison of different methods



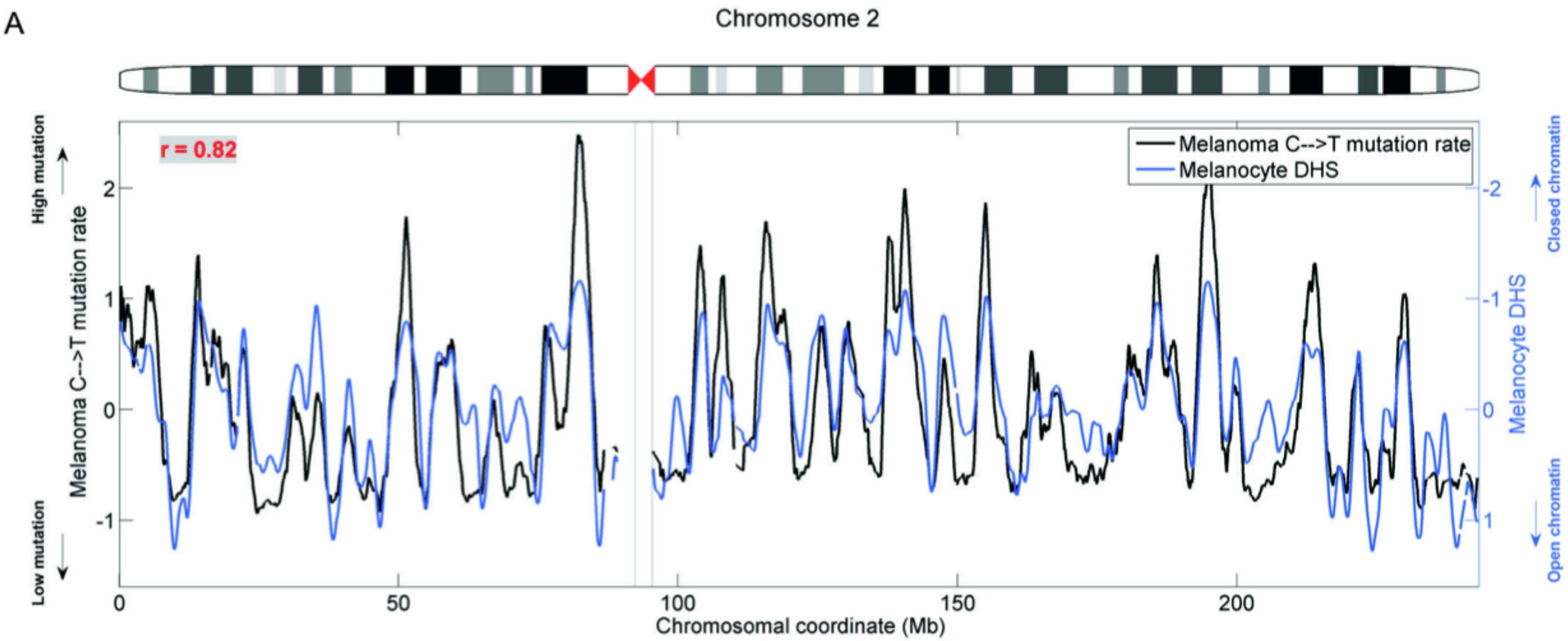
Priyanka Dhingra, Ekta Khurana
(WCMC)

Epigenomic Roadmap Project provides an atlas of epigenomes from 127 adult and fetal tissues



Paz Polak, Getz Lab, MGH/Broad Institute

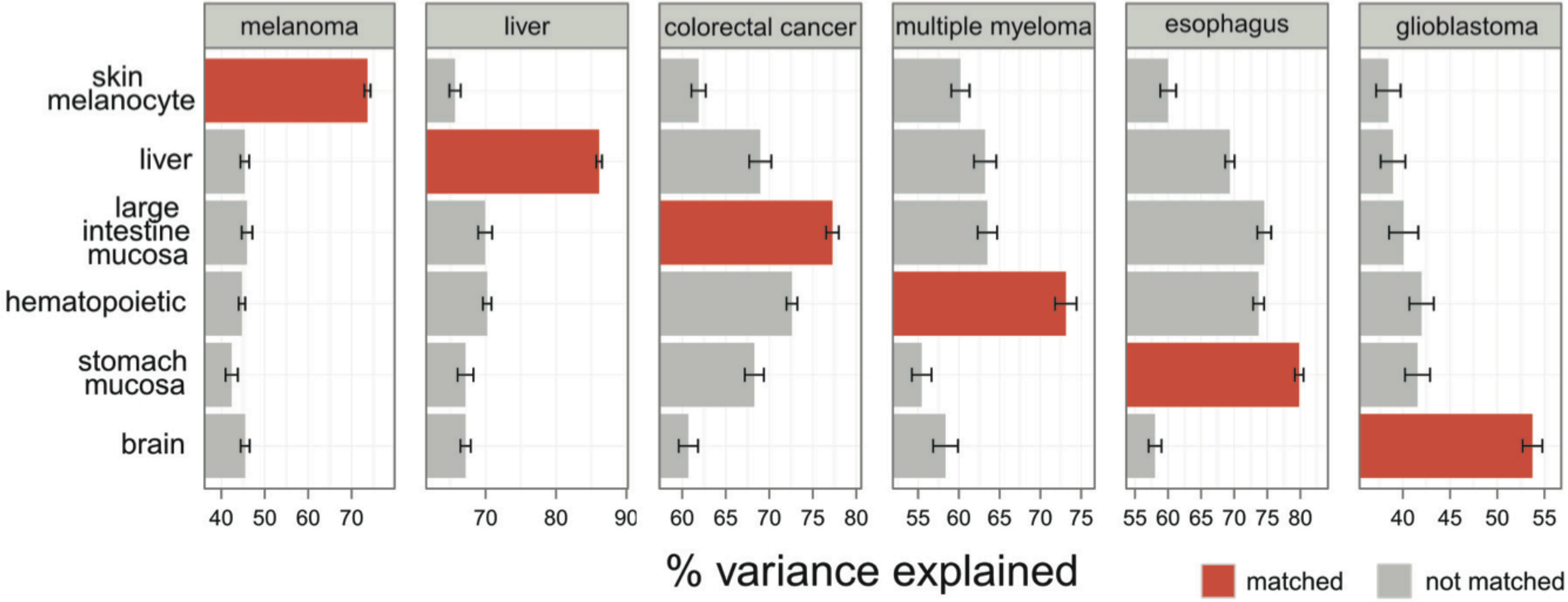
DNaseI profile in normal melanocytes is negatively correlated with melanoma mutation density profile



Melanoma WGS data: Berger et al, 2012
Polak*, Karlic* et al, *Nature*, in press

Paz Polak, Getz Lab, MGH/Broad Institute

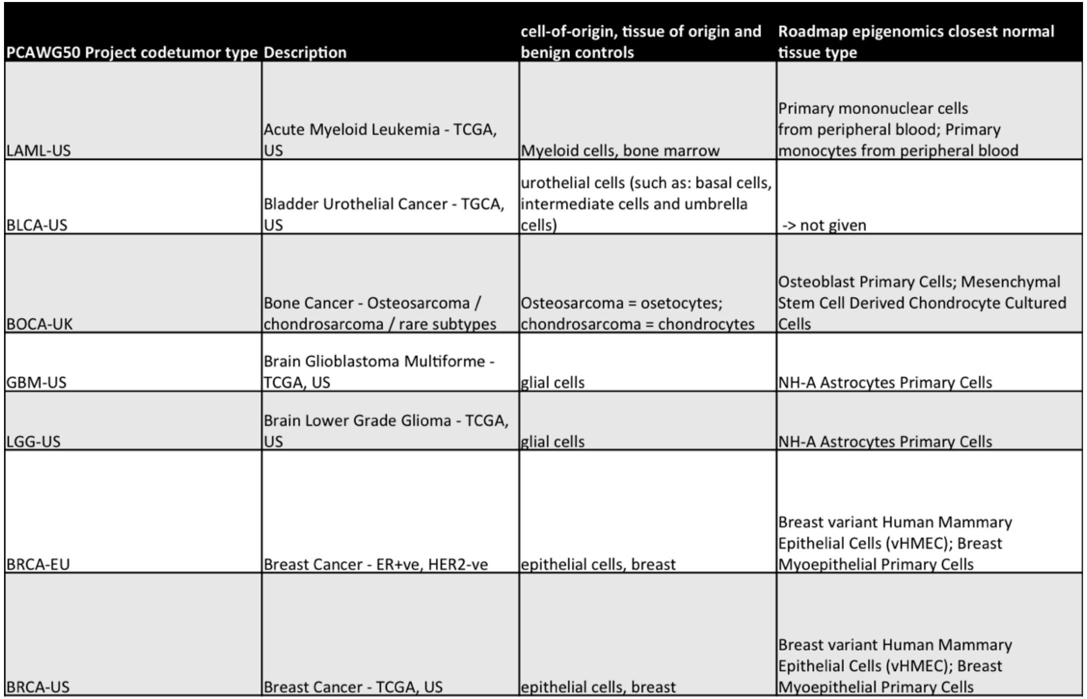
Epigenomes with the highest predictive accuracy correspond to the closest cell-of-origin



Polak*, Karlic* et al, Nature, publication date 2/19/2015

Paz Polak, Getz Lab, MGH/Broad Institute

Annotating and analyzing variants using cell-of-origin epigenomic data



PCAWG50 Project codetumor type	Description	cell-of-origin, tissue of origin and benign controls	Roadmap epigenomics closest normal tissue type
LAML-US	Acute Myeloid Leukemia - TCGA, US	Myeloid cells, bone marrow	Primary mononuclear cells from peripheral blood; Primary monocytes from peripheral blood
BLCA-US	Bladder Urothelial Cancer - TGCA, US	urothelial cells (such as: basal cells, intermediate cells and umbrella cells)	-> not given
BOCA-UK	Bone Cancer - Osteosarcoma / chondrosarcoma / rare subtypes	Osteosarcoma = osetocytes; chondrosarcoma = chondrocytes	Osteoblast Primary Cells; Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
GBM-US	Brain Glioblastoma Multiforme - TCGA, US	glial cells	NH-A Astrocytes Primary Cells
LGG-US	Brain Lower Grade Glioma - TCGA, US	glial cells	NH-A Astrocytes Primary Cells
BRCA-EU	Breast Cancer - ER+ve, HER2-ve	epithelial cells, breast	Breast variant Human Mammary Epithelial Cells (vHMEC); Breast Myoepithelial Primary Cells
BRCA-US	Breast Cancer - TCGA, US	epithelial cells, breast	Breast variant Human Mammary Epithelial Cells (vHMEC); Breast Myoepithelial Primary Cells

Full spreadsheet at http://goo.gl/vp3uLS

Paz Polak, Getz Lab, MGH/Broad Institute

Outline

1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
    a) Collecting common resources
    b) Pilot datasets
    c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
    a) Datasets pilot-50 (Train 1) Broad calls, public-607
    b) Compare submitted annotations
    c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
    a) TCGA-505, GBM-27 (Nuria)
    b) Simulated data (Gaddy, Inigo)
    c) Compare submitted significance analyses (Nuria)
4) Example of downstream analyses
    a) Analysis of non-coding RNA (Jakob) - annotation compilation, miRNA profiling (slides from Todd Johnson)
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

Outline

1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
    a) Collecting common resources
    b) Pilot datasets
    c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
    a) Datasets pilot-50 (Train 1) Broad calls, public-607
    b) Compare submitted annotations
    c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
    a) TCGA-505, GBM-27 (Nuria)
    b) Simulated data (Gaddy, Inigo)
    c) Compare submitted significance analyses (Nuria)
4) Analysis of non-coding RNA (Jakob): annotation compilation, miRNA profiling, etc.
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

WG-14: mutational analysis of ncRNAs

Mutational analysis

Identify ncRNA drivers (miRNAs, lncRNAs, tRNAs, ...)

- mutation enrichment
- functional impact of mutations
- clustering, etc

Mutations in regulatory regions

- promoter regions
- splice sites
- cleavage sites

Expression & epigenetic analysis

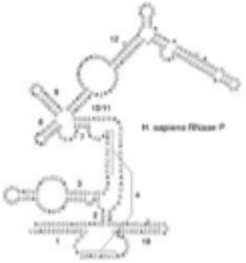
Mutational effect on expression (miRNAs, lncRNAs, ...)

ncRNA perturbation effect on:

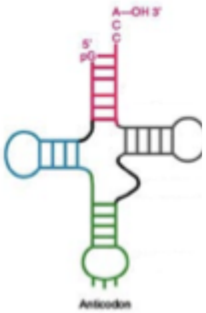
- mRNA expression (miRNAs & lncRNAs)
- methylation patterns (lncRNAs)

miRNA

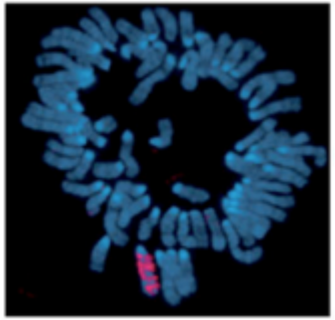
RNase P



tRNA



Xist (~20 kb long)



Group leaders: David Wheeler & Jakob Skou Pedersen

ncRNA annotation sources

Sources with expression evidence

GENCODE, Basic set (v.19):

- mixed ncRNAs (n=39,301)

miRBase (v.20):

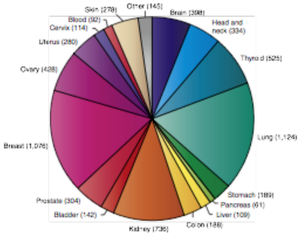
- mature miRNAs (n=2,794)
- miRNA stem-loops (n=1,871)

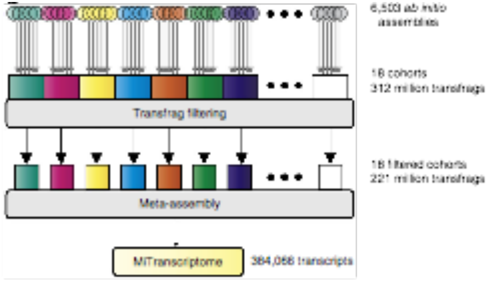
snoRNABase (v.3):

- snoRNAs (n=402)

MiTranscriptome:

- lncRNAs / mixed (n=124,928)





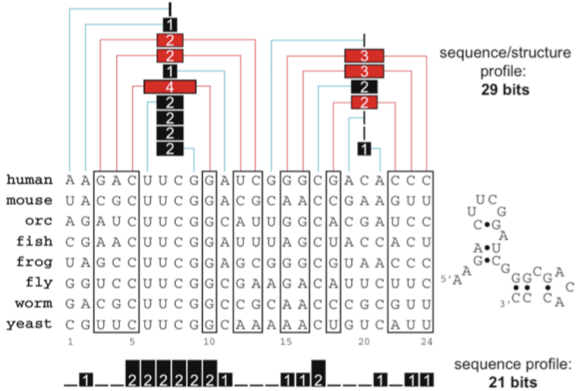
Iyver et al. The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics (2015).

