nature genetics

ANALYSIS

Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations

Carlos L Araya[1,4], Can Cenik[1,4], Jason A Reuter[1], Gert Kiss[2], Vijay S Pande[2], Michael P Snyder[1] & William J Greenleaf[1,3]

Journal Club

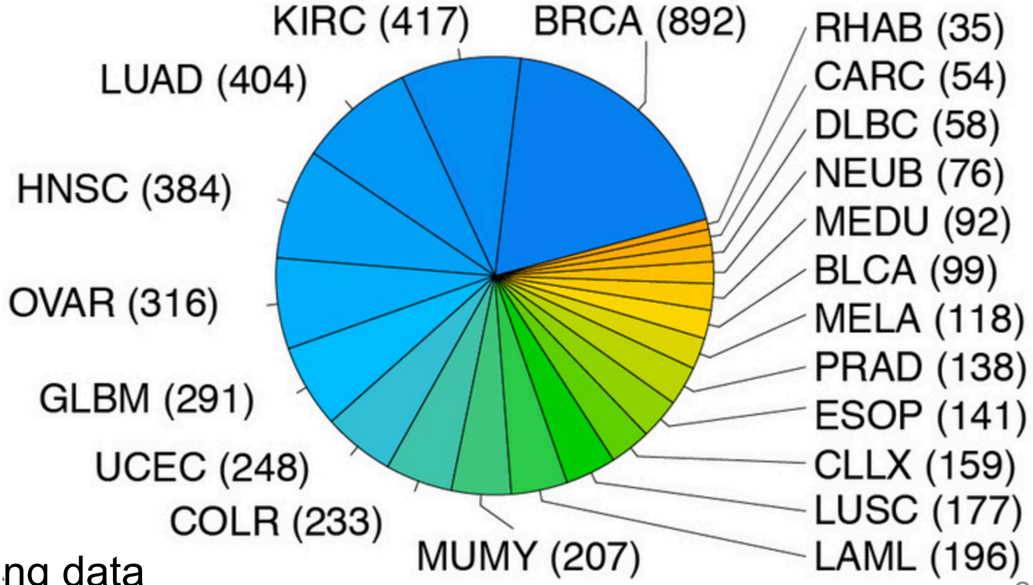
Xiaotong Li

April 26, 2016

NATURE GENETICS VOLUME 48 | NUMBER 2 | FEBRUARY 2016

Background: data summary

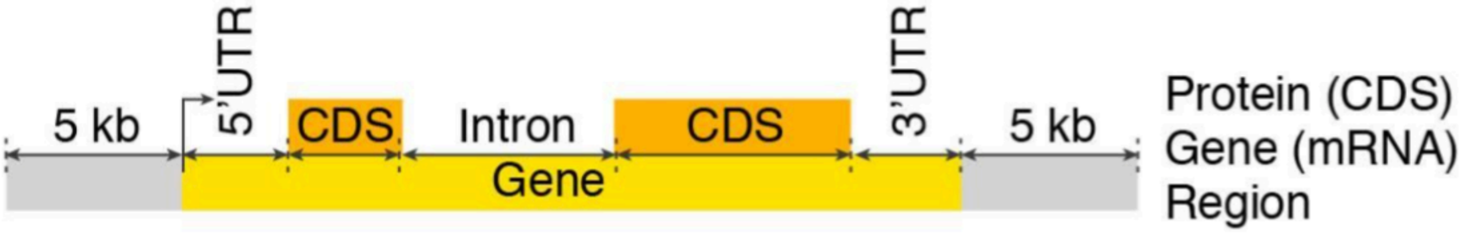
- **Whole-exome sequencing**
  - ✓3,185,590 somatic variant calls
  - ✓from 21 cancer types
- **Whole-genome sequencing**
  - ✓11,461,951 somatic SNV calls
  - ✓23 cancer types



Supply Fig.1a Summary of exome sequencing data

Method: I. uniform variant annotation

- Applied **snpEff** to annotate SNVs (exome & whole genome)
  - ✓ impact in protein-coding regions
  - ✓ impact in transcribed regions
    - ➢ coding, noncoding exons, introns, 5′ UTRs and 3′ UTR
  - ✓ impact in gene-associated regions
    - ➢ transcribed 5 kb upstream and 5 kb downstream
  - ✓ standardize gene name assignments.



