Slides for
WG-2
Papers (A+F) + E
2 x 10 min + 10 min discussion

PCAWG SC call

Sep 12 2016

F: Finding drivers

A: Regulatory drivers

Papers from mega group PCAWG-2-5-9-14

For discussion:

1) Driver discovery (A+F)
(i) methodology (different signals for positive selection, simulations, QQ plots, comparing different methods, combining different methods, multiple hypothesis testing)
(ii) overall survey of genomic elements and mutations in them (broken by element, mutations type and major mutational signatures)
(ii) significant coding (including somatic hyper mutation), 3D structure
(iii) significant regulatory (including UV hotspots), correlation with expression data
(iv) power calculation (detection sensitivity, including some important blind spots, discovery power)

2) non-coding RNAs (B)
significant lncRNAs (MALAT1 + NEAT1), correlation with expression
…

3) Patient-centric view including all drivers (including copy-number and SV) (Paper D)
(i) number of drivers per patient,
(ii) which mutations in each driver are likely functional,
(iii) patients driven by mutations vs. copy-number

4) Pathways paper (Paper C)
(i) Use pathways to find additional drivers

5) Overall burden/funciotnal effect of all mutations. (Paper E)
This may also integrate with the 1st paper, depending on results.

Figure outline for driver paper

Figure 1 - overview figure
A overview of driver detection methodologies, including list of methods
B Cartoon or example QQ plot
C Overview of cohorts analyzed
D Overview of genomic regions, mutations and major signatures

Figure 2 - Combining results from different methods, especially those that are correlated
A example of p-values/significant genes from different methods on coding genes for one representative cancer type
B Illustration of correlation between methods, how to use simulated data to detect and correct for this
C Statistical strategies for combining p-values, and their effect on the list of significant driver genes
D Multiple hypothesis testing: methods and effect on the result gene list

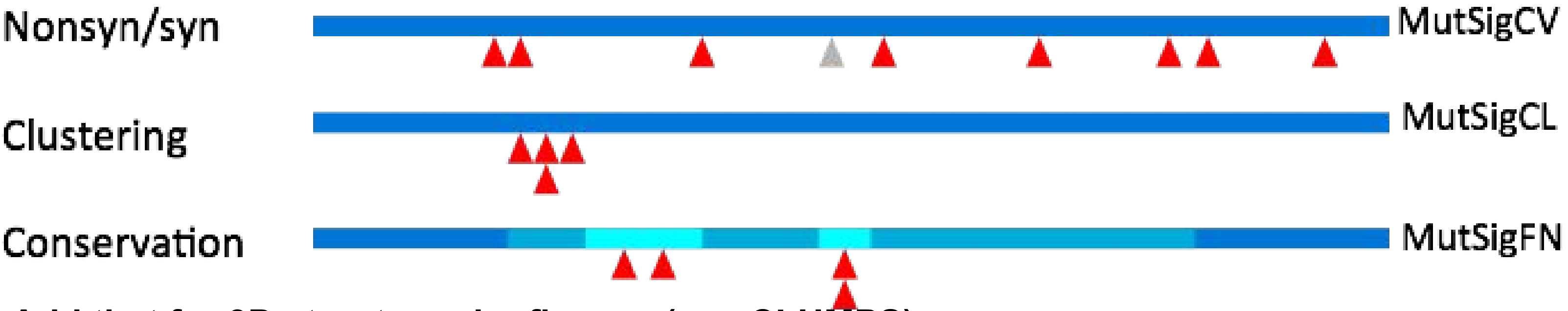
Figure 3 - protein-coding driver genes
A Significant driver genes listed by tumor cohort
B QQ plot/scatter plot of significant driver genes in individual vs combined cohorts
C Impact of coding mutations in novel driver genes (3D structure, phospho sites etc)

Figure 4 - regulatory drivers
A Significant driver elements listed by tumor cohort
B QQ plot/scatter plot of significant driver elements in individual vs combined cohorts
C, (D) Detail of novel results: correlation with expression, impact on TF binding sites, target effects for enhancers

Figure 5 - power analysis (how much can we expect to find in this dataset?)
A Overview explaining detection sensitivity and discovery power
B Detection sensitivity in different cohorts
C List/illustration of lack of sensitivity in known cancer drivers (e.g AKT1, promoters)
D Discovery power in all PCAWG cohorts

Fig 1: Significance analysis
Additional sources of evidence for positive selection

Signals of positive selection



Add that for 3D structure signficance (e.g. CLUMPS)

(1) **Genomic elements**, **somatic mutations** across a **cohort** of patients →
(2) Model for background (i.e passenger) mutations →
(3) Significance of more mutations than expected by chance (burden or dN/dS) →
(4) Correction for multiple hypothesis testing (# of elements) → **q-value**

Fig 1A: Cartoon of significance analysis with various names of methods

Lawrence et al. *Nature* (2014)

Example QQ plot (cohort = GBM, tool = MutSigNC, 5 datasets, promoters)