Sources with homology matches

rfam: structural RNA families (v.11, n=8,825)

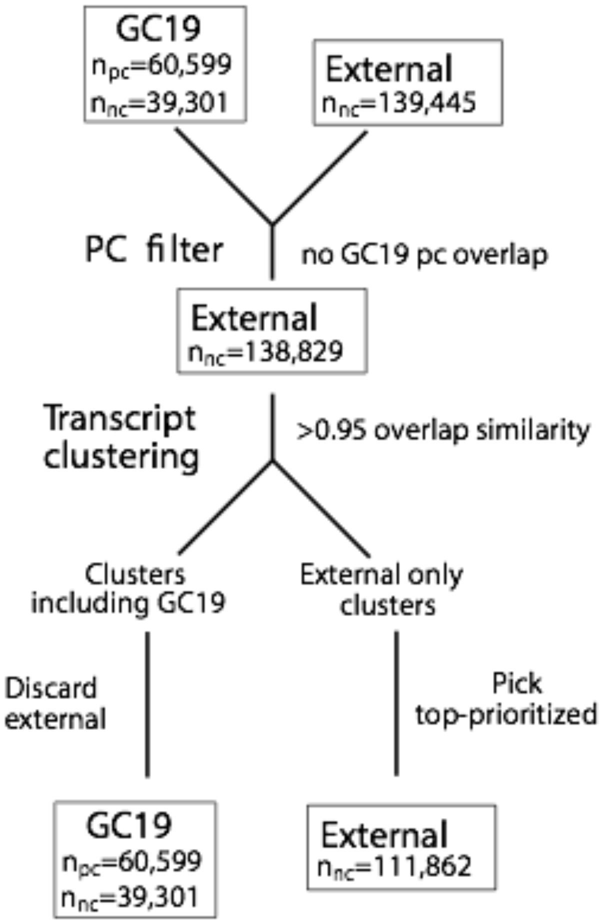
Genomic tRNA Database (hg19, n=625)

RNA structure homology searches



```
human  A A G A C U U C G G A U C G G G C G A C A C C C
mouse  U A C G C U U C G G A C G C A A C C G A A G U U
orc    A G A U C U U C G G C A U U G G C A C G A U C C
fish   C G A A C U U C G G A U U U A G C U A C C A C U
frog   U A G C C U U C G G A G C G G G C G U A A C C C
fly    G G U C C U U C G G C G A A G A C A U U C U U C
worm   G A C G C U U C G G C C G C A A C C C G C G U U
yeast  C G U U C U U C G G C A A A A A C U G U C A U U
       1       5         10        15        20      24
```

Reducing to single comprehensive, non-redundant ncRNA set



GC19 n_{pc}=60,599 n_{nc}=39,301

External n_{nc}=139,445

PC filter

no GC19 pc overlap

External n_{nc}=138,829

Transcript clustering

>0.95 overlap similarity

Clusters including GC19

External only clusters

Discard external

Pick top-prioritized

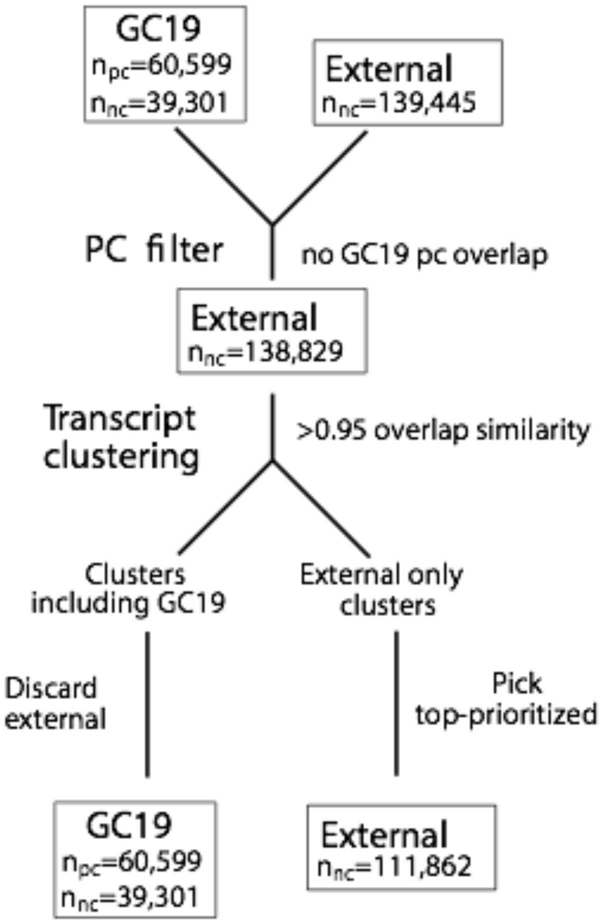
GC19 n_{pc}=60,599 n_{nc}=39,301

External n_{nc}=111,862

1. miRbase	(n=4,665)
2. snoRNAbase	(n=402)
3. MiTranscriptome	(n=124,928)
4. tRNA DB	(n=625)
5. Rfam	(n=8,825)

Morten Muhlig Nielsen (AU)

Reducing to single comprehensive, non-redundant ncRNA set



GC19
n_{pc}=60,599
n_{nc}=39,301

External
n_{nc}=139,445

PC filter

no GC19 pc overlap

External
n_{nc}=138,829

Transcript clustering

>0.95 overlap similarity

Clusters including GC19

External only clusters

Discard external

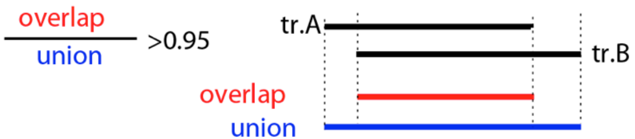
Pick top-prioritized

GC19
n_{pc}=60,599
n_{nc}=39,301

External
n_{nc}=111,862

1. miRbase	(n=4,665)
2. snoRNAbase	(n=402)
3. MiTranscriptome	(n=124,928)
4. tRNA DB	(n=625)
5. Rfam	(n=8,825)

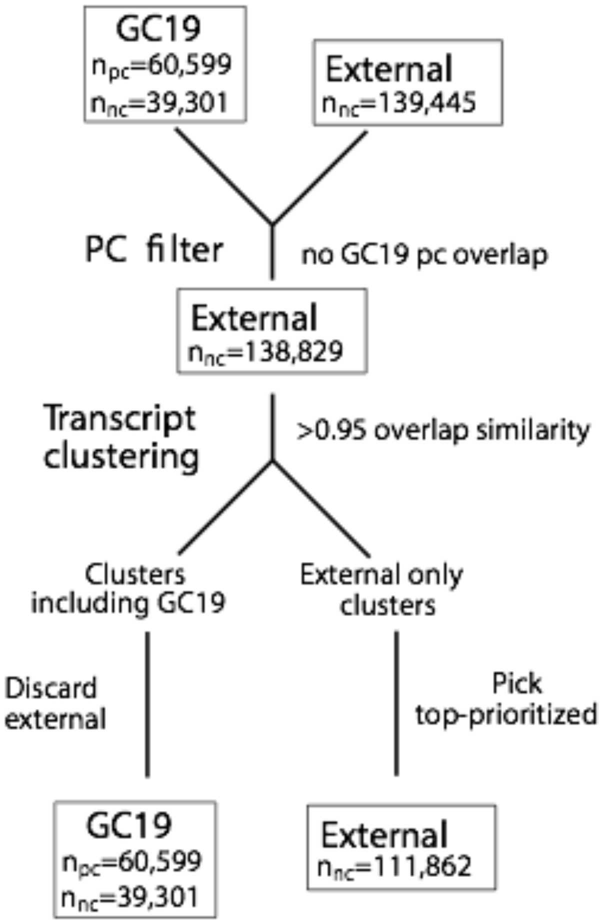
overlap similarity



$$\frac{\text{overlap}}{\text{union}} > 0.95$$

Morten Muhlig Nielsen (AU)

Reducing to single comprehensive, non-redundant ncRNA set



GC19
n_{pc}=60,599
n_{nc}=39,301

External
n_{nc}=139,445

PC filter

no GC19 pc overlap

External
n_{nc}=138,829

Transcript clustering

>0.95 overlap similarity

Clusters including GC19

External only clusters

Discard external

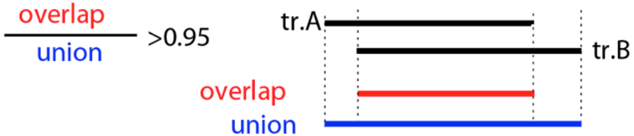
Pick top-prioritized

GC19
n_{pc}=60,599
n_{nc}=39,301

External
n_{nc}=111,862

1. miRbase	(n=4,665)
2. snoRNAbase	(n=402)
3. MiTranscriptome	(n=124,928)
4. tRNA DB	(n=625)
5. Rfam	(n=8,825)

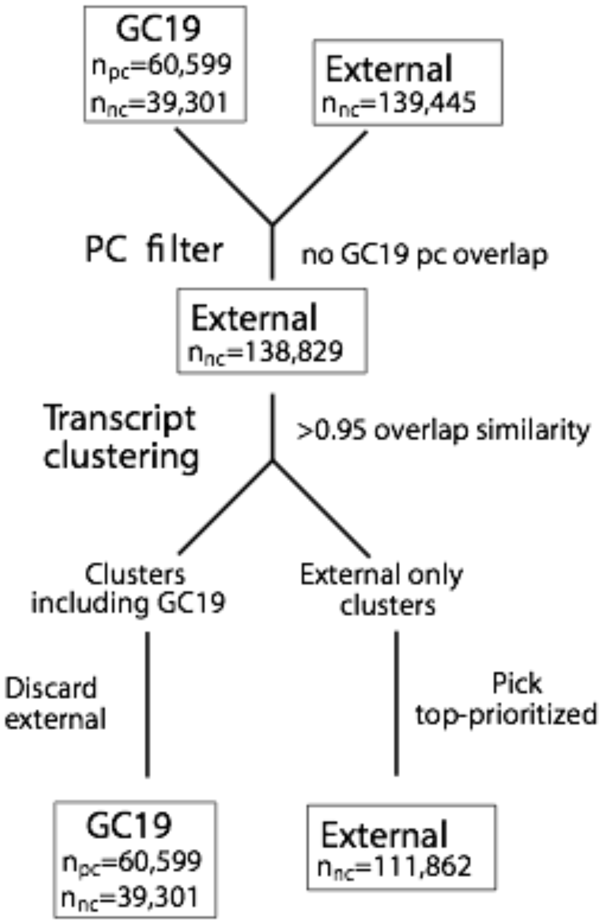
overlap similarity



1. miRbase	(n=3,082)
2. snoRNAbase	(n=16)
3. MiTranscriptome	(n=105,670)
4. tRNA DB	(n=620)
5. Rfam	(n=2,474)

Morten Muhlig Nielsen (AU)