Supply Fig.1b Reference coordinates for mutation impact annotation

Result: I. uniform variant annotation

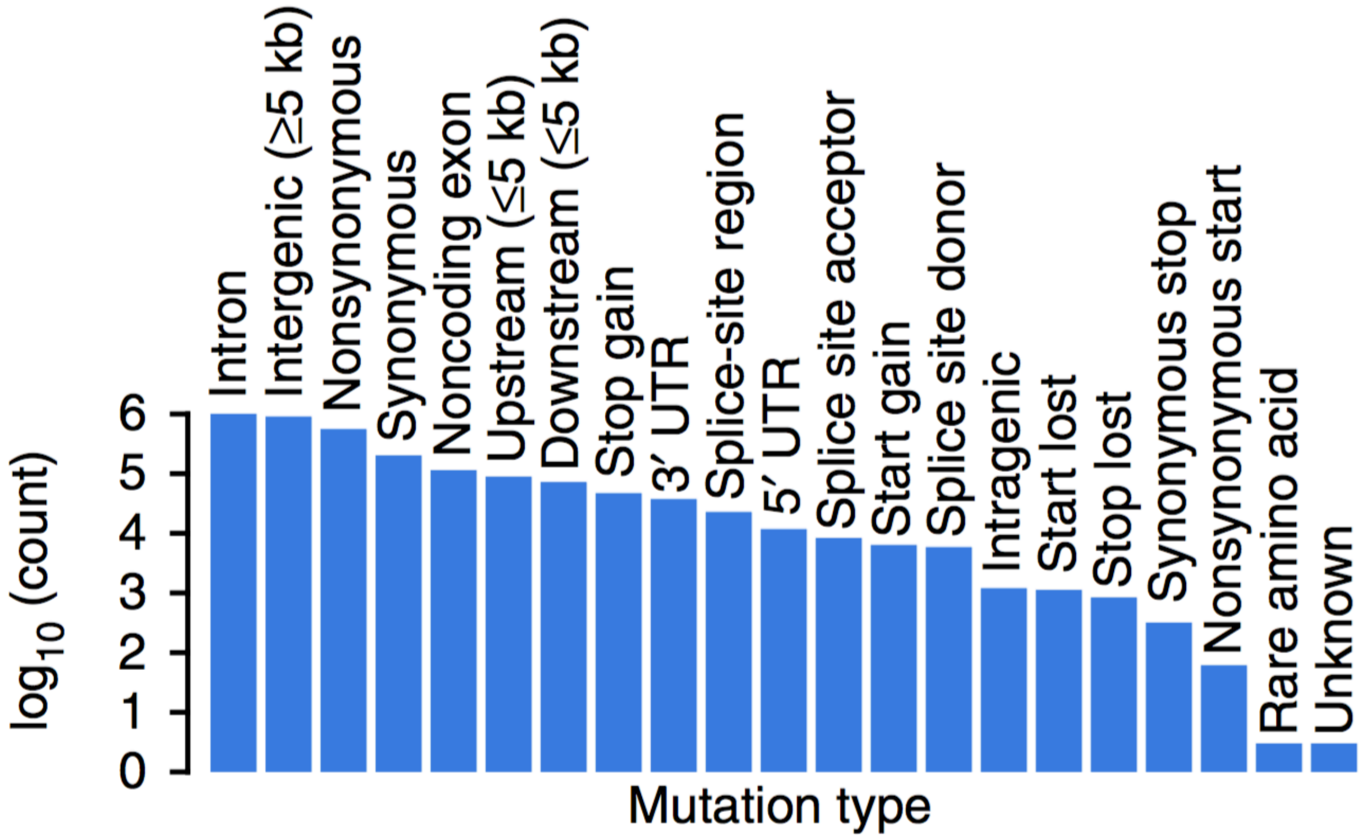
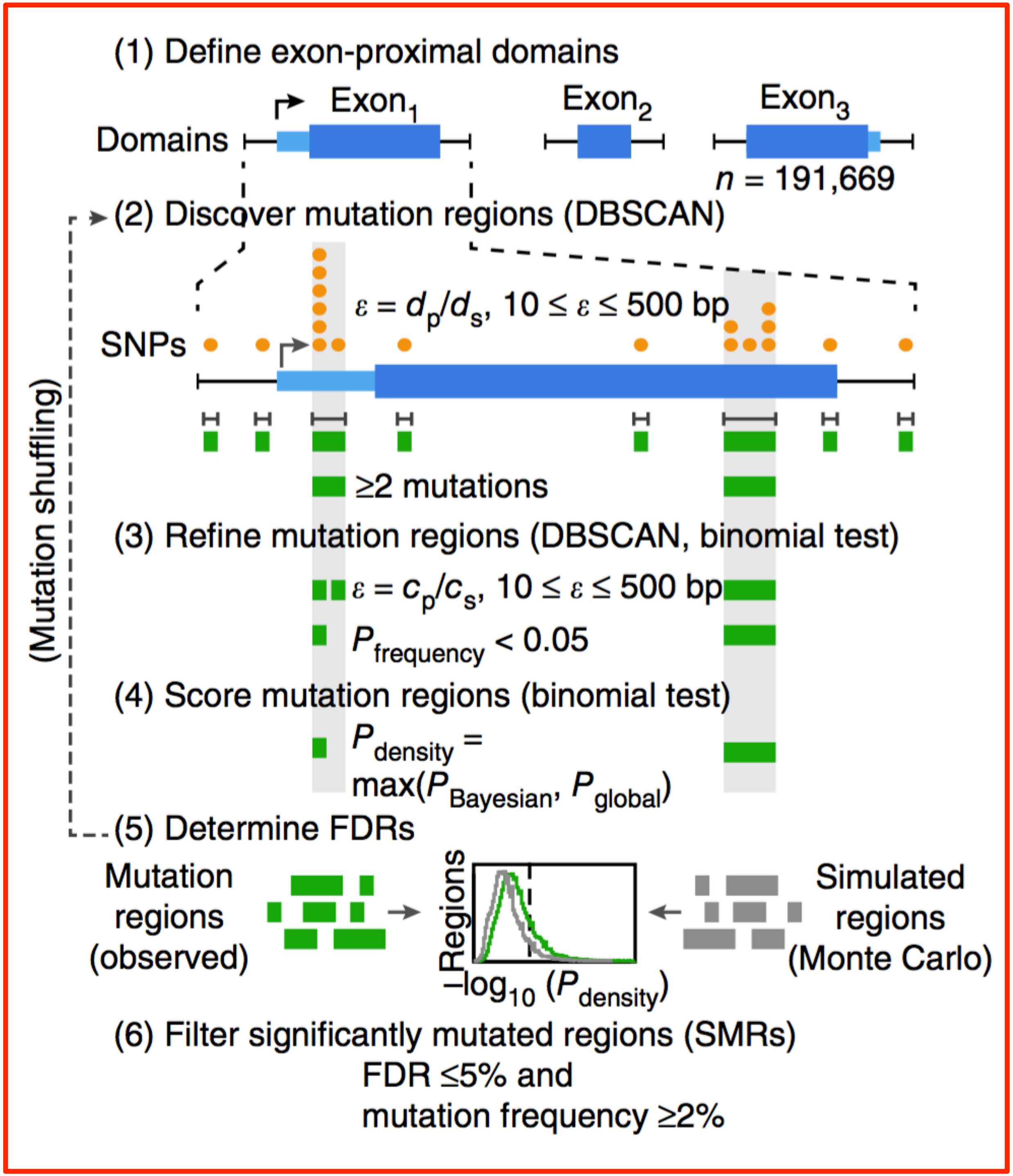


Fig. 1a Pan-cancer distribution of mutation types for *n* = 3,078,482 somatic SNV calls.

Method: II. procedure in calling SMRs

- SMRs: **significantly mutated regions**
- Mutation probability models
  - ✓ **whole-exome sequencing-derived**
    - ➢ **‘exonic’ mutation probability**: frequency of transitions and transversions within the mappable (100-bp), exonic regions
    - ➢ refined by expression levels, replication timing and GC content
    - ➢ **’matched’ mutation probability**: averaged the ‘exonic’ mutation probability per transition/transversion
    - ➢ **’global’ mutation probability**: average probability of transitions and transversions across all genes per tumor type
  - ✓ **whole-genome sequencing-derived**
    - ➢ ”**Bayesian” mutation probability**: binomial distribution