3 null simulated
1 null + drivers simulations
1 observed data

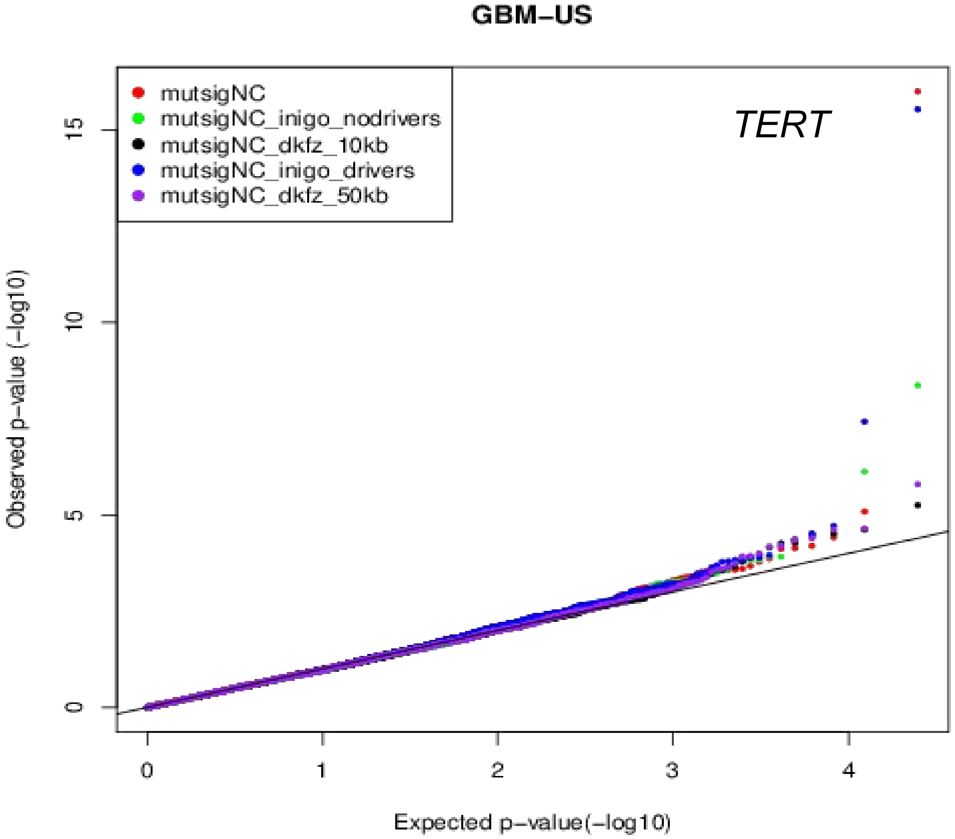
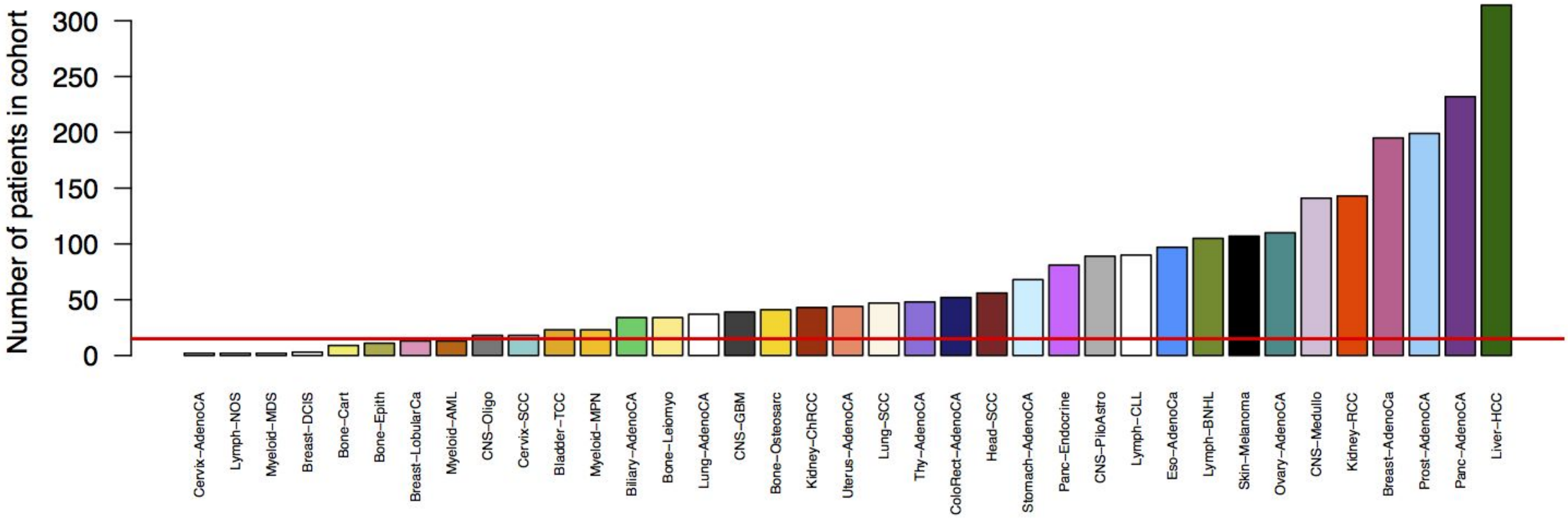


Fig 1B: Cartoon (or real data) Simulations and QQ plots of well calibrated null and well calibrated with significant genes

Esther Rheinbay, Grace Tiao (Getz lab)

Cohorts: 29 individual tumor types + 3 lineages + 1 pan-cancer = 33 cohorts

2583 representative samples
29 cohorts with 2528 cases (>97.5% of all cases)



Red line indicates 15 patients

Fig 1C: Cohorts analyzed

Overall survey of numbers/genomic elements and mutations

Include cartoon of definition of elements

- Coding sequences (20185)
- Promoters (20039)
- Enhancers (30816)
- lncRNAs (5580)
- 5’UTRs (19188)
- 3’UTRs (19369)

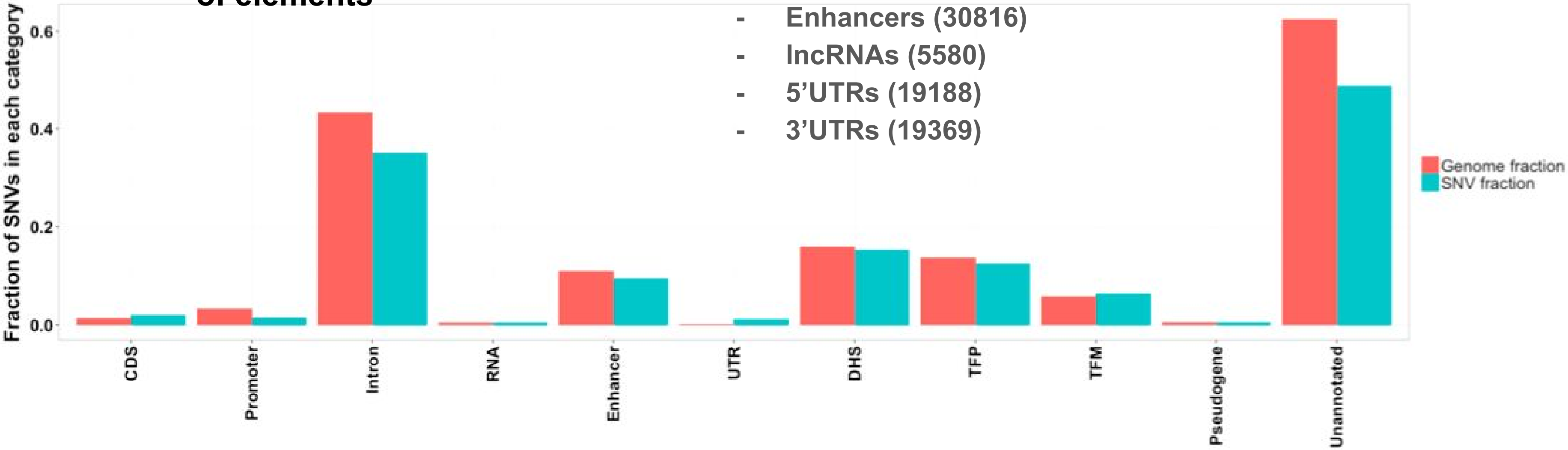


Fig 1D Genomic elements analyzed in paper, number, territory, and breakdown of mutations Mutations broken down by type (SNV, indel), XX major mutational signatures