Reducing to single comprehensive, non-redundant ncRNA set



GC19 n_{pc}=60,599 n_{nc}=39,301

External n_{nc}=139,445

PC filter

no GC19 pc overlap

External n_{nc}=138,829

Transcript clustering

>0.95 overlap similarity

Clusters including GC19

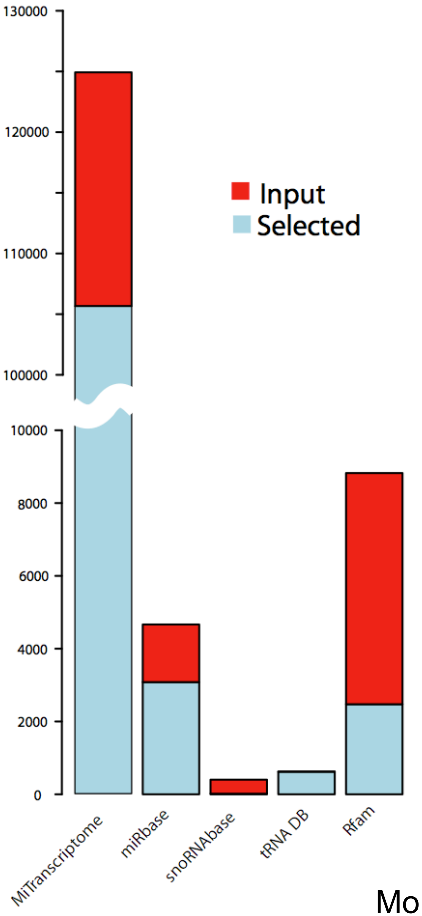
External only clusters

Discard external

Pick top-prioritized

GC19 n_{pc}=60,599 n_{nc}=39,301

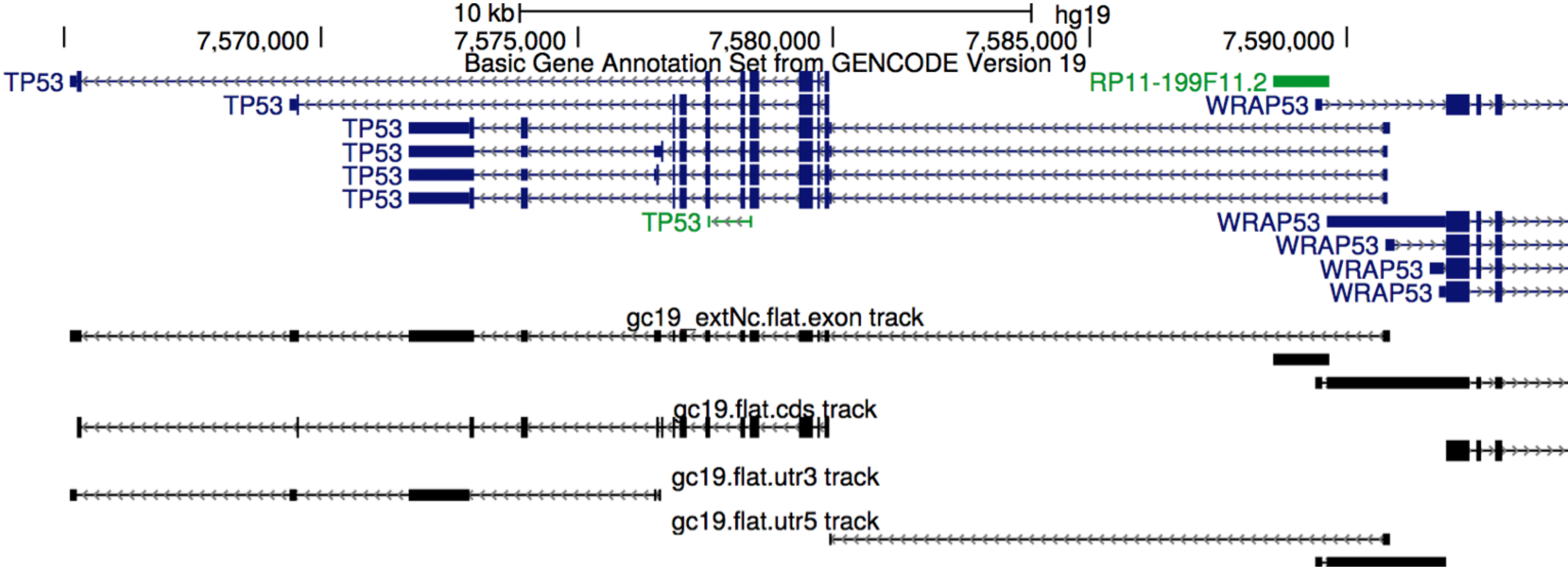
External n_{nc}=111,862



Morten Muhlig Nielsen (AU)

Collapsing isoforms to flattened gene models

For mutational analysis, we decided to work with a single gene model per transcript.



Available from Synapse: https://www.synapse.org/#!Synapse:syn3154572

Morten Muhlig Nielsen (AU)

Source, IDs, and addtional information retained in bed files

geneSet::geneName::geneID::transcriptID::extraAnnotation::extraAnnotation...

Ex.

gencode::SAMD11::ENSG00000187634.6::ENST00000342066.3::protein_coding::KNOWN

Available from Synapse: https://www.synapse.org/#!Synapse:syn3154572

Morten Muhlig Nielsen (AU)

ncRNA / lncRNA expression profiling

Approach:

- Base on RNAseq SOP from WG3
- Profile extended Gencode set

Challenges:

- Families of ncRNA with highly similar members
- Idea: define equivalence classes of highly similar transcripts and combine read counts for comparison between samples.

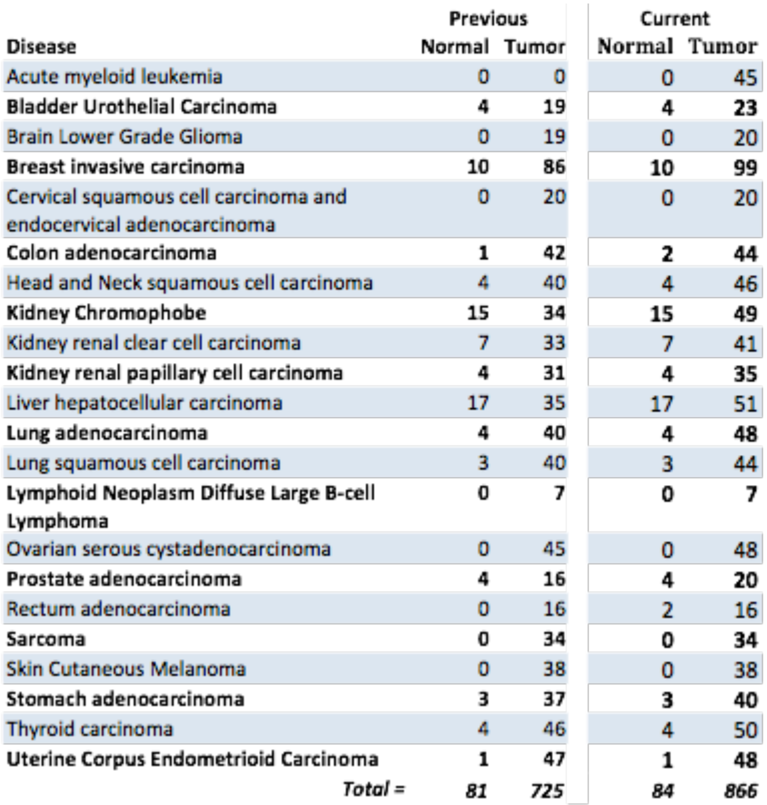
Amin Samir, Lab of Lynda Chin, MD Anderson Cancer Center

miRNA expression profiling

miRNA-seq SOP

- SOP from TCGA (Genome Sciences Centre, BC Cancer Agency)
- Updated with same version of BWA as used for WGS mapping

Overview of samples / patients



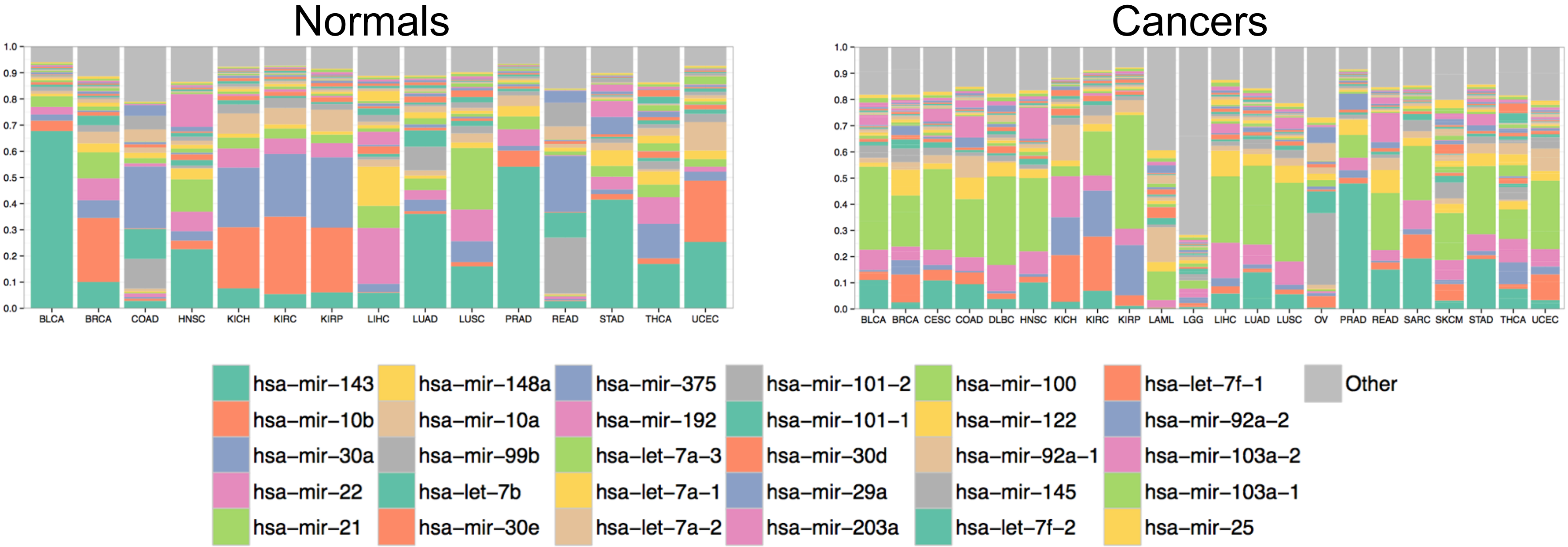
Disease	Previous Normal	Previous Tumor	Current Normal	Current Tumor
Acute myeloid leukemia	0	0	0	45
Bladder Urothelial Carcinoma	4	19	4	23
Brain Lower Grade Glioma	0	19	0	20
Breast invasive carcinoma	10	86	10	99
Cervical squamous cell carcinoma and endocervical adenocarcinoma	0	20	0	20
Colon adenocarcinoma	1	42	2	44
Head and Neck squamous cell carcinoma	4	40	4	46
Kidney Chromophobe	15	34	15	49
Kidney renal clear cell carcinoma	7	33	7	41
Kidney renal papillary cell carcinoma	4	31	4	35
Liver hepatocellular carcinoma	17	35	17	51
Lung adenocarcinoma	4	40	4	48
Lung squamous cell carcinoma	3	40	3	44
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	0	7	0	7
Ovarian serous cystadenocarcinoma	0	45	0	48
Prostate adenocarcinoma	4	16	4	20
Rectum adenocarcinoma	0	16	2	16
Sarcoma	0	34	0	34
Skin Cutaneous Melanoma	0	38	0	38
Stomach adenocarcinoma	3	37	3	40
Thyroid carcinoma	4	46	4	50
Uterine Corpus Endometrioid Carcinoma	1	47	1	48
Total =	*81*	*725*	*84*	*866*

All from TCGA. Awaiting 44 from ICGC.

Data collection: DCC & Sergei Iakhnin heads collection of metadata
Expression profiling: Todd Johnson et al. Riken.

miRNA expression profiling - first results

Relative expression of miRNAs across tumour types

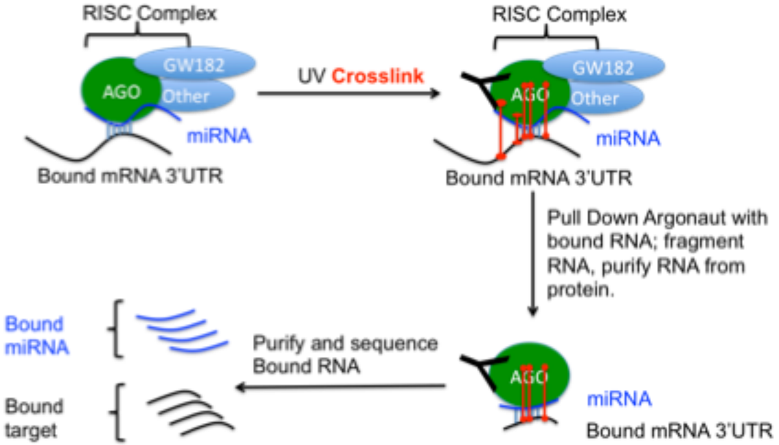


Todd Johnson et al. Riken.

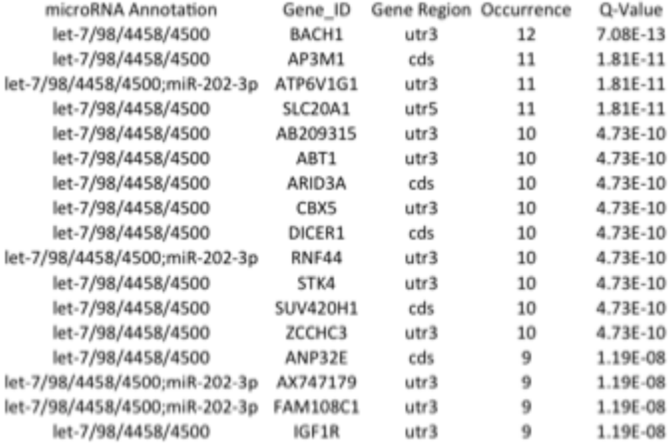
Experimentally defined miRNA binding sites across cell lines / cancer types

miRNA-AGO-CLIP Target Atlas:
Experimental screens of tumour cell lines (n>20), xenografts, etc.