(1) Define exon-proximal domains
Exon$_1$
Exon$_2$
Exon$_3$
Domains
n = 191,669
(2) Discover mutation regions (DBSCAN)
$\varepsilon = d_p/d_s$, $10 \le \varepsilon \le 500$ bp
SNPs
≥2 mutations
(3) Refine mutation regions (DBSCAN, binomial test)
$\varepsilon = c_p/c_s$, $10 \le \varepsilon \le 500$ bp
$P_{frequency} < 0.05$
(4) Score mutation regions (binomial test)
$P_{density}$ = max($P_{Bayesian}$, P_{global})
(5) Determine FDRs
(Mutation shuffling)
Mutation regions (observed)
Regions
$-\log_{10}$ ($P_{density}$)
Simulated regions (Monte Carlo)
(6) Filter significantly mutated regions (SMRs)
FDR ≤5% and
mutation frequency ≥2%

exon-proximal domains: **±1,000 bp**

DBSCAN: Density-based spatial clustering of applications with noise

Distance parameter **ε** is dynamically defined as the average distance of mutated positions (**d_p**) in the domain size (**d_s**)

Subclusters: higher mutation densities ($P < 0.05$, binomial test)

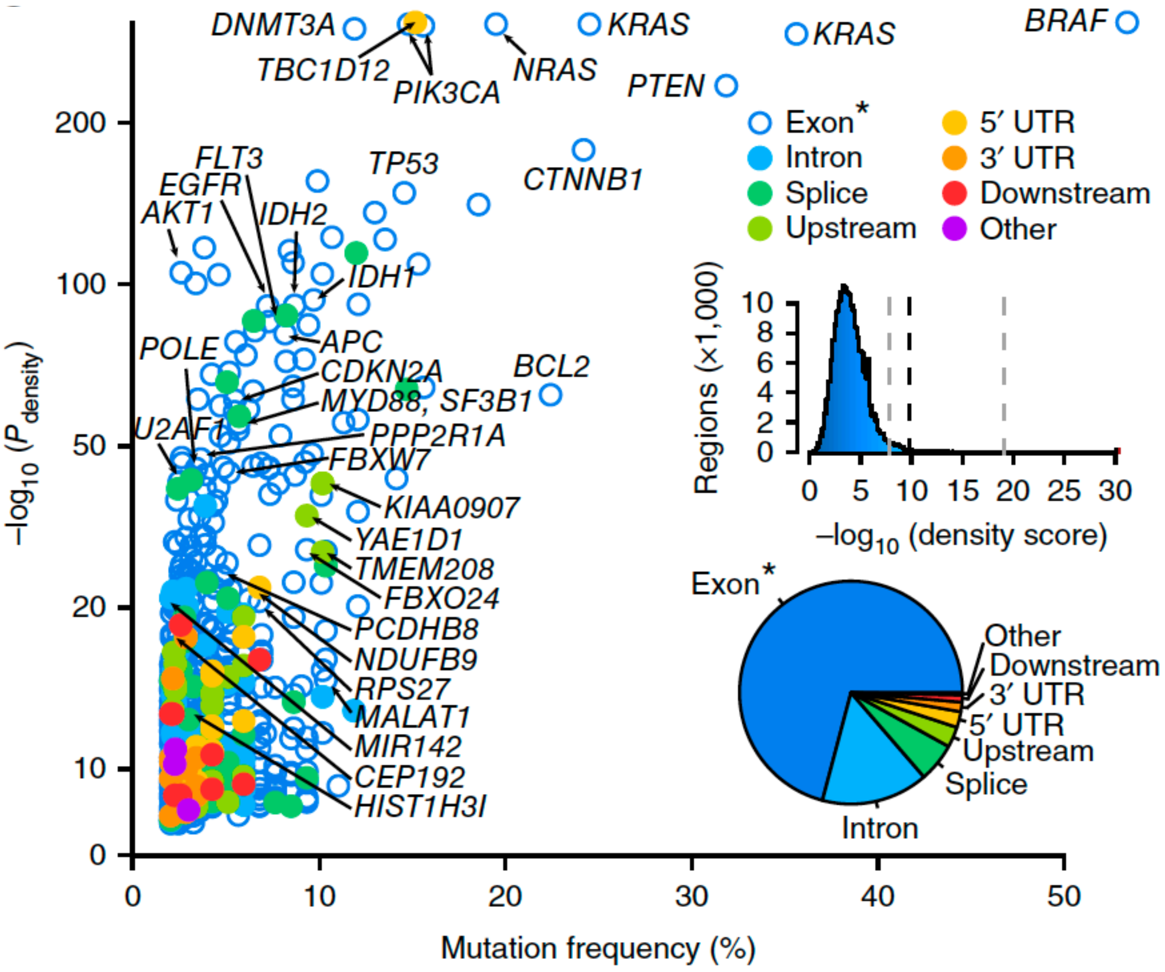
Select the most conservative density scores

Empirical FDRs calculated from simulations

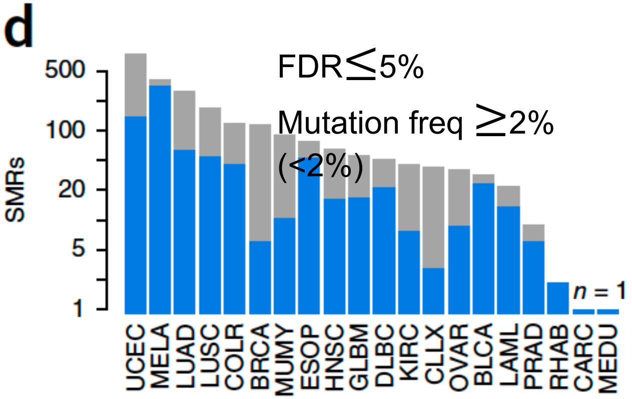
Computed the density score ($P_{density}$) threshold that guarantees FDR ≤5%

Output: **872** SMRs, from 735 unique genomic regions, in 20 distinct cancer types.