Interval lists compiled by Morten Nielsen, Jakob Skou Pedersen and Nicholas Sinnott-Armstrong

Ekta Khurana

Fig 2: Comparison and combination of p-values from different driver detection methods

Fig 2A: significant elements of different methods for example interval list (coding)

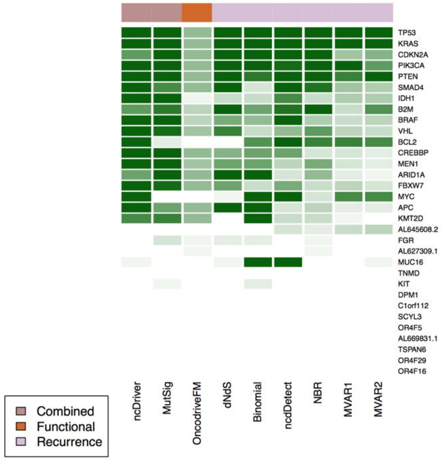
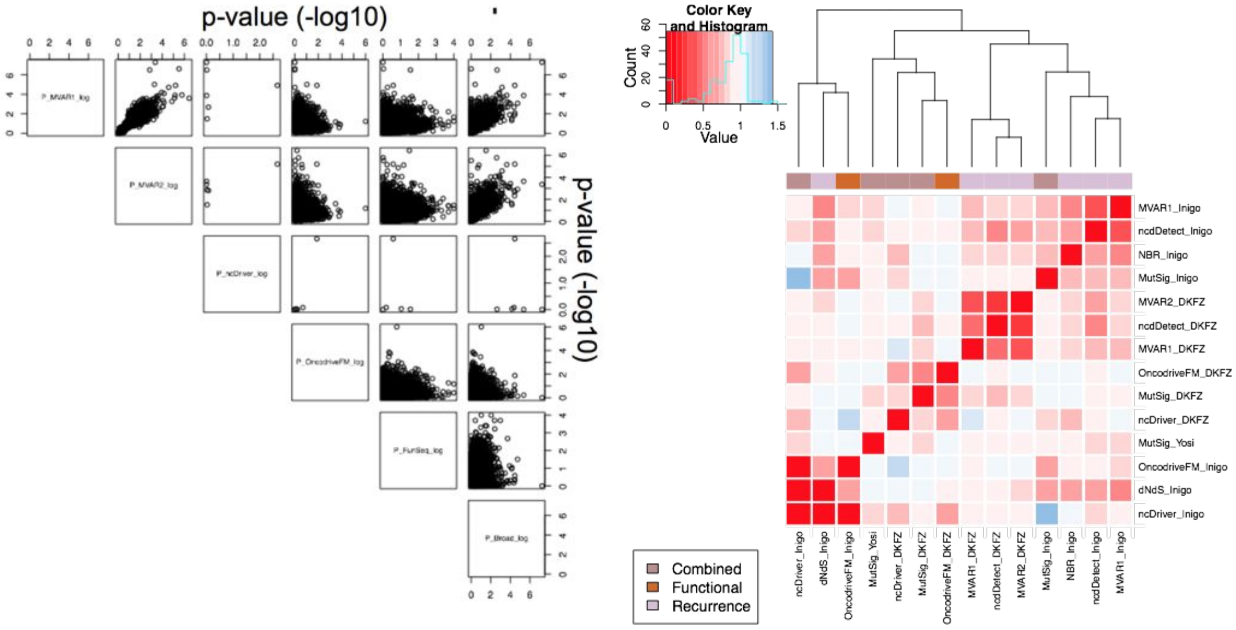


Fig 2B: Correlation structure on simulated and real data



Esther Rheinbay, Grace Tiao (Getz lab)

Fig 2: Comparison and combination of p-values

Fig 2C: Methods for combining p-values

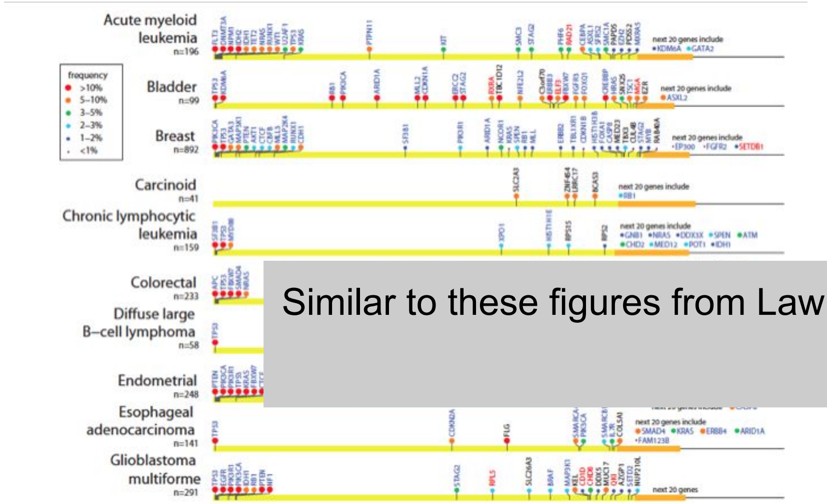
Show example results from different methods for formally combining
p-values
Compare -log p-values of different methods