AGO crosslinking and CLIP



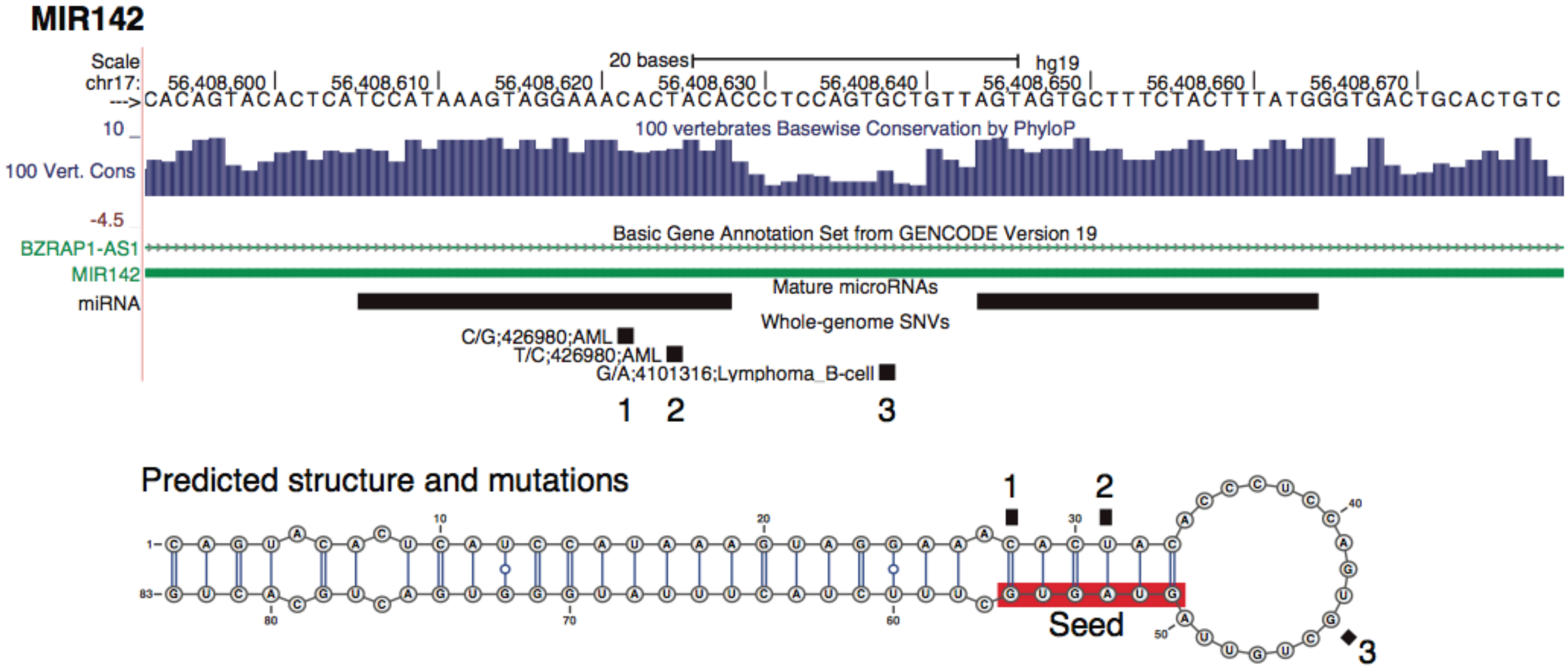
miRNA-mRNA network data



microRNA Annotation	Gene_ID	Gene Region	Occurrence	Q-Value
let-7/98/4458/4500	BACH1	utr3	12	7.08E-13
let-7/98/4458/4500	AP3M1	cds	11	1.81E-11
let-7/98/4458/4500;miR-202-3p	ATP6V1G1	utr3	11	1.81E-11
let-7/98/4458/4500	SLC20A1	utr5	11	1.81E-11
let-7/98/4458/4500	AB209315	utr3	10	4.73E-10
let-7/98/4458/4500	ABT1	utr3	10	4.73E-10
let-7/98/4458/4500	ARID3A	cds	10	4.73E-10
let-7/98/4458/4500	CBX5	utr3	10	4.73E-10
let-7/98/4458/4500	DICER1	cds	10	4.73E-10
let-7/98/4458/4500;miR-202-3p	RNF44	utr3	10	4.73E-10
let-7/98/4458/4500	STK4	utr3	10	4.73E-10
let-7/98/4458/4500	SUV420H1	cds	10	4.73E-10
let-7/98/4458/4500	ZCCHC3	utr3	10	4.73E-10
let-7/98/4458/4500	ANP32E	cds	9	1.19E-08
let-7/98/4458/4500;miR-202-3p	AX747179	utr3	9	1.19E-08
let-7/98/4458/4500;miR-202-3p	FAM108C1	utr3	9	1.19E-08
let-7/98/4458/4500	IGF1R	utr3	9	1.19E-08

Mark Hamilton, Lab of Sean McGuire, Baylor College of Medicine.
(Target Atlas: Hamilton et. al., Nature Communications, 2013)

Example: mutated miRNA


MIR142
Scale
20 bases
hg19
chr17:
56,408,600
56,408,610
56,408,620
56,408,630
56,408,640
56,408,650
56,408,660
56,408,670
---> CACAGTACACTCATCCATAAAGTAGGAAACACTACACCCTCCAGTGCTGTTAGTAGTGCTTTCTACTTTATGGGTGACTGCACTGTC
10
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
-4.5
Basic Gene Annotation Set from GENCODE Version 19
BZRAP1-AS1
MIR142
Mature microRNAs
miRNA
Whole-genome SNVs
C/G;426980;AML
T/C;426980;AML
G/A;4101316;Lymphoma_B-cell
1
2
3
Predicted structure and mutations
1
2
1
10
20
30
40
50
60
70
80
83
Seed
3

Ref: Henrik Hornshøj et al., ncDriver. In preparation.

Alexandrov et al. data (n=507).

Driver mutations in miR142 previously reported based on TCGA exome data:

Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. TCGA. The New England Journal of Medicine.. 2013.

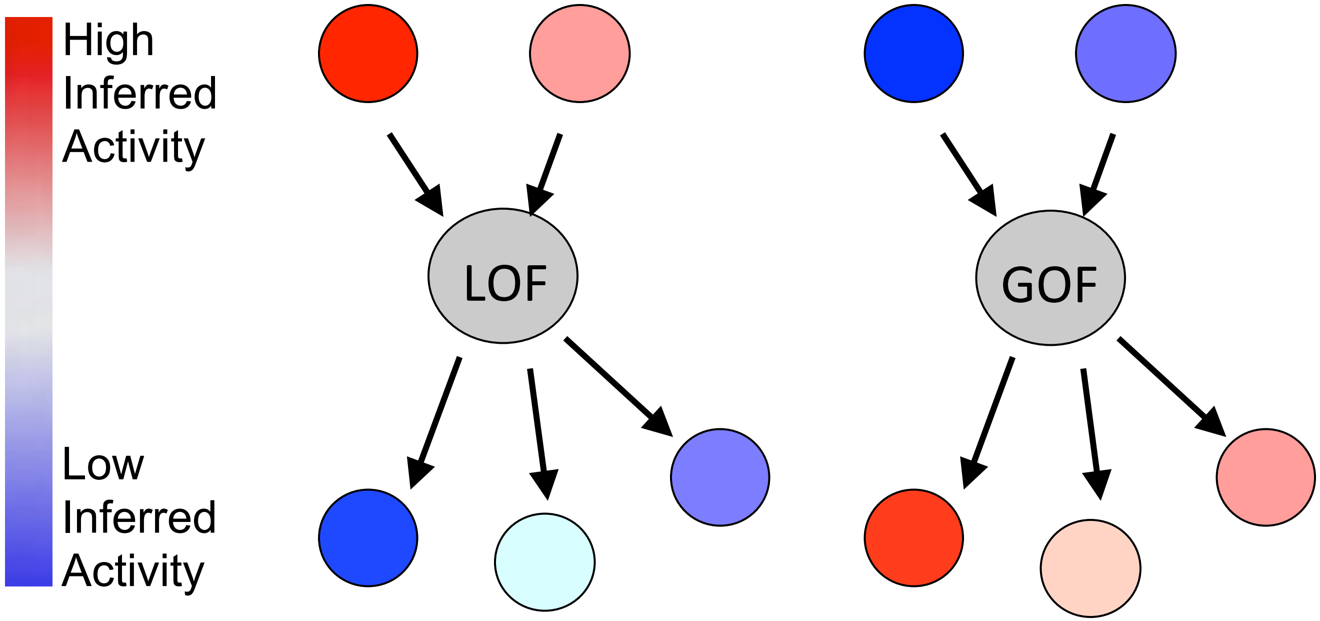
With miRNA and mRNA expression and miRNA binding sites, now possible to:

- evaluate (statistical) effect of mutations on miRNA expression
- evaluate (statistical) effect of mutations on target transcripts

Outline

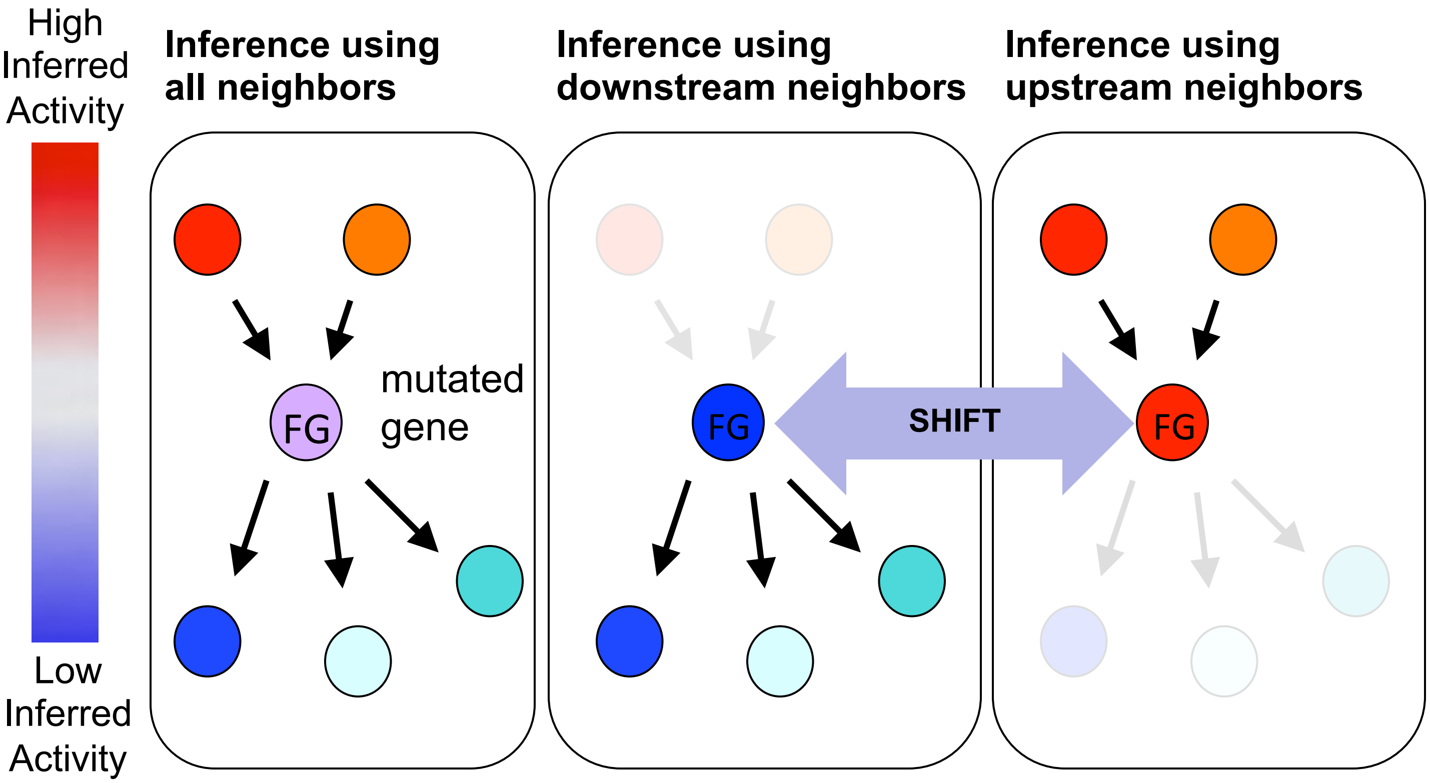
1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
    a) Collecting common resources
    b) Pilot datasets
    c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
    a) Datasets pilot-50 (Train 1) Broad calls, public-607
    b) Compare submitted annotations
    c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
    a) TCGA-505, GBM-27 (Nuria)
    b) Simulated data (Gaddy, Inigo)
    c) Compare submitted significance analyses (Nuria)
4) Example of downstream analyses
    a) Analysis of non-coding RNA (Jakob) - annotation compilation, miRNA profiling (slides from Todd Johnson)
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