Mutation frequency and density scores for the SMRs discovered



- color-coded by type
- labeled by associated gene
- Top: distribution of density scores in evaluated regions
- Bottom: distribution of SMR region types
- Dashed lines: the minimum, median and maximum density score FDR (5%) thresholds.
- “Exon*”: coding exons & noncoding genes



FDR≤5%

Mutation freq ≥2%

(<2%)

Fig.1d Number of SMRs in each cancer type

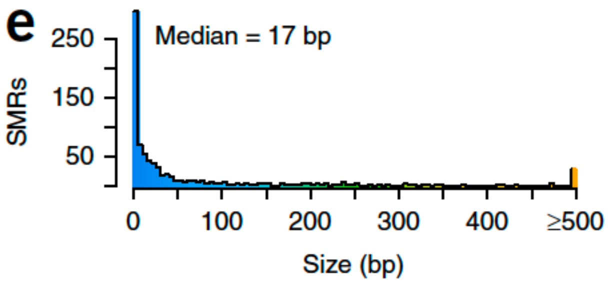


Fig.1e SMR size distribution

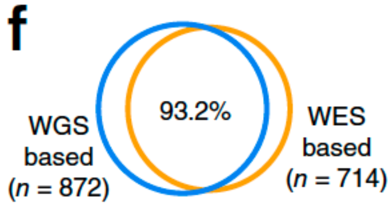


Fig.1f Concordance of SMRs

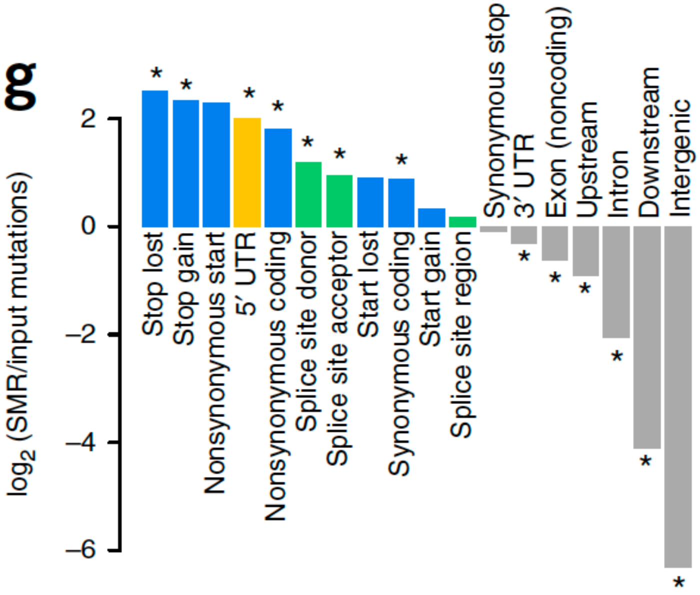


Fig.1g Categories with significant fold change between SMR-associated and input mutations (*P < 0.01)

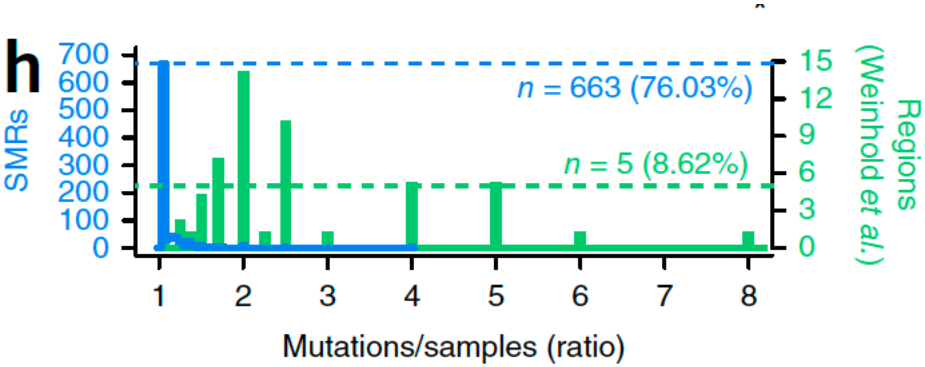
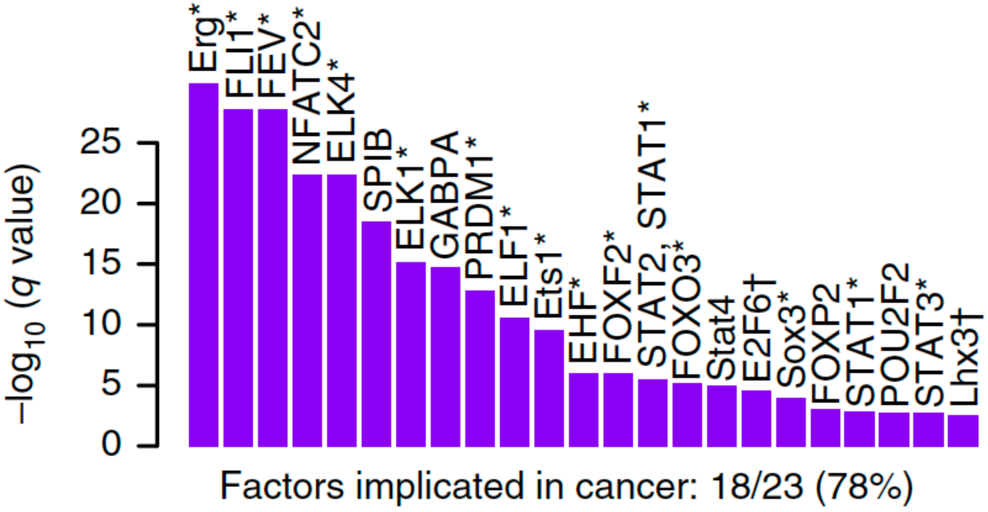


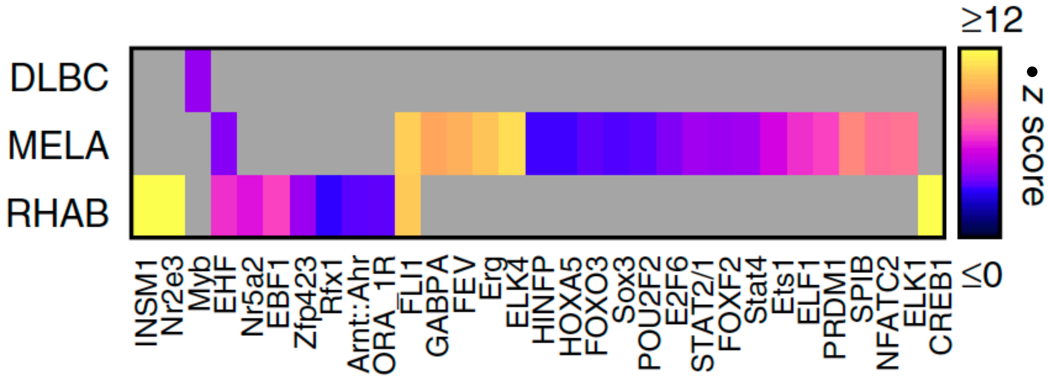
Fig.1h Distribution of number of mutations per sample in SMRs and 58 recurrently altered noncoding regions

Horizontal lines: the number of regions where mutations derive from distinct samples