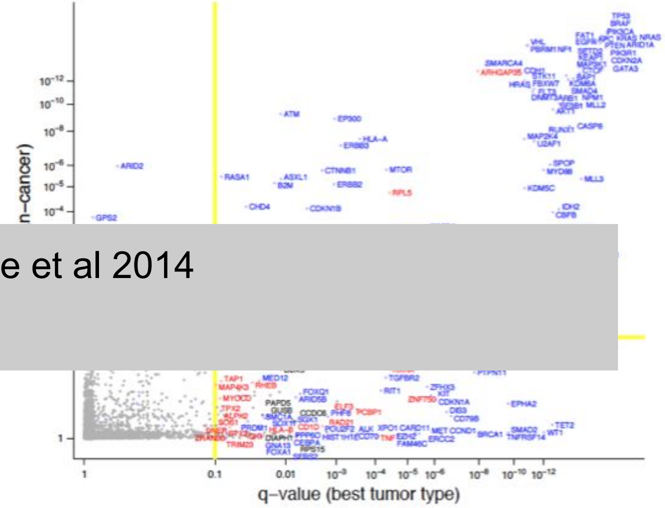
Fig 2D: Multiple hypothesis testing

Show example results for significance using restricted hypothesis testing, stratified testing, weighted hypothesis testing (IHW), standard BH

Fig 3: Significant coding regions

Fig 3A,B: Significant genes in coding regions across sets
Highlighting new findings in individual tumor types, combined cohorts or pan-cancer analysis

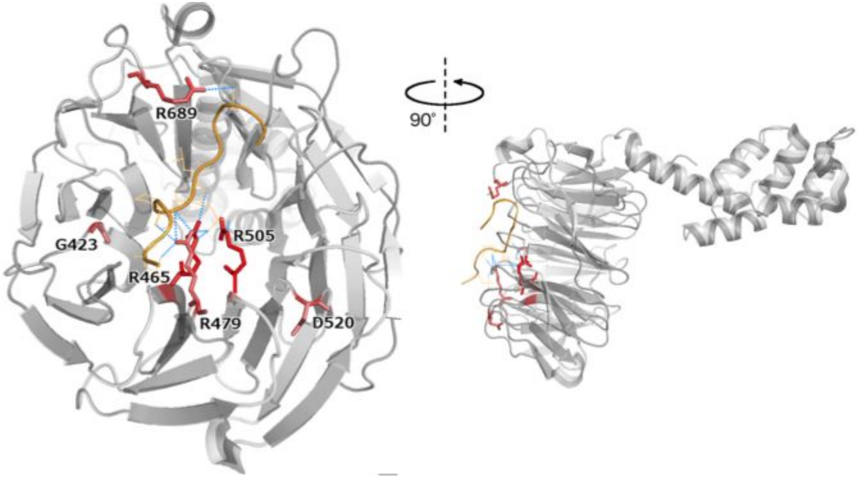


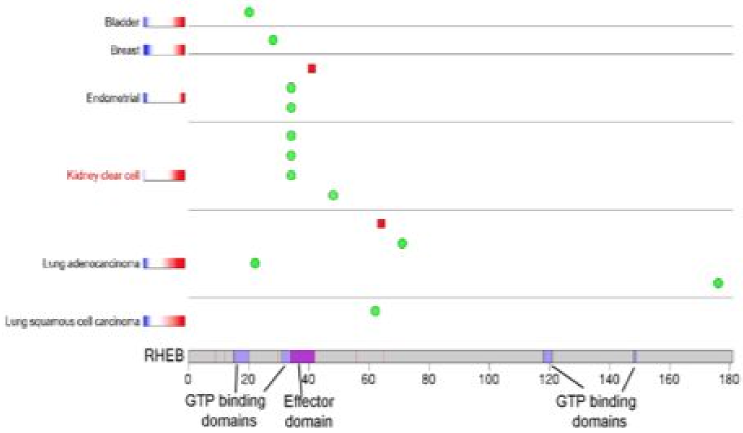


Similar to these figures from Lawrence et al 2014

Fig 3: Significant coding regions

Fig 3C: Protein structure and stick-figures for new or interesting genes, e.g. if they have different patterns in different cohorts. (novel findings)



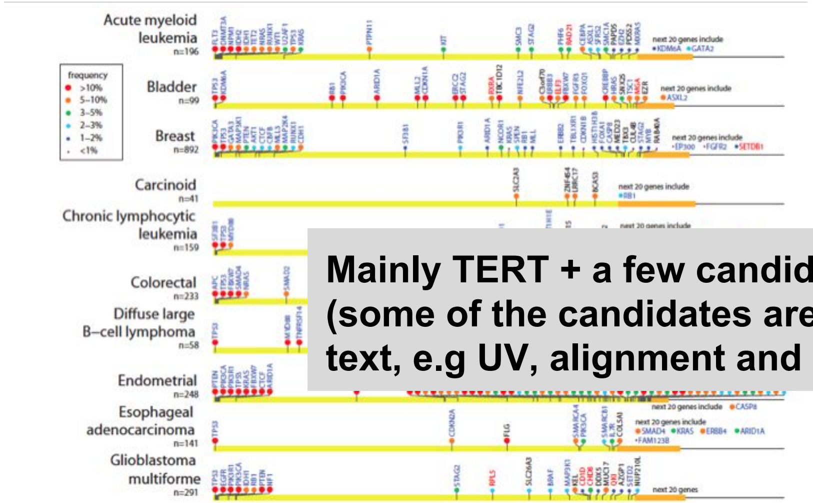


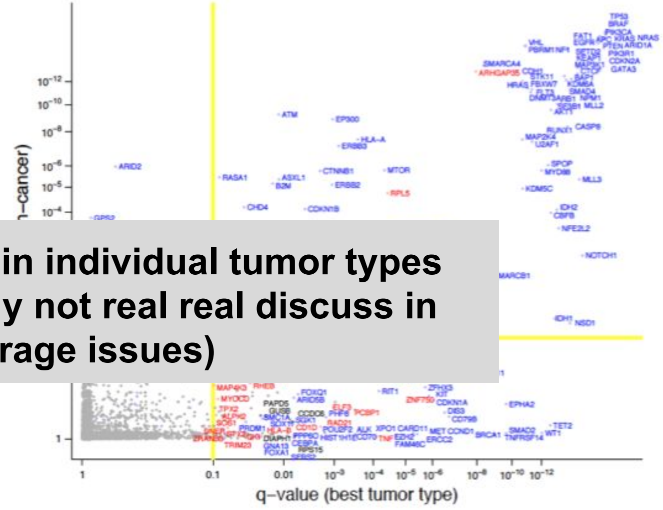
Place holders for figures for new genes

Fig 4A,B: Significant regulatory elements

Fig 4A,B : Significant regulatory elements
Highlighting new findings in individual tumor types, combined cohorts or pan-cancer analysis