Paradigm-Shift: Consequences of gain and loss -of-function on pathways



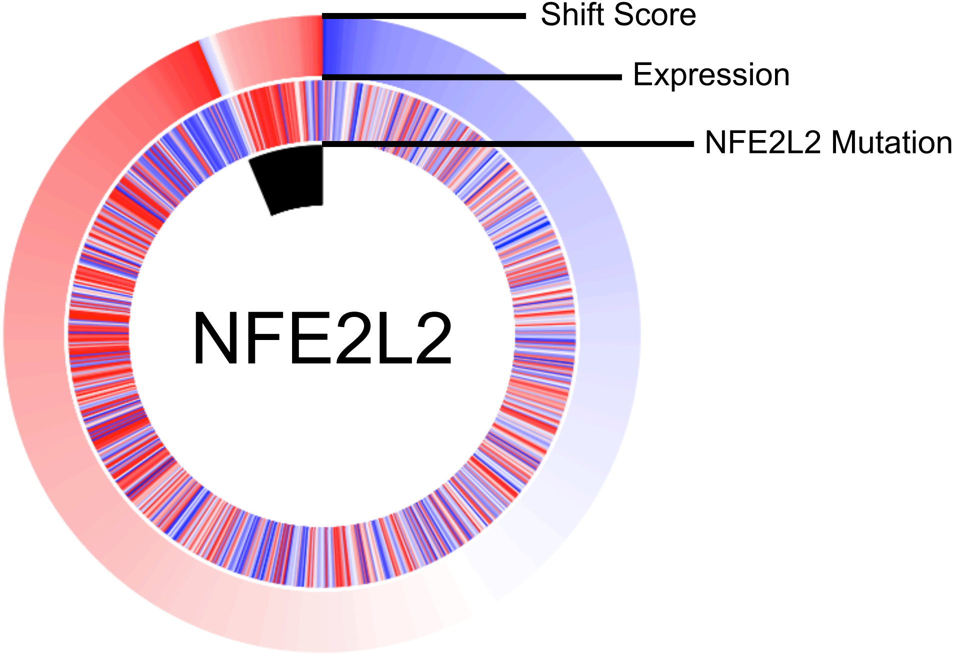
Ng et al. 2012. *Bioinformatics*

PARADIGM-SHIFT: Predicting the Impact of Mutations On Genetic Pathways

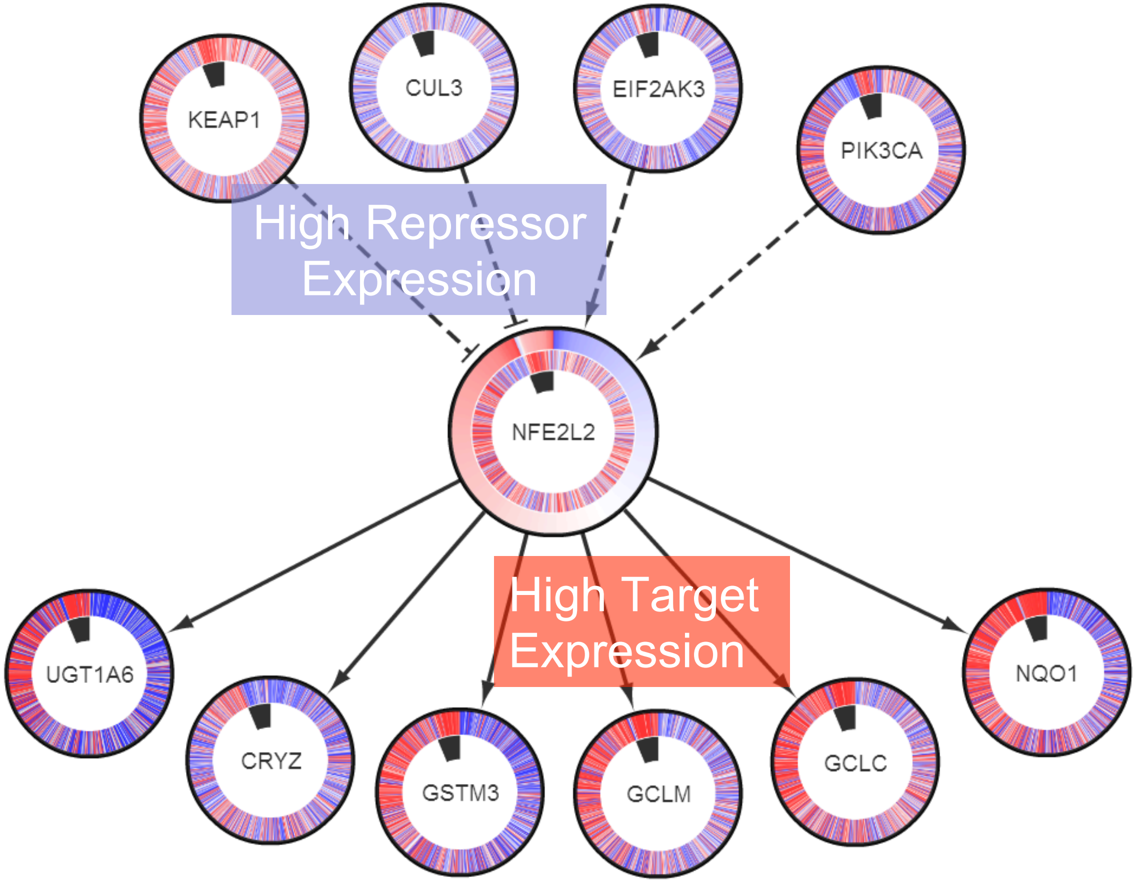


Ng et al. 2012. *Bioinformatics*

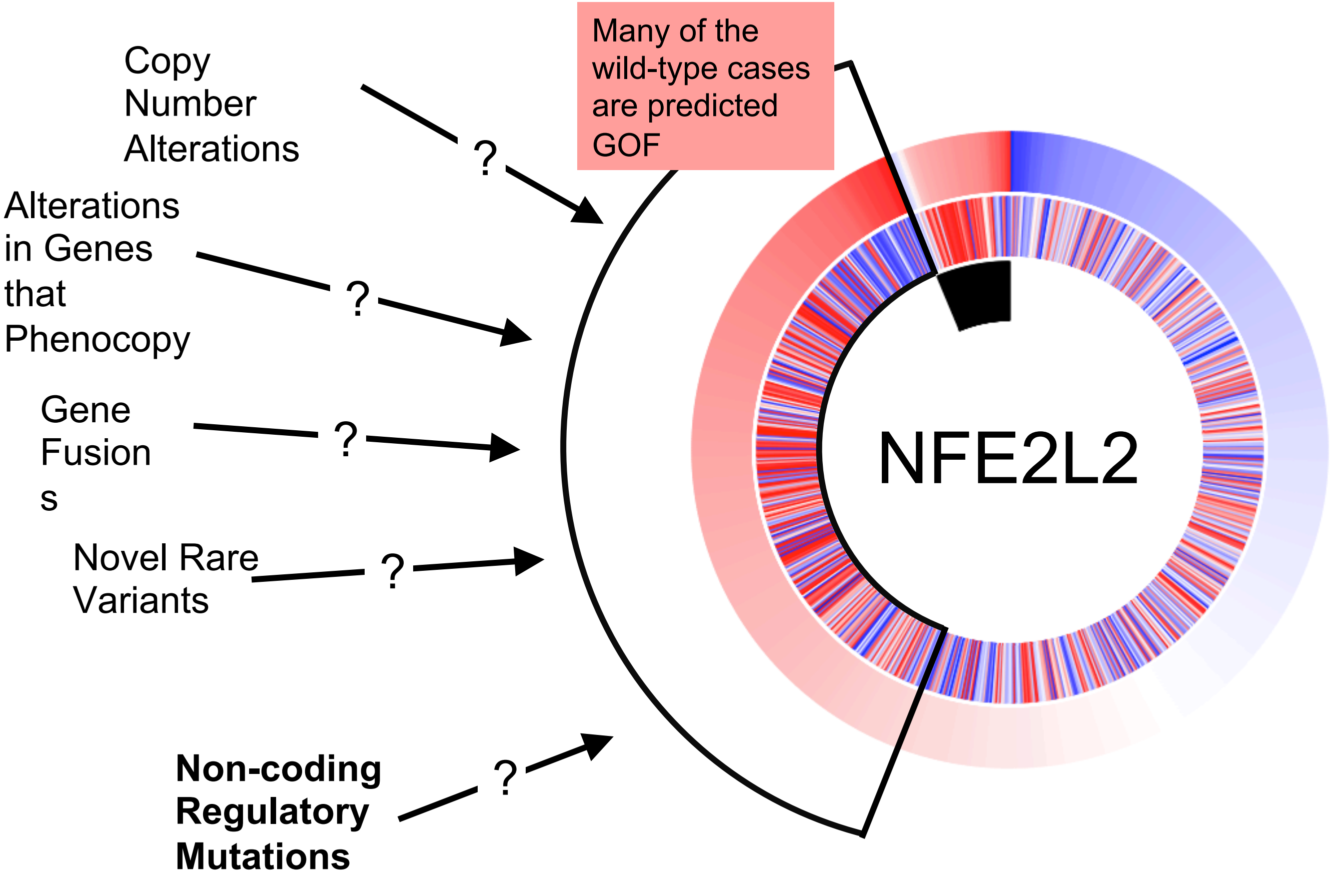
PARADIGM-SHIFT predicts gain-of-function of NFE2L2 across Pan-Cancer 12



Surrounding pathway around NFE2L2 shows transcriptional activation of targets in mutant patients



PARADIGM-SHIFT predicts gain-of-function for many NFE2L2 wild-type patients



Many of the wild-type cases are predicted GOF

Copy Number Alterations

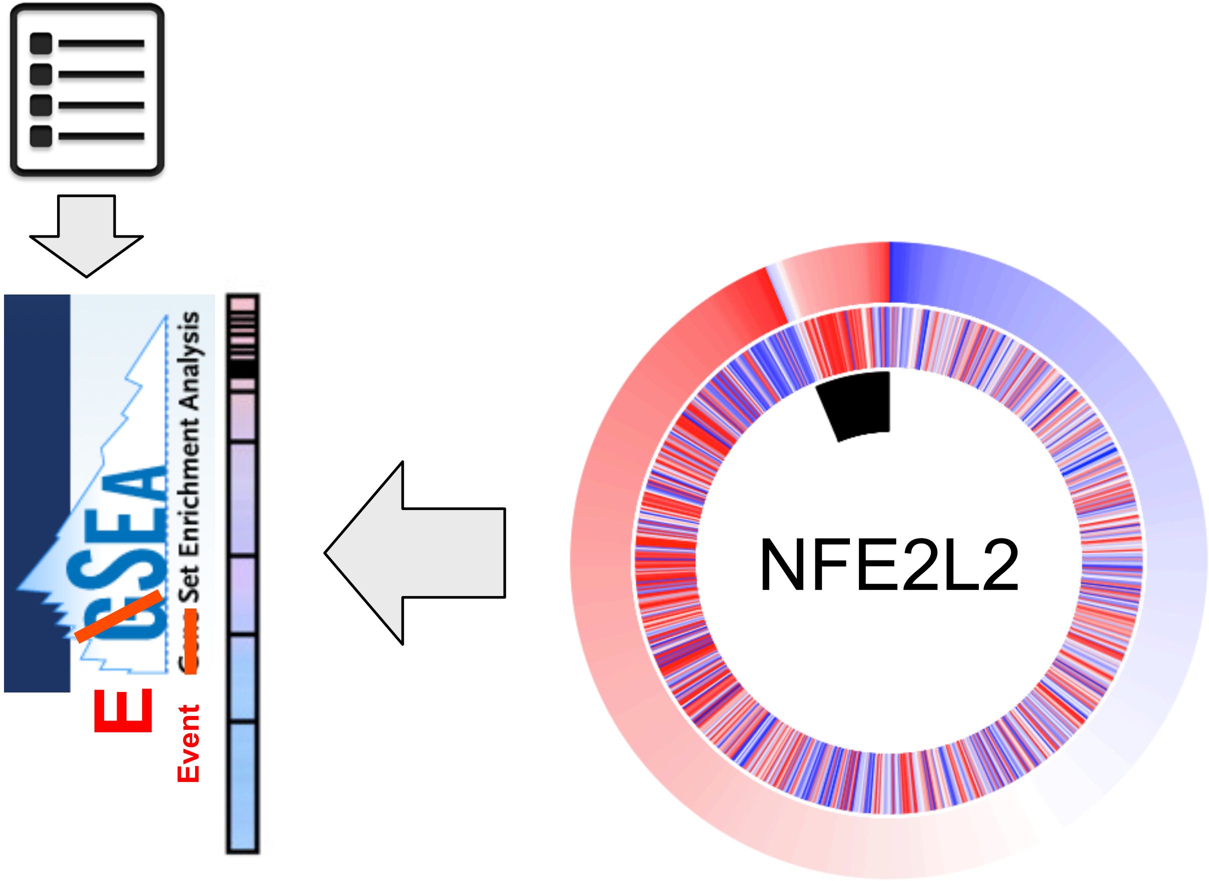
Alterations in Genes that Phenocopy

Gene Fusions

Novel Rare Variants

Non-coding Regulatory Mutations

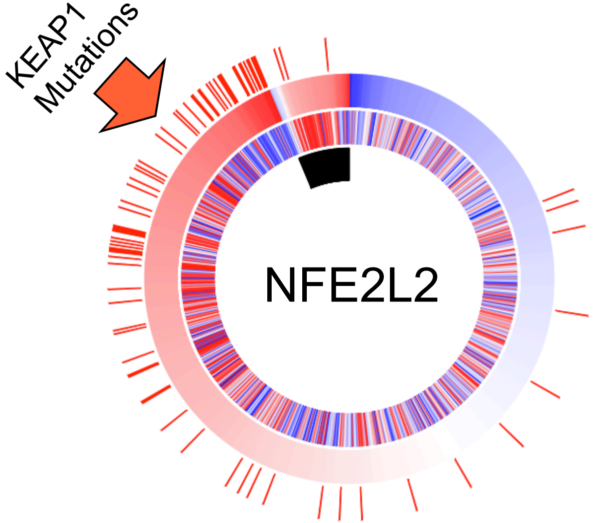
Identifying associated events that can explain PARADIGM-SHIFT predictions in wild-type cases



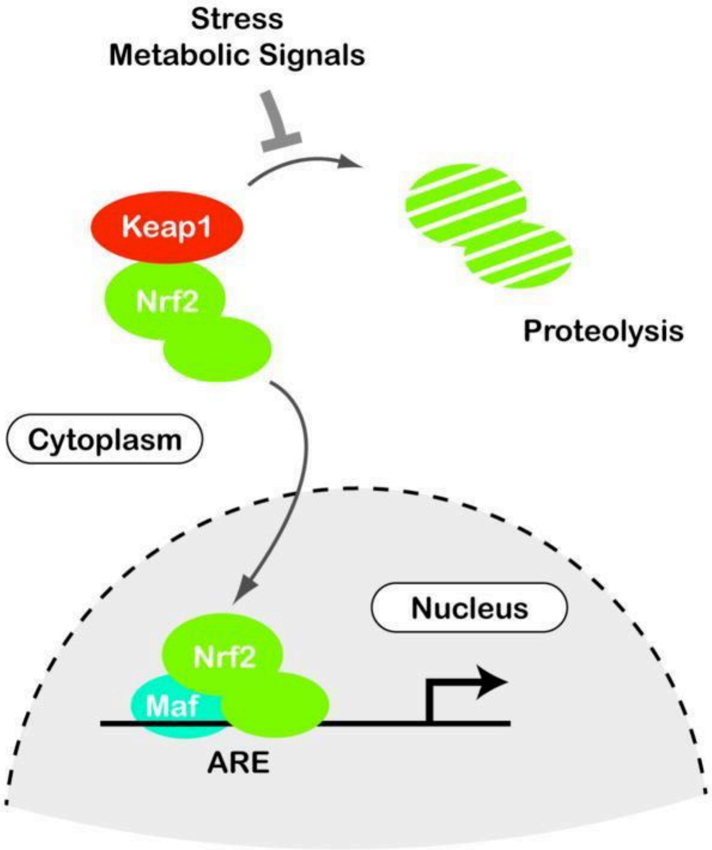
Mutations in KEAP1 are significantly associated with predicted Nrf2 (NFE2L2) pathway activation

Gene	ESEA Score	P-value
KEAP1	**0.68**	**< 0.0001**
WASH3P	-0.66	< 0.0001
COL6A6	0.45	0.0003
MYH6	0.48	0.0004