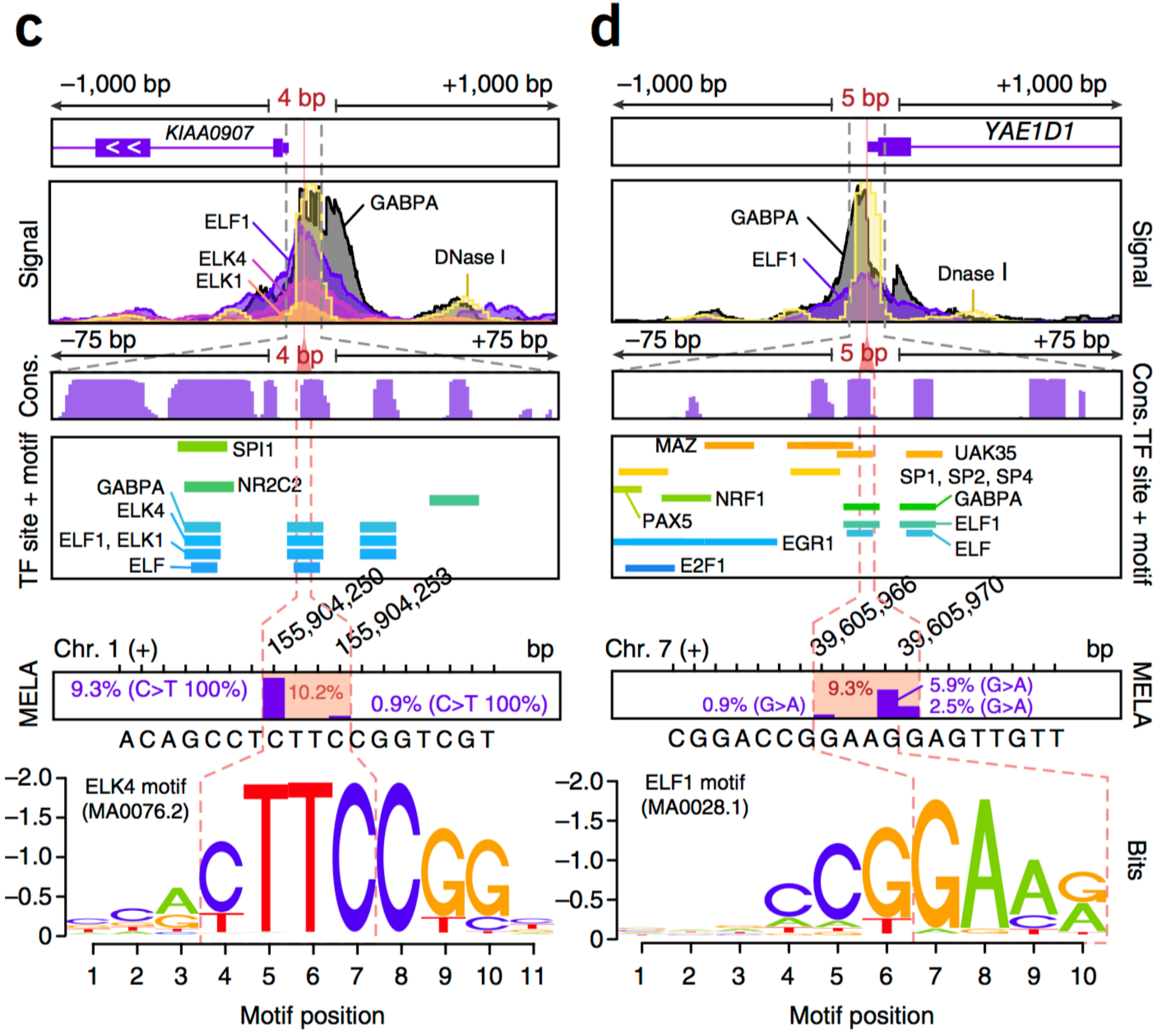
Noncoding SMRs recurrently alter promoters and 5' UTRs



- Transcription factors with enriched (q<0.01) motifs in small SMRs (<=25 bp)
  - ✓ 18/23 TFs: cancer or cell cycle control associated, developmental



- Cancer-specific motif enrichment
  - ✓ DLBC: diffuse large B cell lymphoma
  - ✓ MELA: melanoma
  - ✓ RHAB: rhabdoid tumor



Melanoma SMRs in *KIAA0907* (**c**) and *YAE1D1* (**d**) **promoter** regions

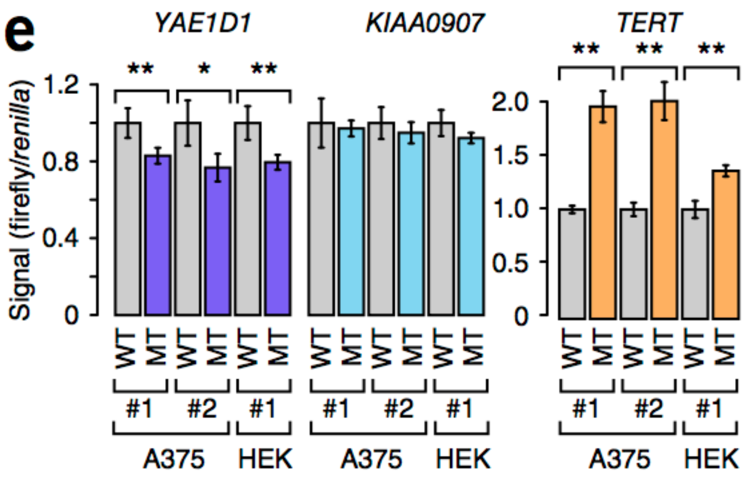


Fig.2e Luciferase reporter signal from wild-type (WT) and mutant (MT) promoters

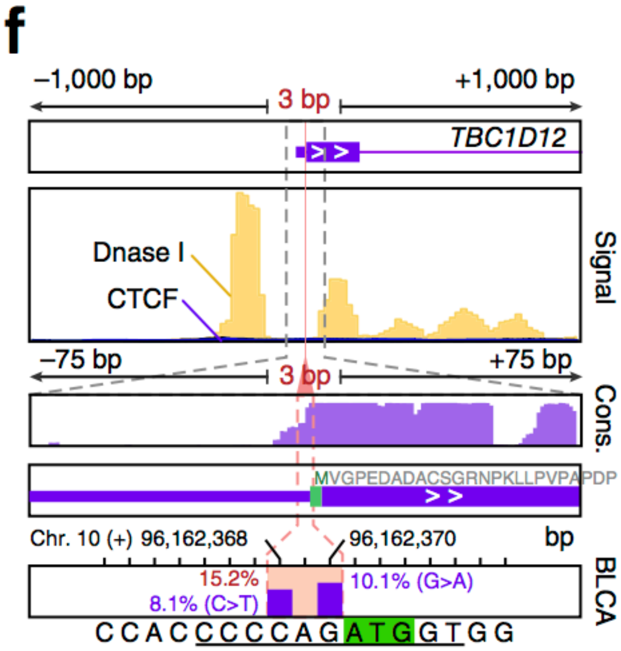


Fig.2f Bladder cancer SMR in the 5′ UTR of TBC1D12

- YAE1D1 promoter mutations reduced reporter gene expression
- no detectable changes in reporter gene expression with the mutant KIAA0907 promoter