Similar to this figures
From Lawrence et al



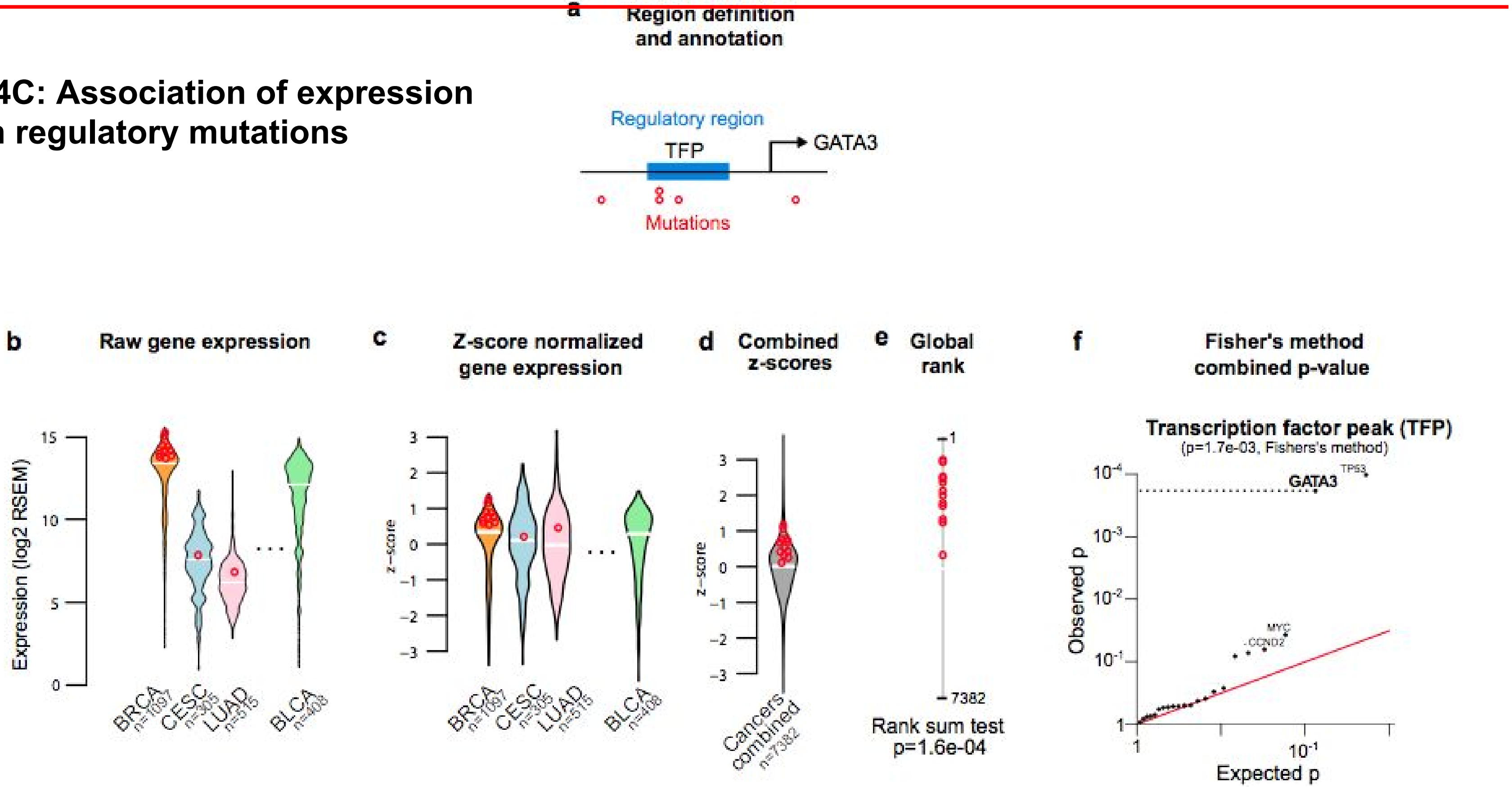


Mainly TERT + a few candidates in individual tumor types (some of the candidates are likely not real real discuss in text, e.g UV, alignment and coverage issues)

Esther Rheinbay, Gad Getz

Fig 4C: Expression data provides additional evidence for functional effect of mutations

Fig 4C: Association of expression With regulatory mutations



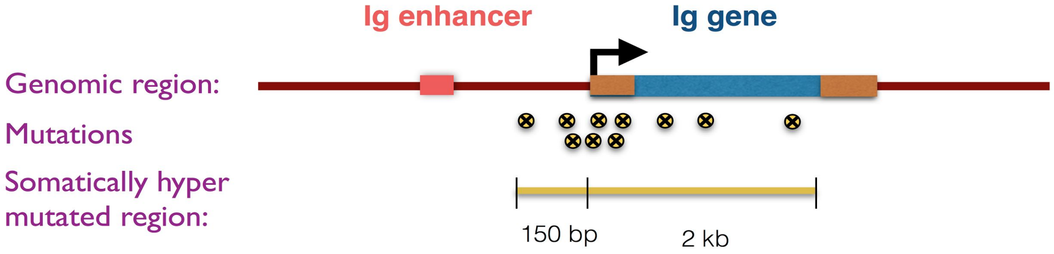
Morten Muhlig Nielsen, Henrik Hornshøj, & Jakob Skou Pedersen (*In preparation*).

Supp Fig 1: Effect of Somatic hypermutations in lymphomas

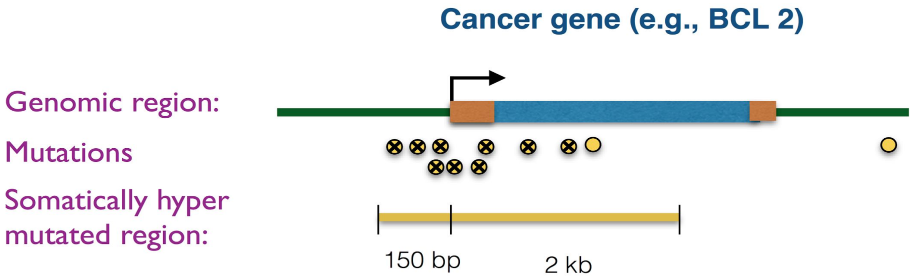
○ Normal somatic mutation
⊗ Somatic hyper mutation