* Benjamini-Hochberg, q-value < 0.05

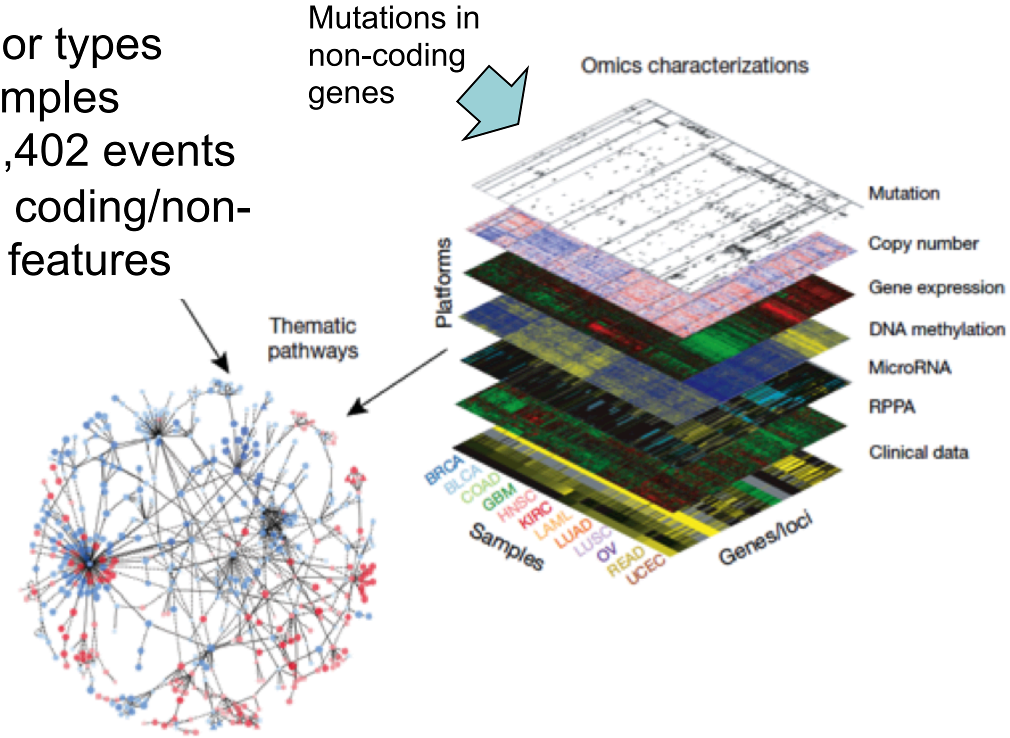


KEAP1 regulates the degradation of Nrf2 (NFE2L2)



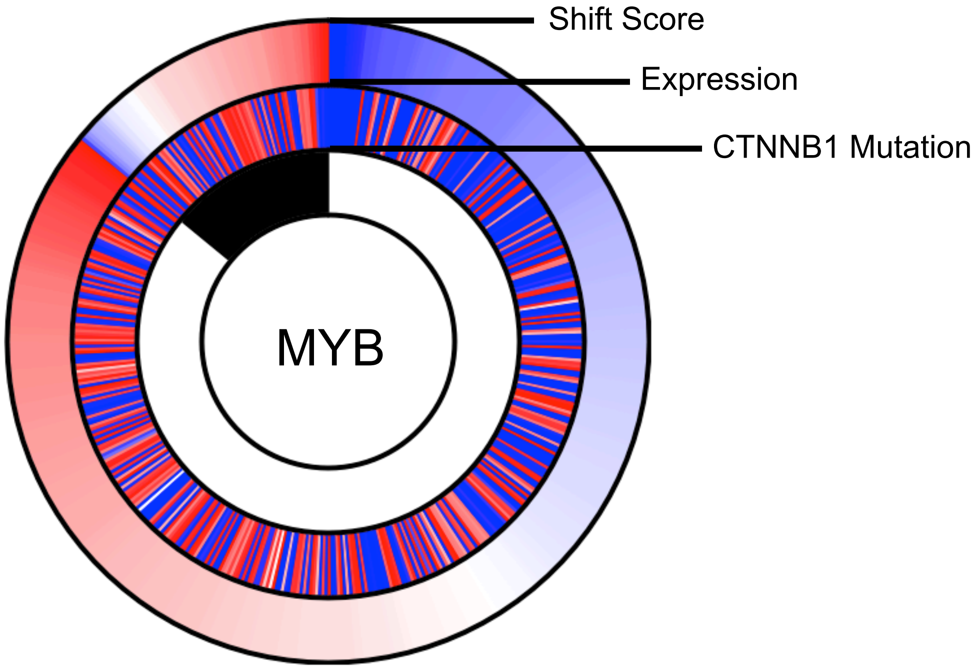
Application of Paradigm-Shift to Pilot-505

- 14 tumor types
- 505 samples
- 18,497,402 events
- 31,350 coding/non-coding features

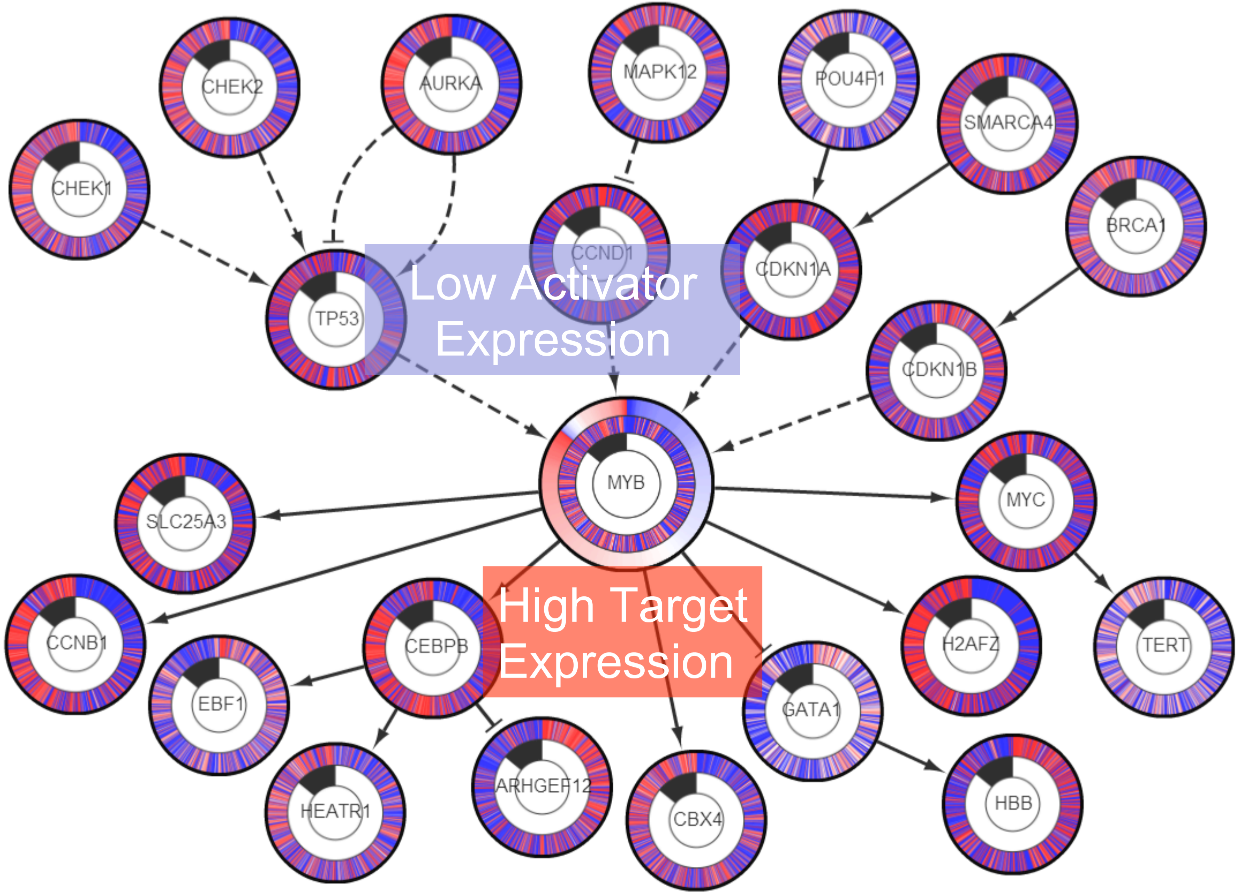

Mutations in non-coding genes

Fredriksson et al. 2014. *Nature Genetics*

Mutations in MYB in the Pilot-505 are predicted by PARADIGM-SHIFT as activating



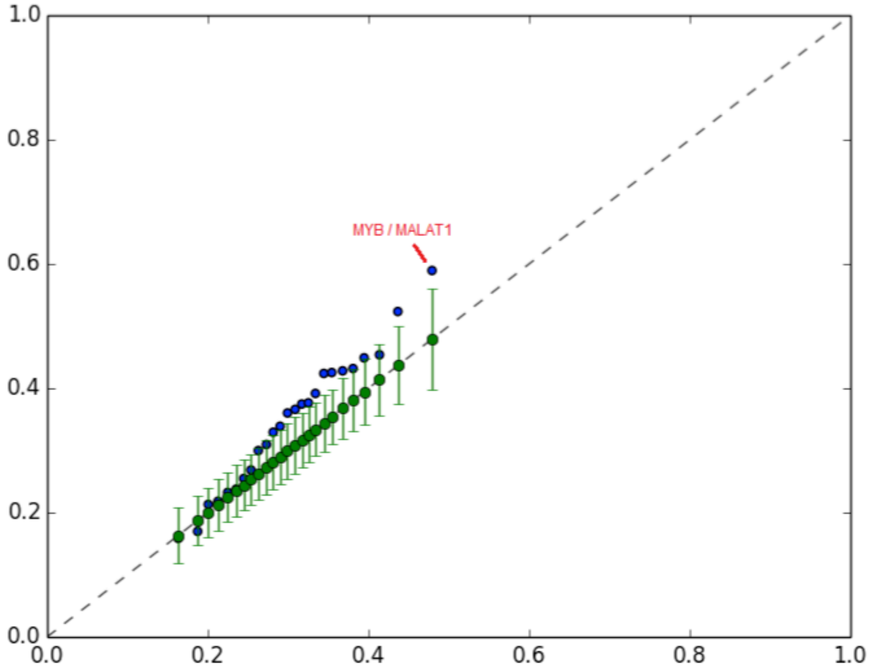
Neighborhood view of the Myb Activating Prediction (data is pilot-505)



Low Activator Expression

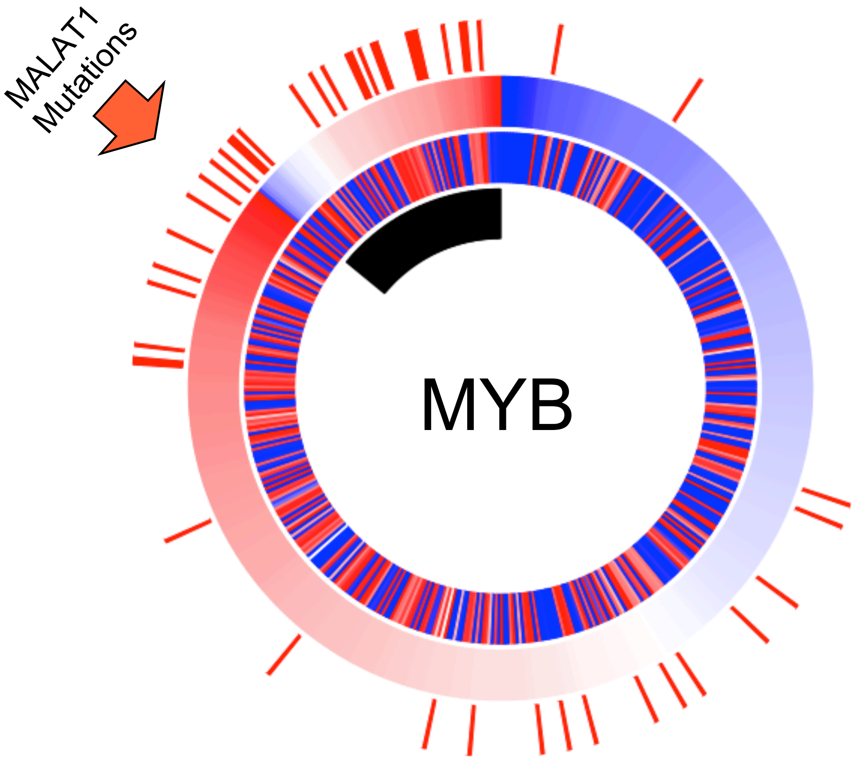
High Target Expression

Mutations in the lncRNA MALAT1 are correlated with predicted MYB pathway activation in pilot-505