- Bladder tumors with mutations in this SMR displayed altered p90RSK phosphorylation
  - a signal of increased cell cycle proliferation
- Altered α-tubulin levels

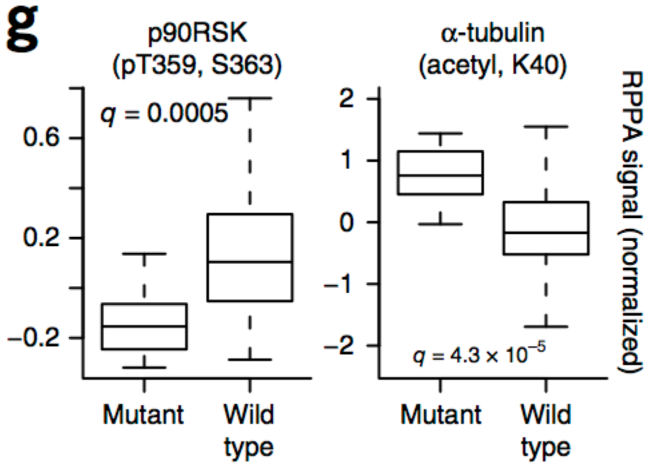
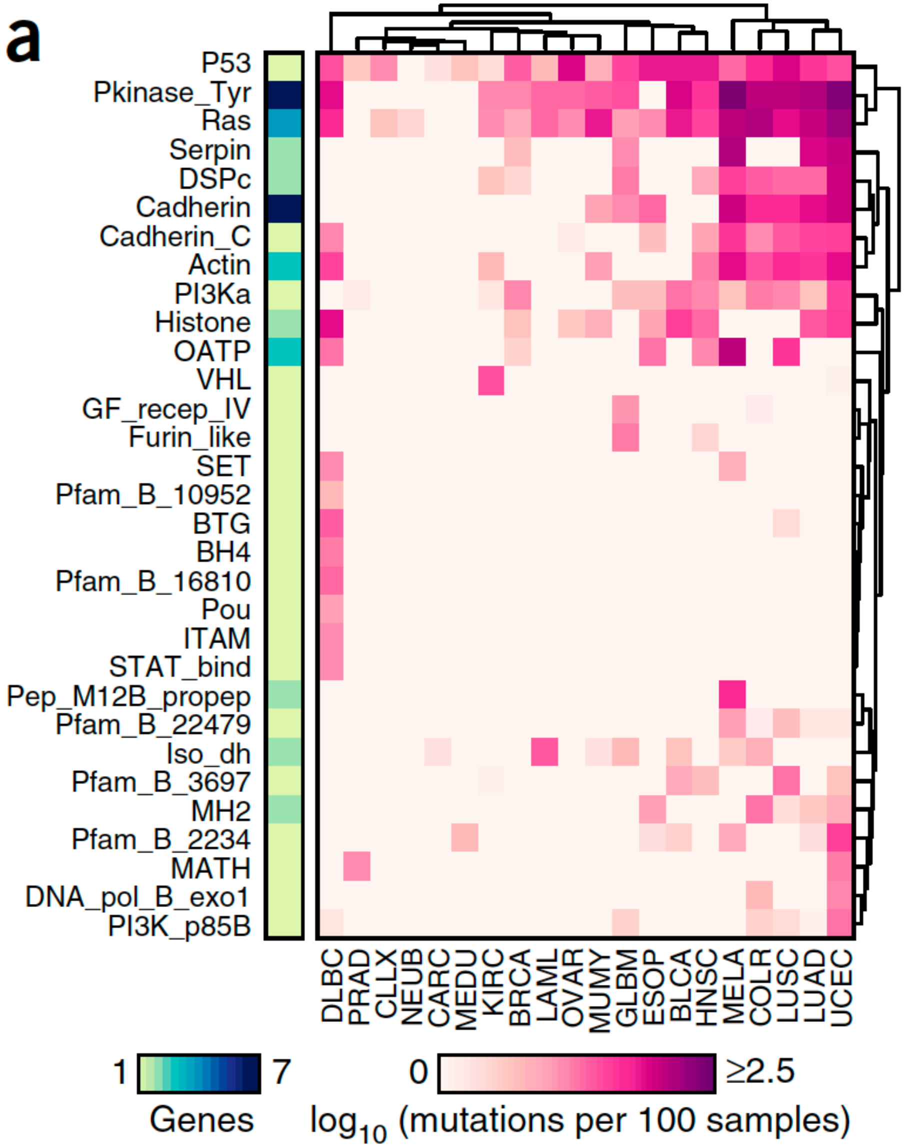