Supp Fig 1A: cartoon

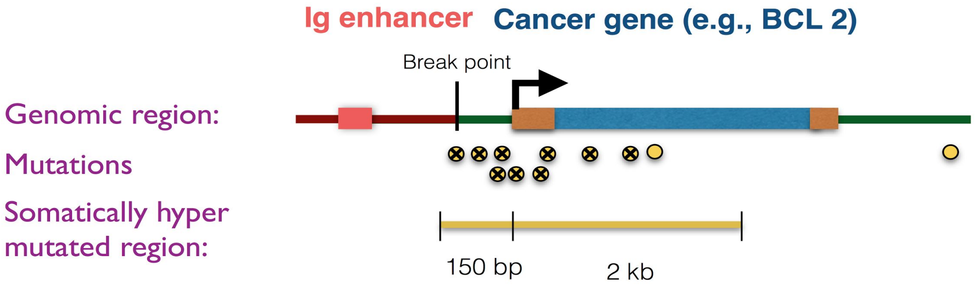
Somatic hypermutation of transcription start site regions of immunoglobulin loci in B-cells



Aberrant **off-target** somatic hypermutations in B-cell derived cancers



Translocation of cancer genes to Ig loci also causes somatic hypermutation of cancer genes



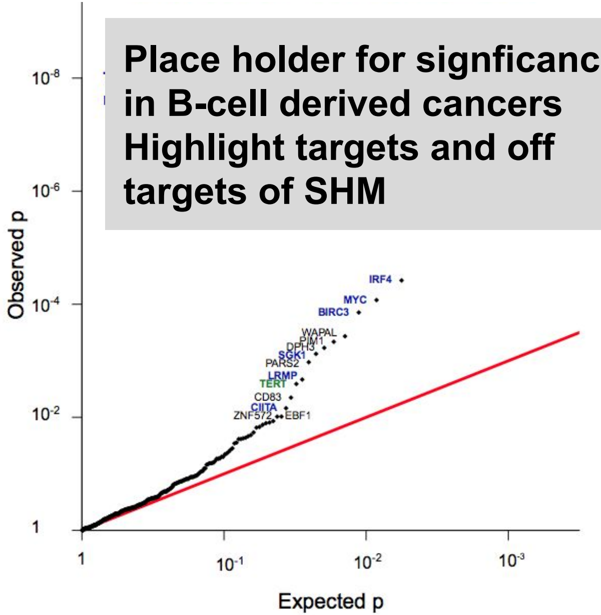
Slide prepared by Jakob Skou Pedersen

Supp Fig 1B,C: Significance analysis and effect on expression

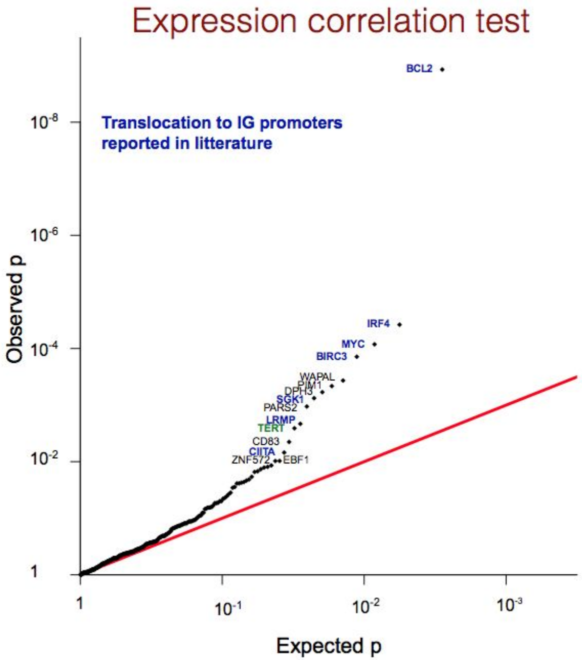
○ Normal somatic mutation

⊗ Somatic hyper mutation

Signficance analysis in B-cell derived cells highlighting targets and off-targets of SHM, results w and w/o using AID signature mutations.



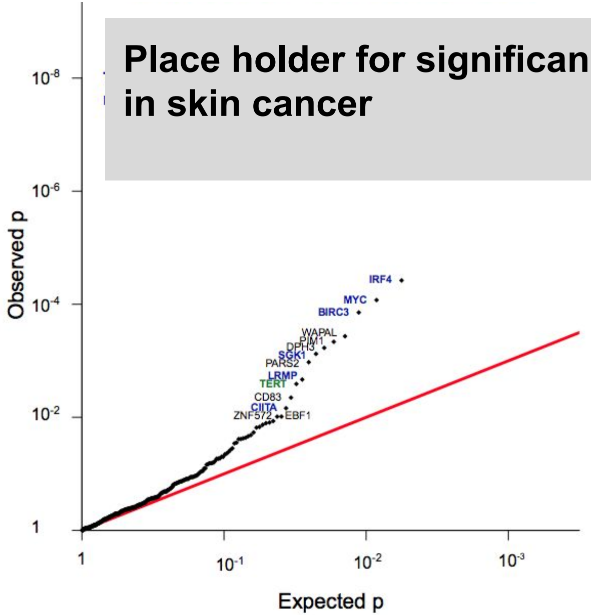
Place holder for signficance in B-cell derived cancers Highlight targets and off targets of SHM


Expression correlation test

- Some cancer genes use immunoglobulin (Ig) promoters after translocation
- TSS region of Ig genes somatically hyper mutated in lymphomas
- Mutations in these genes may therefore not be causative