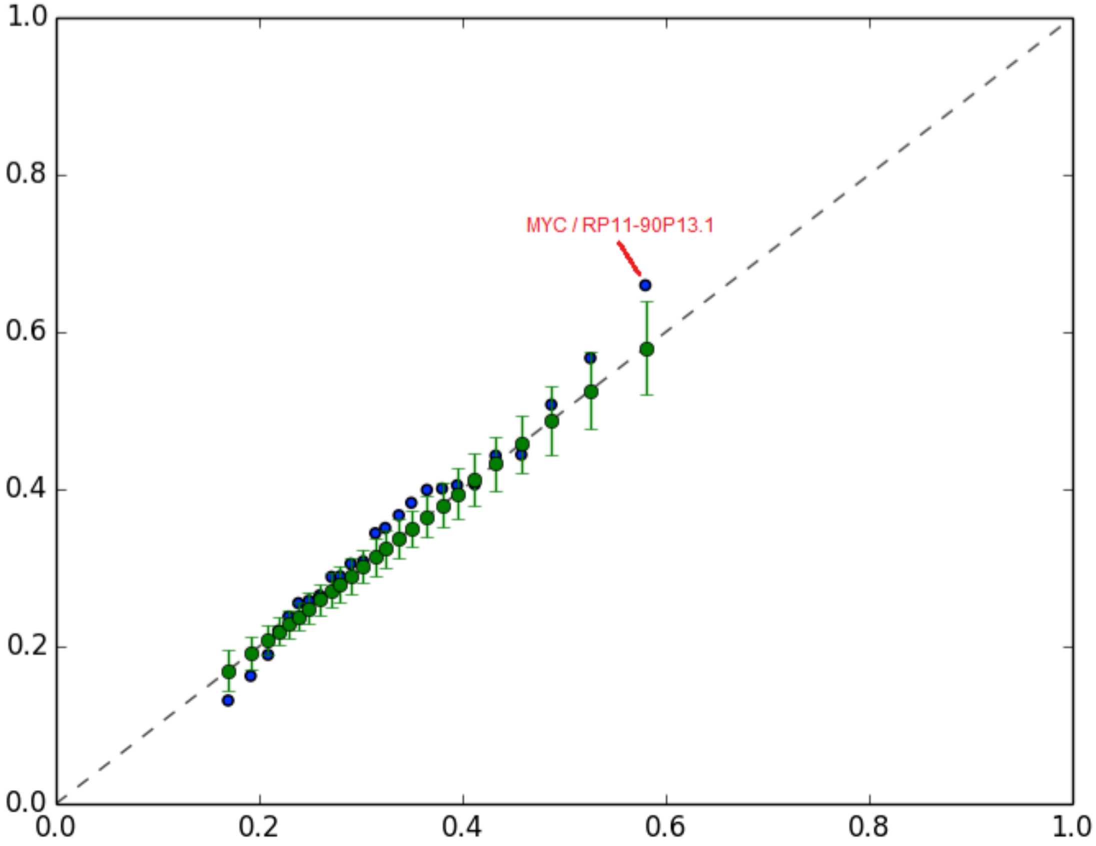


- QQ-plot shown for observed vs expected ESEA (GSEA) score
- Expected scores generated by producing 100 balanced permutations (balanced by permuting the same number of events given tumor type) then plotting the average score of each quantile across the 100 permutations
- Error bars indicate P < 0.05 based on the variance of scores across each quantile for the 100 balanced permutations

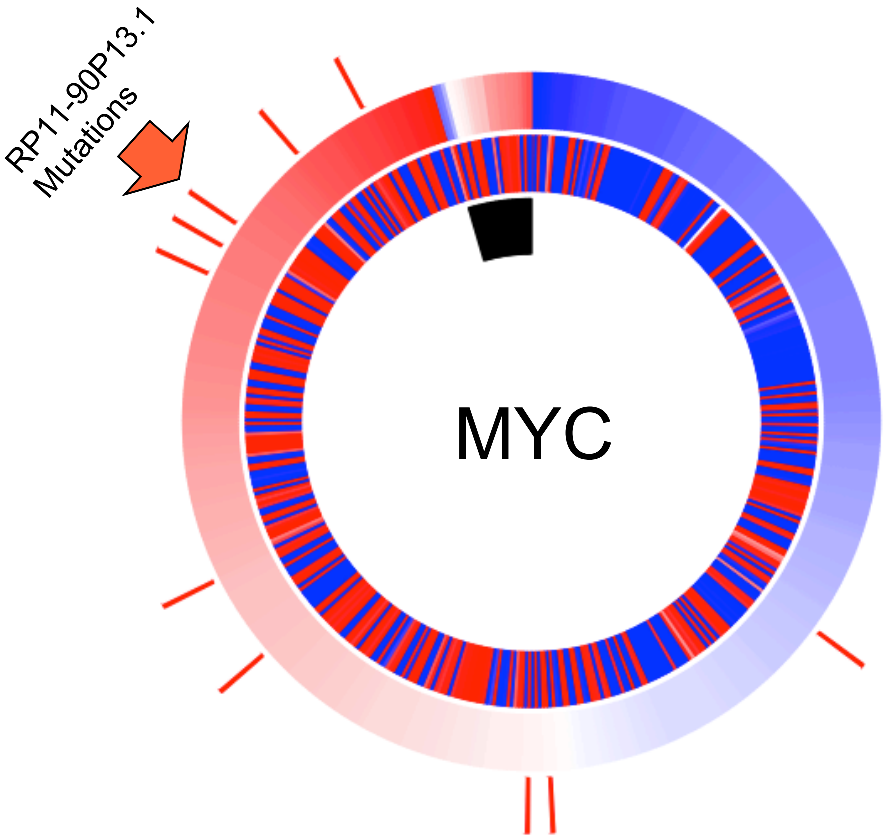
Mutations in the lncRNA MALAT1 are correlated with predicted MYB pathway activation in Pilot-505



Putative association between RP11-90P13.1 and MYC pathway activation (Pilot-505)



Putative association between RP11-90P13.1 and MYC pathway activation (Pilot-505)

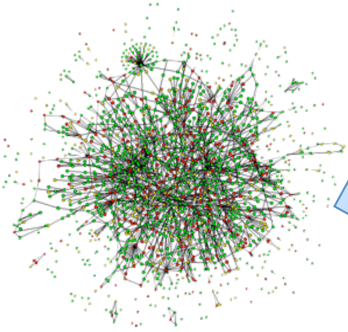


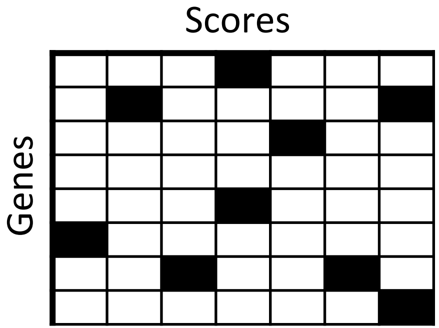
HotNet2

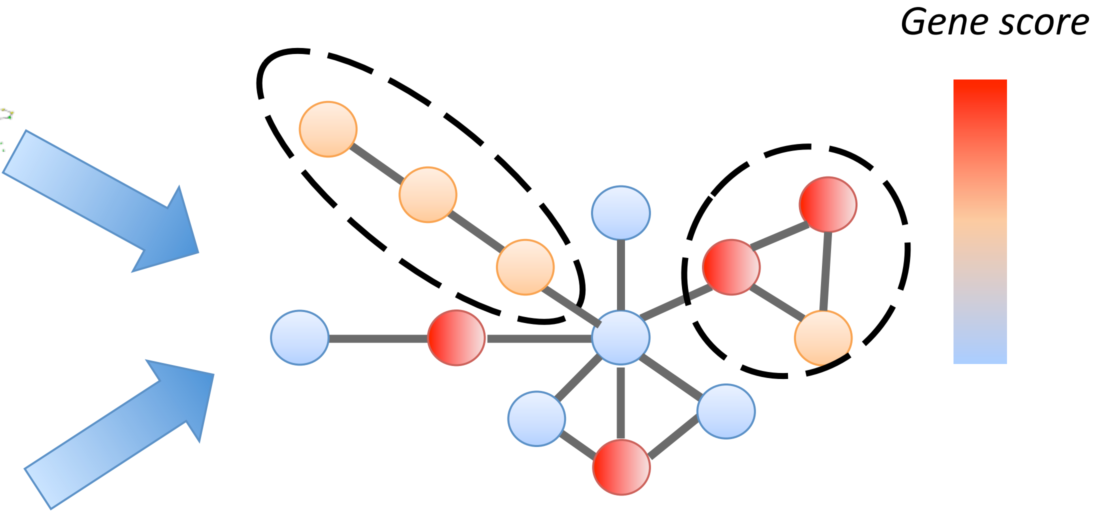
Significantly Altered Subnetworks

Question: Given network labeled with vertex scores, are these scores clustered on network?

Nodes = genes/proteins
Edges = (pairwise) interactions



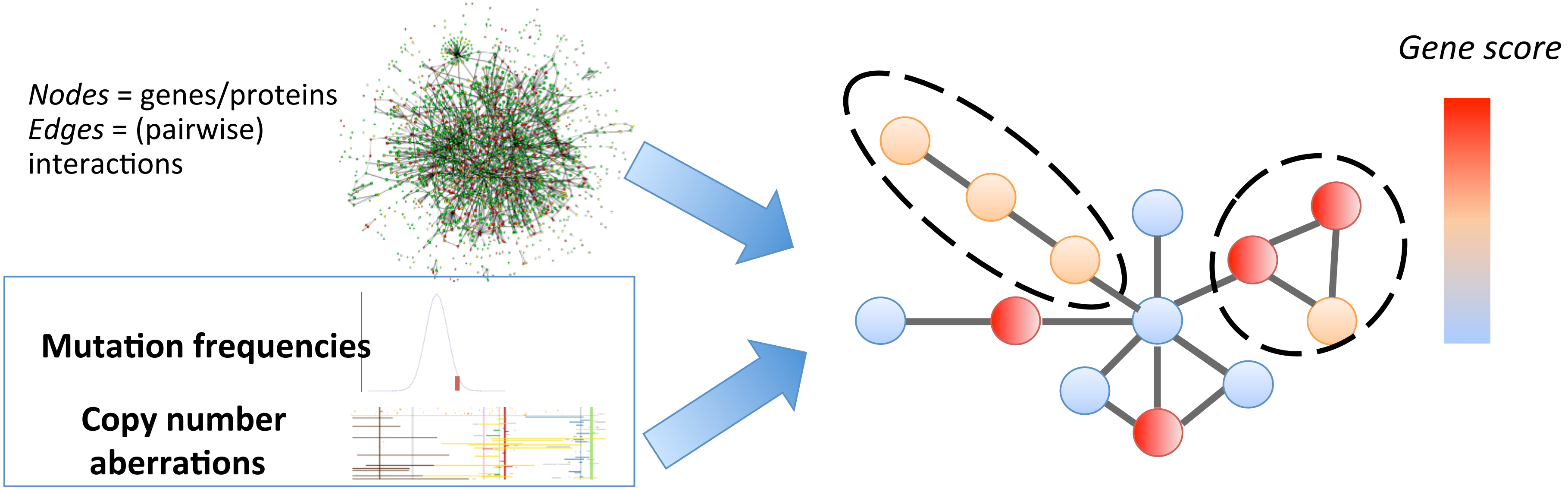




Leiserson, Vandin, et al. *Nature Genetics* (2015)

HotNet2

Significantly Mutated Subnetworks



Nodes = genes/proteins
Edges = (pairwise) interactions

Mutation frequencies

Copy number aberrations

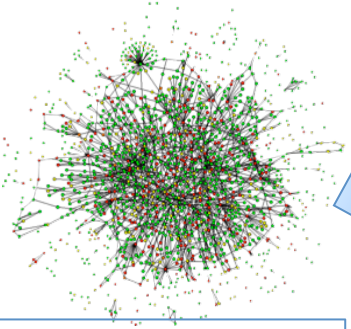
multiple TCGA papers…

Leiserson, Vandin, et al. *Nature Genetics* (2015)

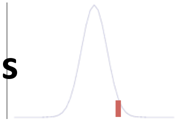
HotNet2

Significantly Mutated Subnetworks

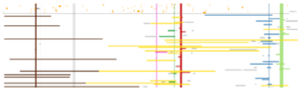
Nodes = genes/proteins
Edges = (pairwise) interactions



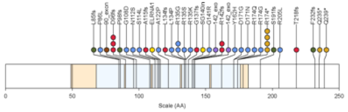
Mutation frequencies



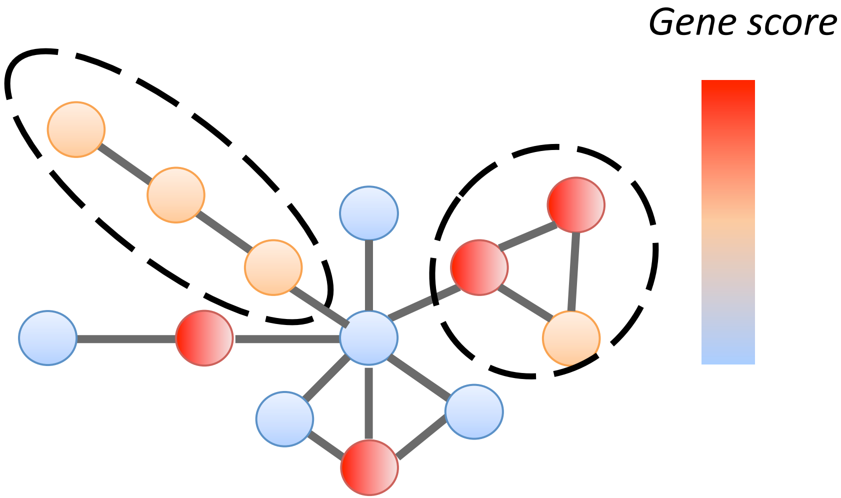
Copy number aberrations



Driver gene scores



MutSigCV, Music, Oncodrive-FM…



Pilot-505 with Oncodrive-FM scores:
In progress…

Leiserson, Vandin, et al. *Nature Genetics* (2015)

Outline

1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
    a) Collecting common resources
    b) Pilot datasets
    c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
    a) Datasets pilot-50 (Train 1) Broad calls, public-607
    b) Compare submitted annotations
    c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
    a) TCGA-505, GBM-27 (Nuria)
    b) Simulated data (Gaddy, Inigo)
    c) Compare submitted significance analyses (Nuria)
4) Example of downstream analyses
    a) Analysis of non-coding RNA (Jakob) - annotation compilation, miRNA profiling (slides from Todd Johnson)
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

Hypothesis testing

- FDR on all hypotheses (even for genes w/o mutations)
- Restricted hypothesis testing (RHT) -- Lawrence et al. *Nature* (2014)
- Potential approach -- use weighted BH FDR   Genovese et al (Biometrika (2006), 93, 3, pp. 509–524)

wBH:
1) Define W_i such that the average of them is 1
2) Calculate weighted p-values wP_i = P_i / W_i
3) perform standard BH on wP_i using standard cutoff

How to choose W_i? It is our choice
Since the average needs to be 1 we have the equation
300*x + (20000-300)*y = 20000
where x is the weight for the pan-can genes (~300) and y is for the rest.
Still this leaves one degree of freedom so I recommend we split the 20000 evenly to the two components, i.e
300*x = 10000 --> x = 33.333
19700*y = 10000 --> y = 0.5076
Basically, all p-values of the pan-can genes are decreased by a factor of 33.333 and the p-values of the non pan-can genes are roughly doubled.