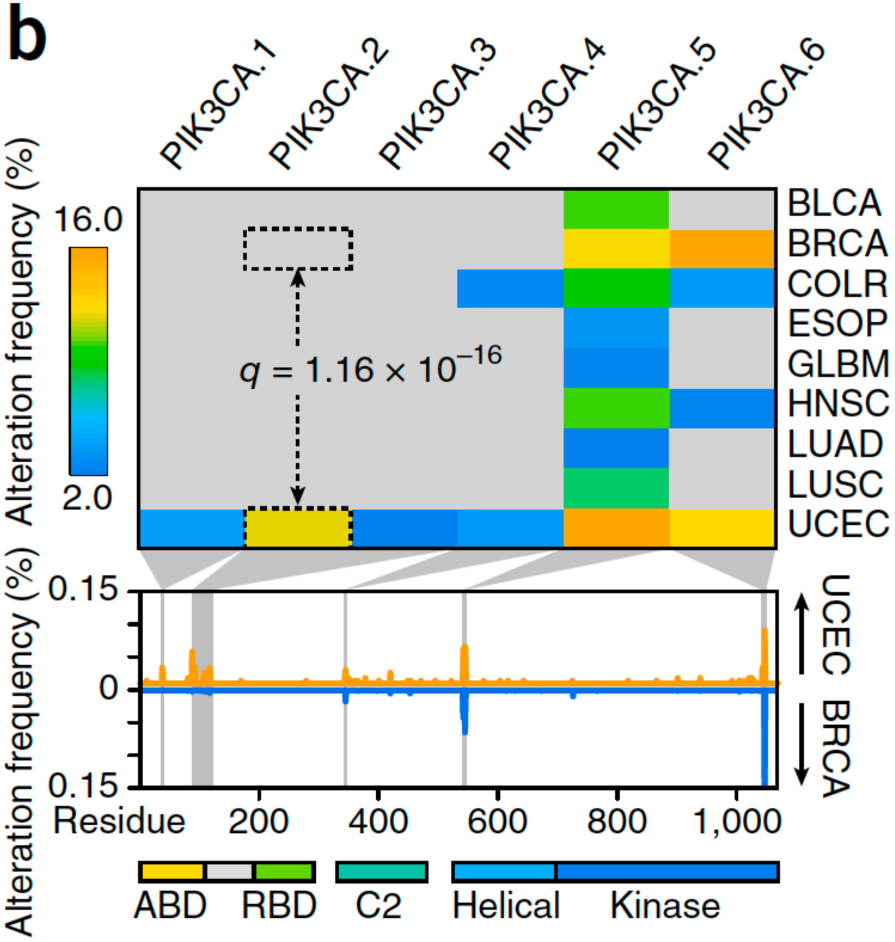
Fig.2g Relative protein and post translational modification signals

Structural mapping of SMRs onto proteins and complexes



- Nonsynonymous mutation frequency per PFAM protein domain per cancer, per residue
- Many protein domains showed high burdens of somatic alteration in multiple cancers
- Protein domains can show remarkable cancer type specificity in burdens of alteration
  - ✓ VHL in kidney clear cell carcinoma
  - ✓ SET in diffuse large B cell lymphoma

Alteration frequency matrix of PIK3CA SMRs



- Detected **six SMRs** in PIK3CA across **eight** cancer types
  - ✓ PIK3CA.1: Adaptor-binding domain (ABD)
  - ✓ PIK3CA.2 & .3: α-helix region between ABD and linker region between ABD and Ras-binding domain (RBD)
  - ✓ PIK3CA.4: C2
  - ✓ PIK3CA.5: helical domain
  - ✓ PIK3CA.6: kinase domain
- Significant differences in PIK3CA.2 alteration frequencies in endometrial and breast cancers
  - ✓ further validated in whole-genome sequences
  - ✓ differences in total *PIK3CA* mutation frequency between endometrial and breast cancers could, in part, be localized to this region

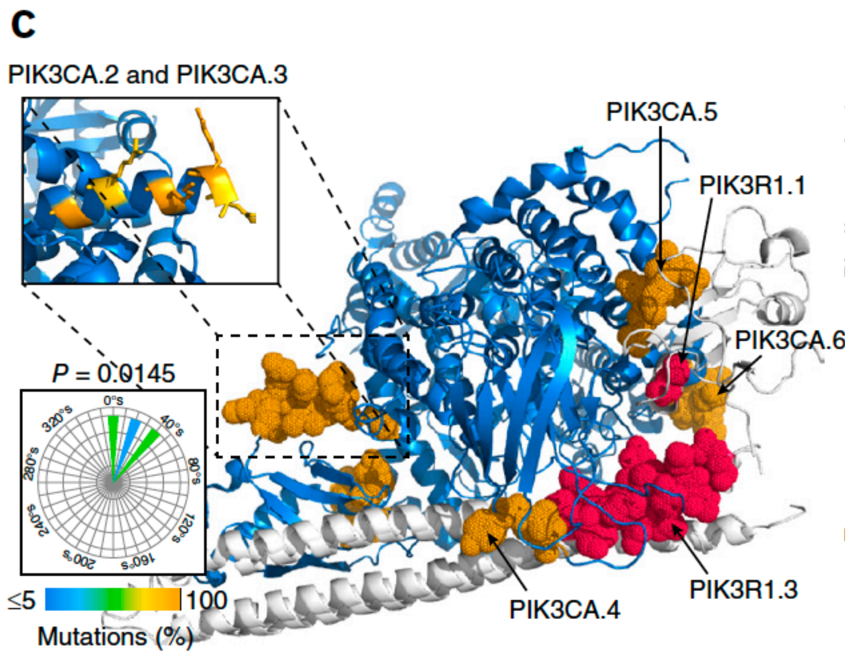


Fig.3c Co-crystal structure of PIK3CA and PIK3R1 interaction

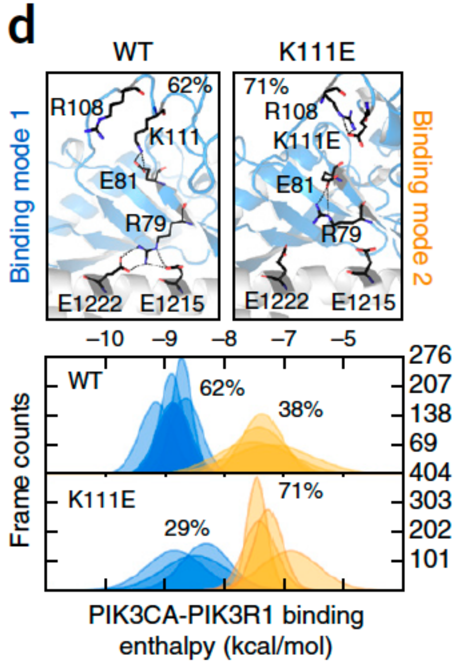


Fig. 3d Mutations within the PIK3CA.2, PIK3CA.3 SMR α-helix interfere with Arg79-binding