Supp. Fig 2: Effect of UV, promoter hotspots in skin cancer

Supp Fig 2A



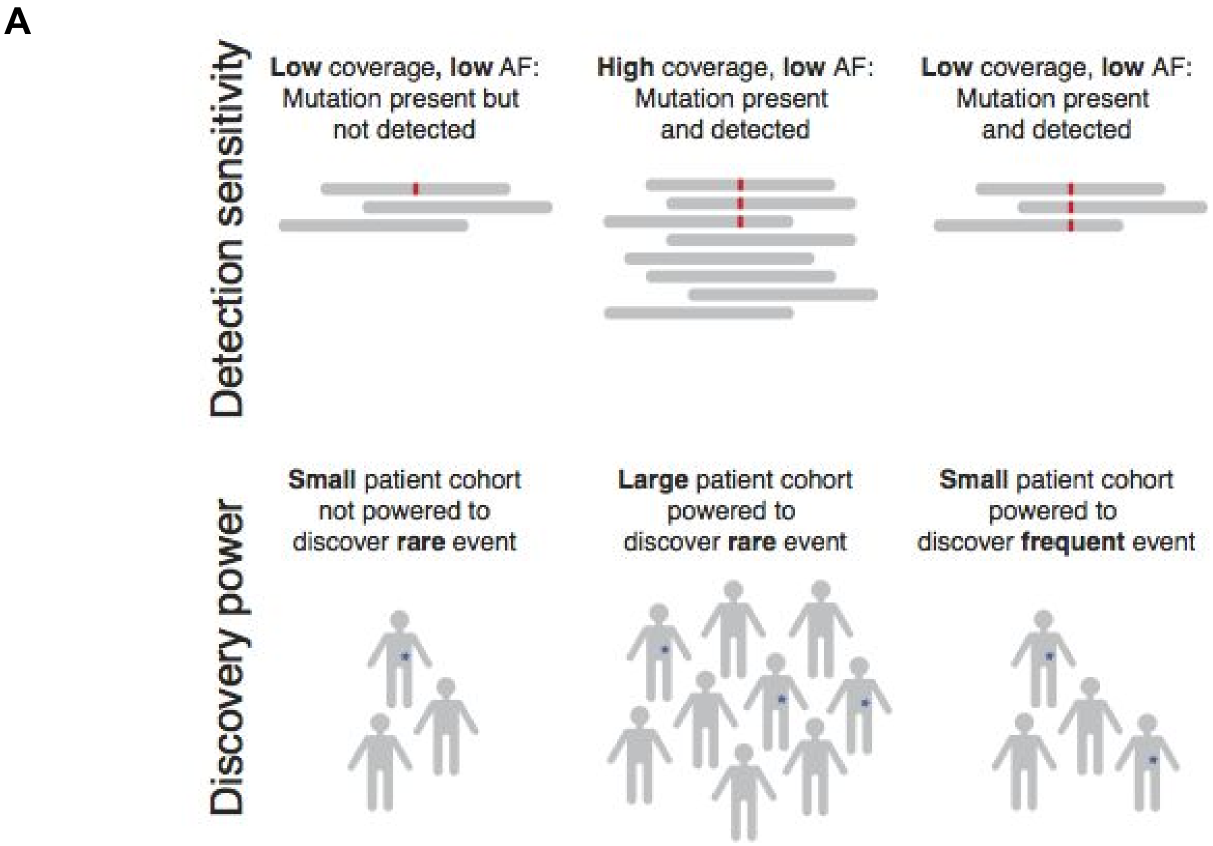
Place holder for significance in skin cancer

Supp Fig 2B

Place holder for analysis of hotspot mutations in melanoma

Esther Rheinbay, Gad Getz

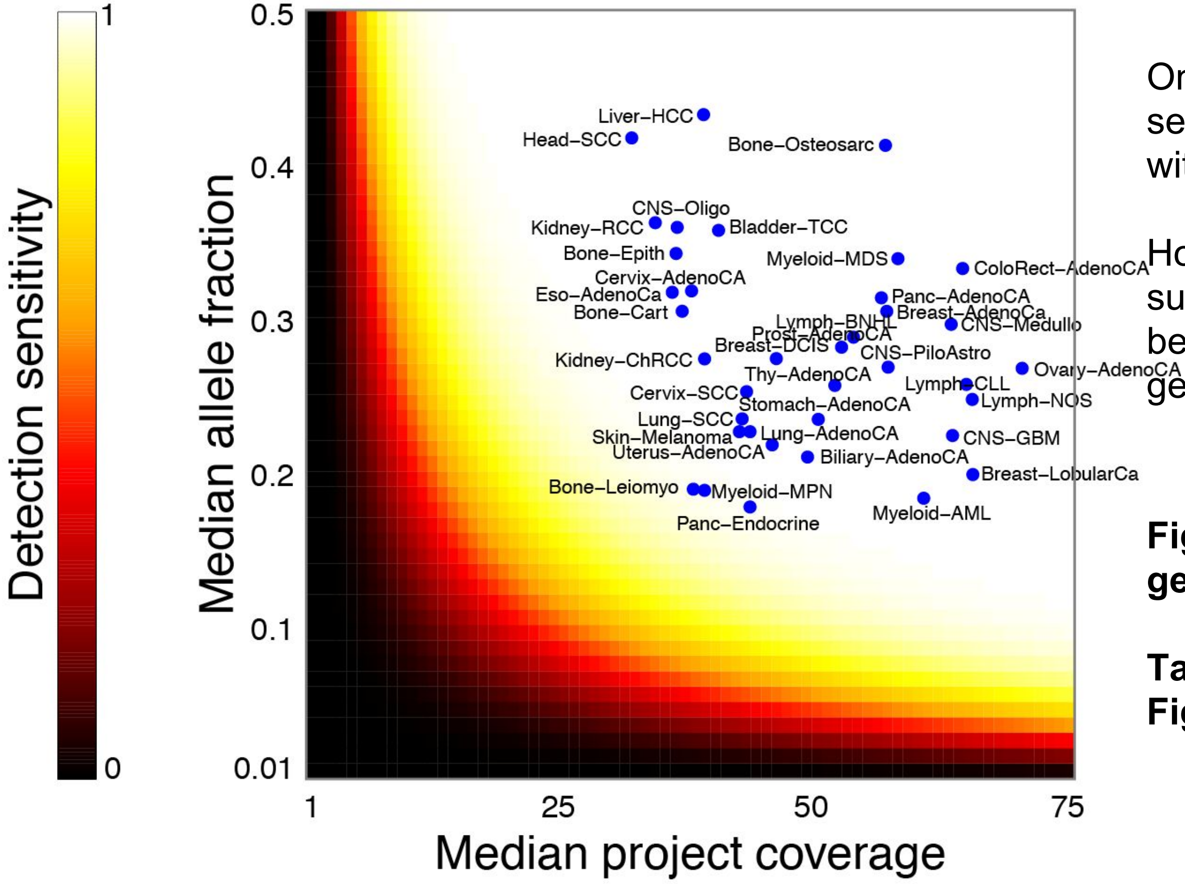
Fig 5A: power analysis


A
Detection sensitivity
Low coverage, low AF: Mutation present but not detected
High coverage, low AF: Mutation present and detected
Low coverage, low AF: Mutation present and detected
Discovery power
Small patient cohort not powered to discover rare event
Large patient cohort powered to discover rare event
Small patient cohort powered to discover frequent event

Esther Rheinbay, Gad Getz

Fig 5B,C: power analysis

Fig 5B



On average, detection sensitivity is sufficient (>99%) within all tumor types

However, there may be substantial differences between patients and different genomic regions!