Outline

1) Overview of the meta-group 2-5-9-14 (Nuria / Ekta)
    a) Collecting common resources
    b) Pilot datasets
    c) Reference Annotations sub-group (Ekta)
2) Annotation exercise (Ekta, Esther)
    a) Datasets pilot-50 (Train 1) Broad calls, public-607
    b) Compare submitted annotations
    c) Annotation tracks and mapping to ENCODE (Gaddy, Paz)
3) Signals for positive selection exercise
    a) TCGA-505, GBM-27 (Nuria)
    b) Simulated data (Gaddy, Inigo)
    c) Compare submitted significance analyses (Nuria)
4) Example of downstream analyses
    a) Analysis of non-coding RNA (Jakob) - annotation compilation, miRNA profiling (slides from Todd Johnson)
5) Pathway analyses (Josh Stuart, Ben Raphael)
6) Discuss staged statistical analysis to maximize potential discoveries (Gaddy)
7) Next Steps

Next Steps

- Standardize VCF for submissions
- Define common annotation table formats
- Generate consensus variant annotations
- Use Synapse with “annotations” and “provenance”

Acknowledgements

PCAWG-2

Gad Getz
Mark Gerstein
Ekta Khurana
Esther Rheinbay
Benedikt Brors
Lars Feuerbach
Carl Herrmann
Jan Korbel
Yao Fu
Todd Johnson

Jing Zhang
Sushant Kumar
Vasilia Rudneva
Hiroyuki Aburatani
Kenji Tatsuno
Hiroki Ueda
Jung Kyoon Choi
Woojin Yang
Youngil Koh
Jong-Sun Jung

Ewan Birney
Ian Dunham
Sandro Morganella
Alfonso Valencia
Federico Abascal
Li Ding
Matthew A. Wyczalkowski
Michael D. McLellan
Reyka Jayasinghe

PCAWG-9

Michael Lawrence
Nuria Bigas-Lopez
Gad Getz
Julian Hess
Nick Haradhvala
Petar Stojanov
Abel Gonzalez-Perez
Li Ding
Hidewaki Nakagawa
Tatsuhiko Tsunoda

Nick Haradhvala
Petar Stojanov
Beifang Niu
Ivo Gut
Roderic Guigó
Rory Johnson
Jose MG Izarzugaza
Cenk Sahinalp
Ken Chen
Loris Mularoni

Sabari Radhakrishnan
Jose MG Izarzugaza
Andrés Arturo Lanzós
Simon C Heath
Keunchil Park
Alfonso Valencia
...

PCAWG-5

Ben Raphael
Josh Stuart
Kevin White
Wei Jiao
Guanming Wu
Lucas Lochovsky
Mark Rubin
Rodero Guigo
Chis Benz

Christina Yau
Ted Goldstein
Max Leiserson
Hsin-Ta Wu
Fabio Vandin
Gary Bader
Juri Reimand
Mohammed Helmy
Christian Von Mering

Abdullah Kahraman
Miguel Vazquez
Victor de la Torre
Marc Marti-Renom
Ivo Gut
Francisco Martinez-Jimenez
David de Juan
Steven Van Laere
Jan Fostier

Kathleen Marchal
Hyung Lae-Kim
Sungsoo Yoon
Youngwook Kim
Kiejung Park

PCAWG-14

David Wheeler
Jakob Skou Pedersen
Todd A. Johnson
Samir B. Amin
Mark P. Hamilton
Hannah Cheung
Henrik Hornshøj
Morten Muhlig Nielsen
Johanna Bertl
Asger Hbolth

Guo Qianyun
Song Cao
Sean E. McGuire
Reyka Jayasinghe
Matthew Wyczalkowski
Michael D. McLellan
Richard Sallari
Tatsuhiko Tsunoda
Li Ding
Nuria Bigas-Lopez

EXTRA SLIDES

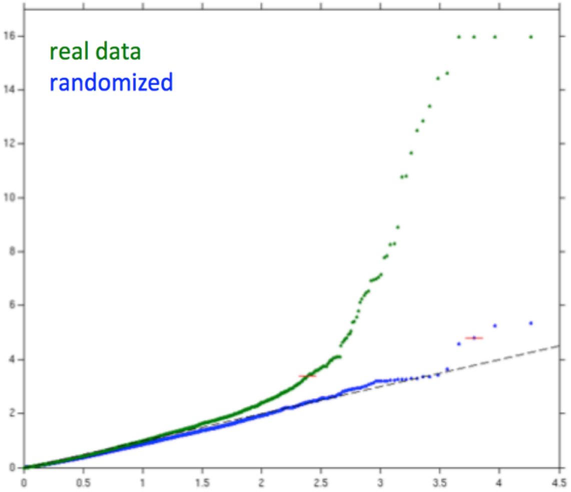
Boston Slides

Link to Boston slides (for reference)
https://docs.google.com/a/upf.edu/presentation/d/1VLrmkNCVuTVsD9xrGbranm-VrEhfeOUL7NsoP8ZTbt8/edit#slide=id.g4a046587b_00

PCAWG 505 pilot

mutation significance analysis

PANCAN 505 coding

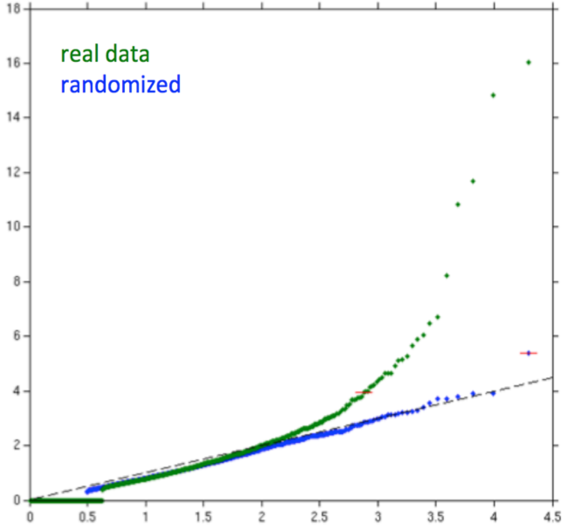


TP53
PIK3CA
BRAF
IDH1
CDKN2A
PTEN
NFE2L2
VHL
APC
KRAS
STK11
NRAS
AKT1
MUC4
HLA-DRB1
ARID1A
RNF43
PIK3R1
ARHGAP35
HLA-B
CBFB
B2M
PRSS3
FBXW7
POLDIP2
LARP4B

CDH1
MPHOSPH9
PITRM1
MUC17
KRTAP4-5
ACVR1B
MKL2
LCT
SLC38A1
USP6
ZFP36L2
BCLAF1
ATXN1
IL23R
C15orf23
THEMIS
TBXA2R
HRC
CTCF
INPPL1
IL7R
PCDHGA1
EYA4
TGFBR2
MYH7B
SVIL

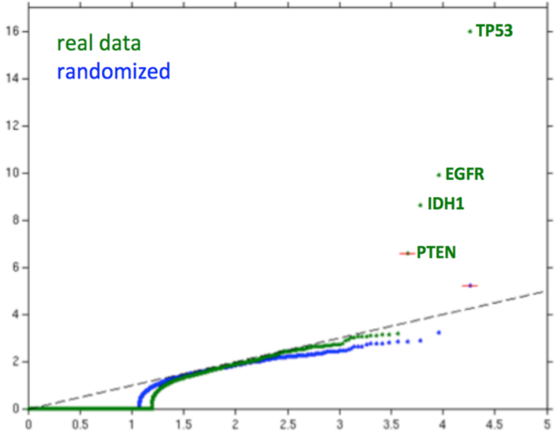
EGFR
DNAH12
DCAF4L1
AVPR1B
HRNR
HRAS
PABPC1
NOTCH1
HLA-A
OSMR
ATP1A4
MRPL32
FGFR3
TNN
IL10RB
PZP
TACC3
WNT16
CDH23
APOL2
ACOT4
CTNNB1

PANCAN 505 promoters

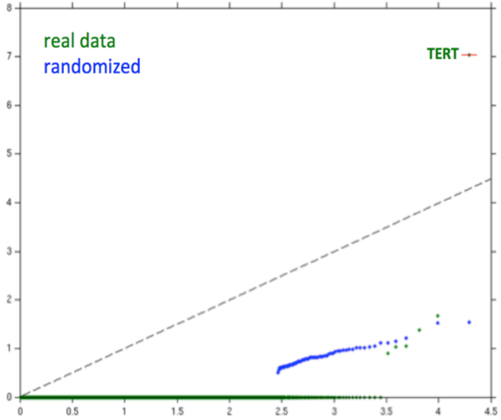


NEXN
TERT
SNX32
ZNF717
CSF2RA
SMG1
TTC40
MUC12
AKAP17A
TRIOBP
RP11-1220K2.2
BCL9
MROH2A
RNF219
CCDC66
RIC8A
AP2A1
C10orf112
RNF17
TDRD15
KNTC1
SEMA6B
STAG3
ATP10D
EIF2AK3
GOLGA6L2

GBM 27 coding



GBM 27 promoters



MutSig2CV results

Julian Hess
Nick Haradhvala
Esther Rheinbay
Mike Lawrence
Gaddy Getz

The importance of calibrated statistical tests

Methods that search for cancer genes (ie. ones that show evidence of positive selection) are based on rejecting the **null hypothesis** that the observed mutations in a gene/region are **all passengers mutations**.

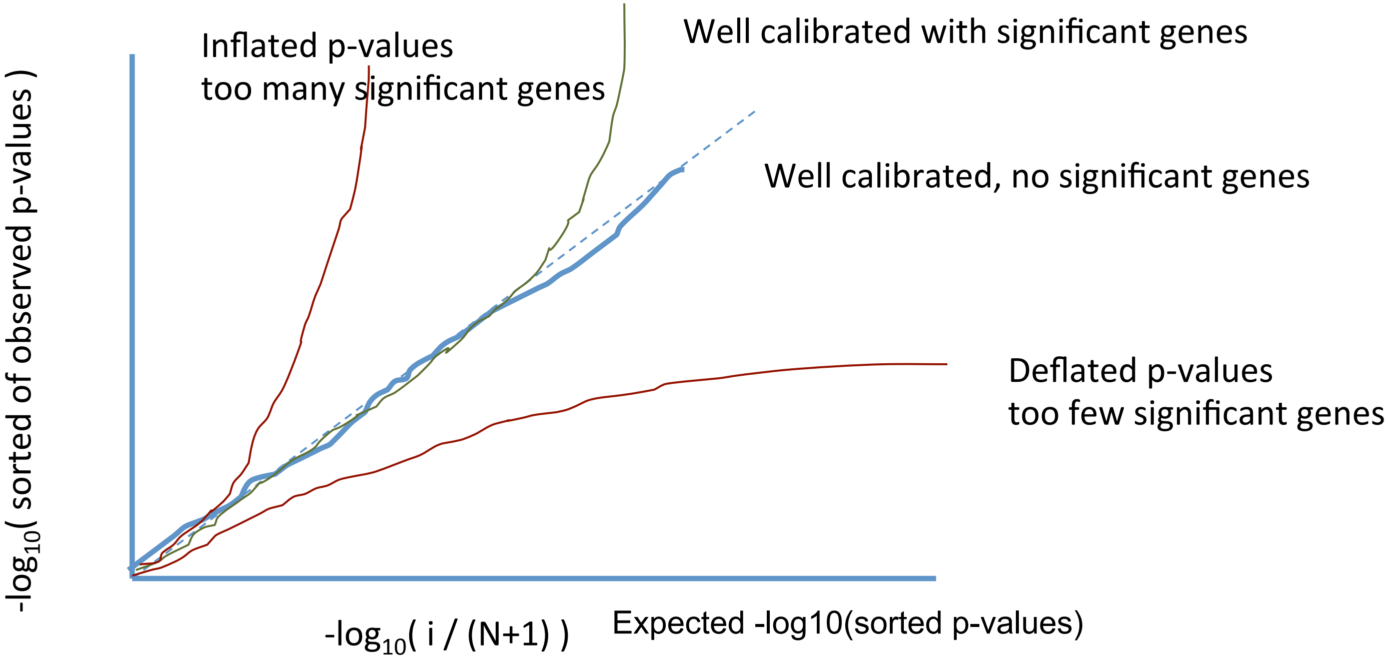
The standard procedure involves: (i) calculating **p-values** for each gene/region; (ii) correcting for **multiple hypothesis testing** (e.g. using the BH procedure); (iii) listing all genes/regions with **FDR q ≤ 0.1** (or some other accepted cutoff) as **candidate cancer genes**.

➔ The expected fraction of false positives in the list is < 10%.

For this procedure to be **valid**, the p-values should indeed reflect the null hypothesis.

Since we believe that most genes/regions do not harbor driver events, we expect the p-values of most genes/regions to be uniformly distributed (ie. follow the null hypothesis).

Example QQ plots



Do: (i) Provide a QQ plot for your test;
(ii) Carefully assess the number of hypotheses you are testing
(ii) Use a standard q-value cutoff (e.g. 0.05, 0.1, 0.25)

Don't: (i) Select a q-value cutoff that will contain only your favorite genes (e.g $q<0.001$);
(ii) Remove from your list genes that don't make biological sense

Gad Getz, Broad/MGH

PAWG-5 Pathway Analyses

Link to UCSC slides:
https://docs.google.com/presentation/d/12CoXGlbtuUSUoqI0ARqsl_wTRTkSqrZfGriDZ5UcoMU/edit?usp=sharing