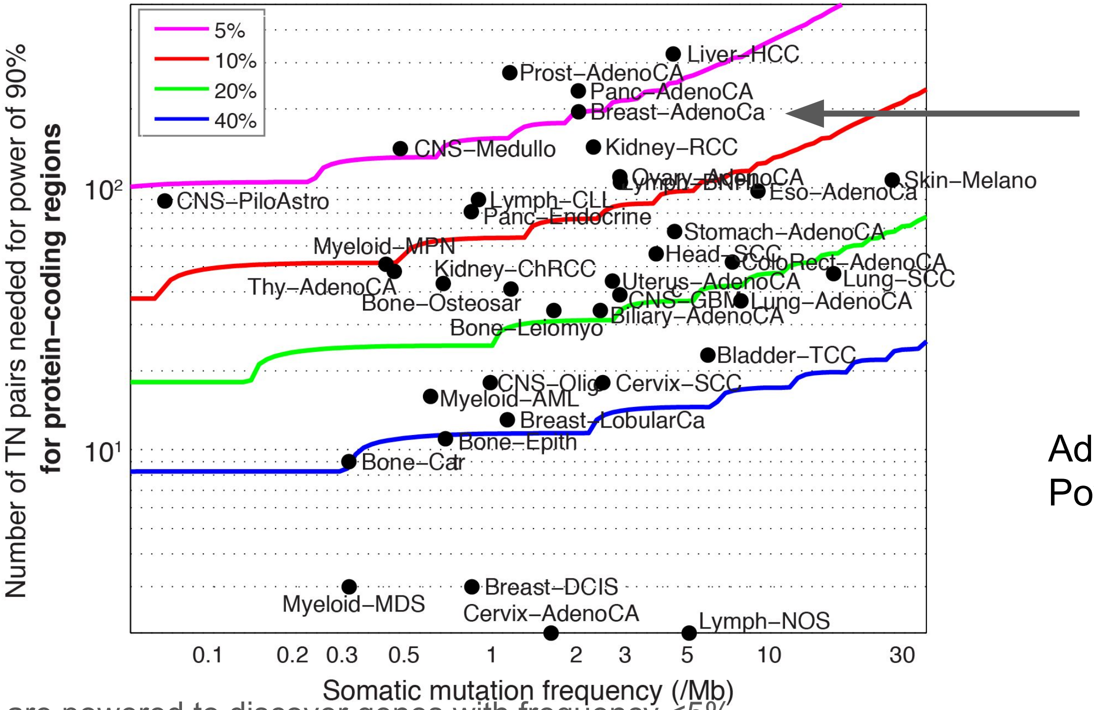
Fig 5C Regions of known cancer genes that are not covered well

Table
Figures of coverage in Supp Fig 3

Esther Rheinbay, Gad Getz

Fig 5D: Discovery power analysis on PCAWG tumor types

Fig 5D



Breast-AdenoCa Cohort powered to Identify protein-coding genes present in 5% of patients

Add in Supp Figures Power for other elements

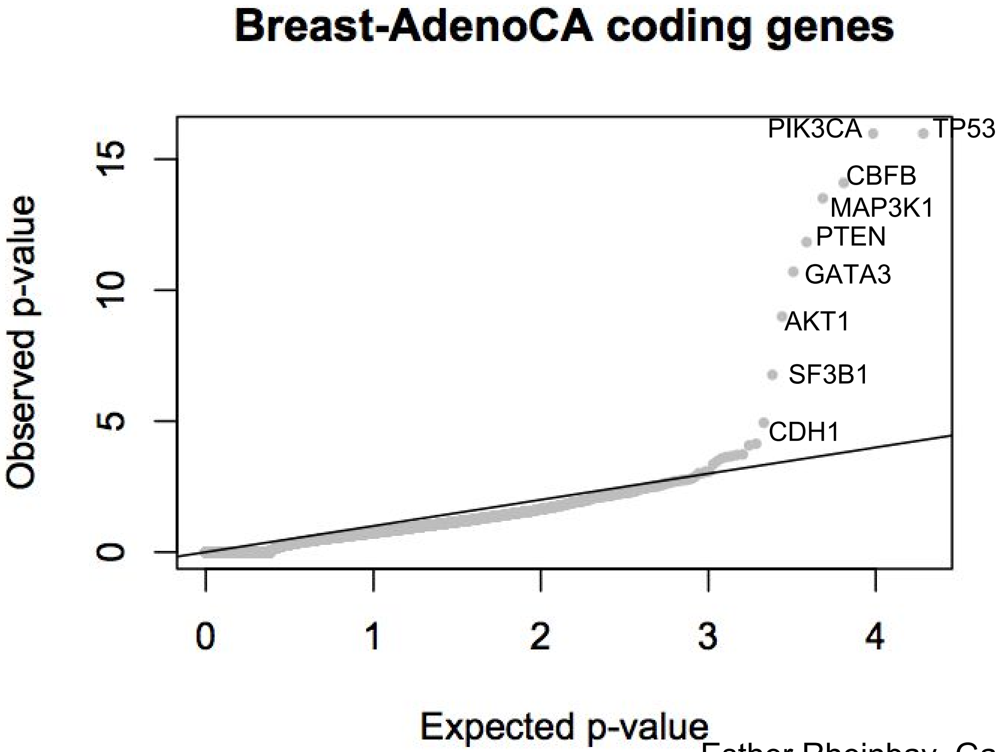
- Very few cohorts are powered to discover genes with frequency <5%
- Most cohorts are powered to find genes >40%; but we should expect to find few genes in bladder cancer, oligodendroglioma, AML
- No major differences in discovery power between element lists (promoters, UTRs, enhancers, lincRNAs)

Esther Rheinbay, Gad Getz

Supp Fig 4: What does the 5% discovery power threshold mean?

BRCA significantly mutated genes with frequency ≥2% from Lawrence et al, 2014

PIK3CA (32.6%)
TP53 (31.5%)
GATA3 (9.4%)
MAP3K1 (6.6%)
MLL3 (6.6%)
CDH1 (6.4%)
NCOR1 (3.8%)
MAP2K4 (3.7%)
PTEN (3.5%)
RUNX1 (2.8%)
PIK3R1 (2.5%)
CTCF (2.2%)
AKT1 (2.1%)
CBFB (2.0%)
SPEN (2.0%)
TBX3 (2.0%)



Esther Rheinbay, Gad Getz