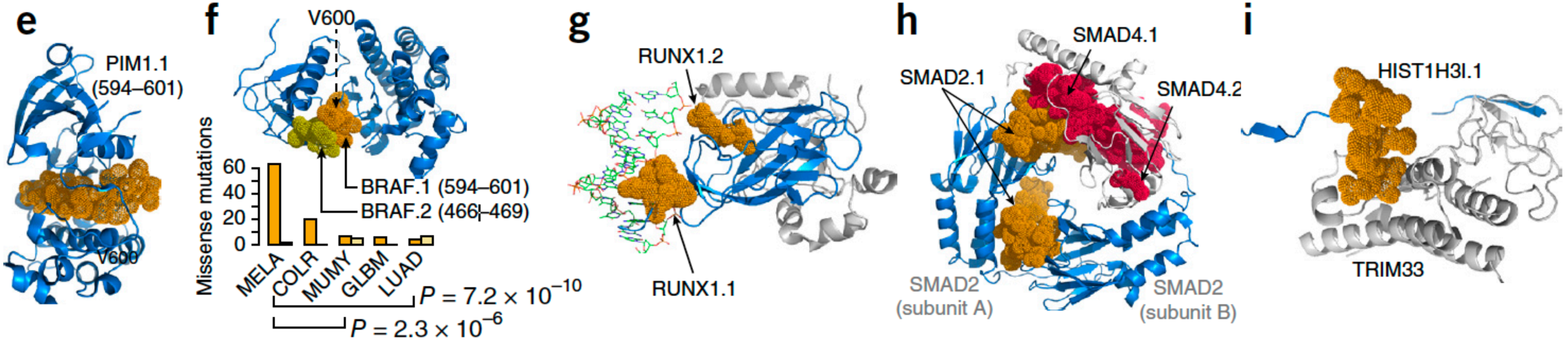
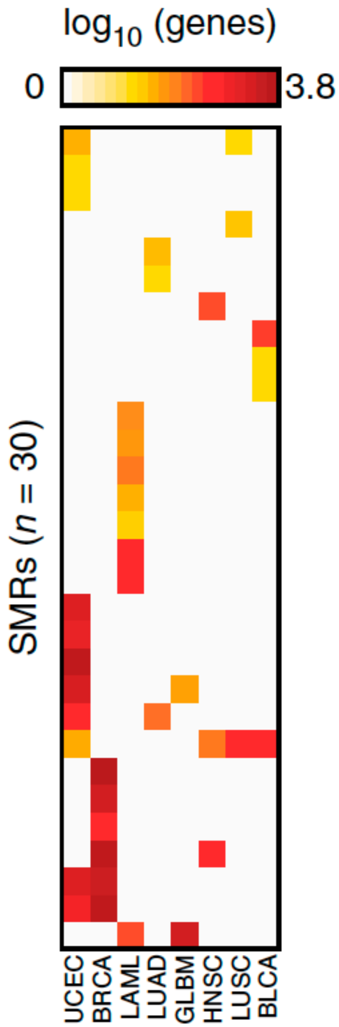
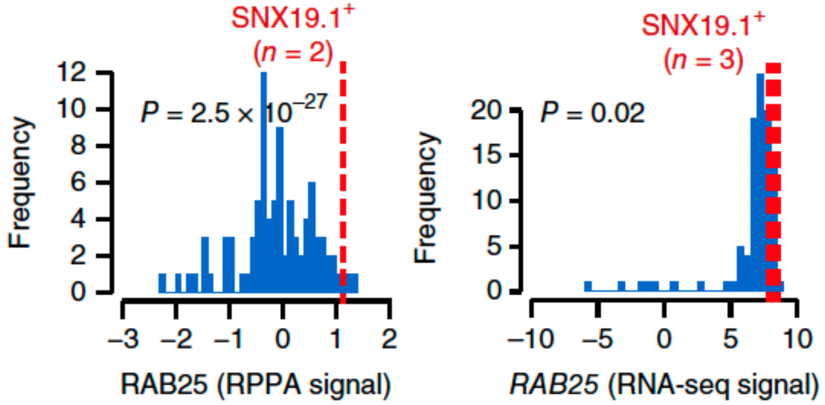


Fig.3e-i Molecular structures are shown spatially clustered alterations

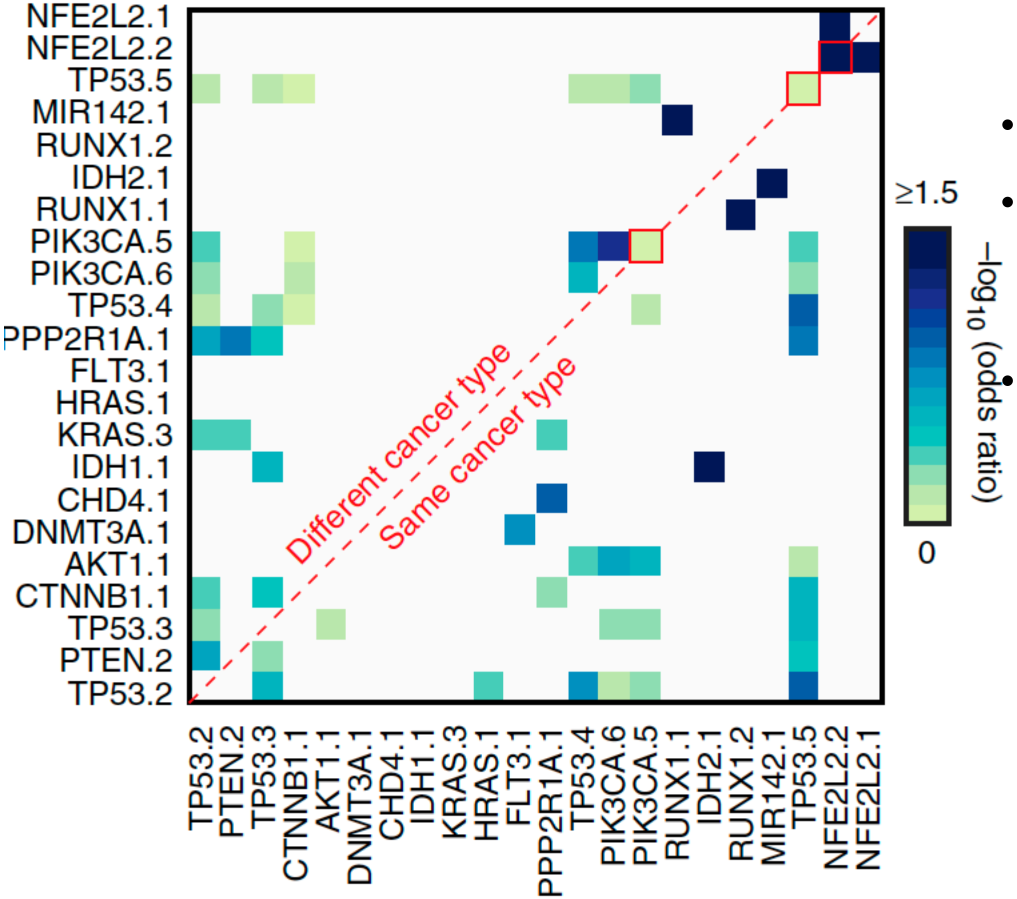
SMRs are associated with distinct molecular signatures



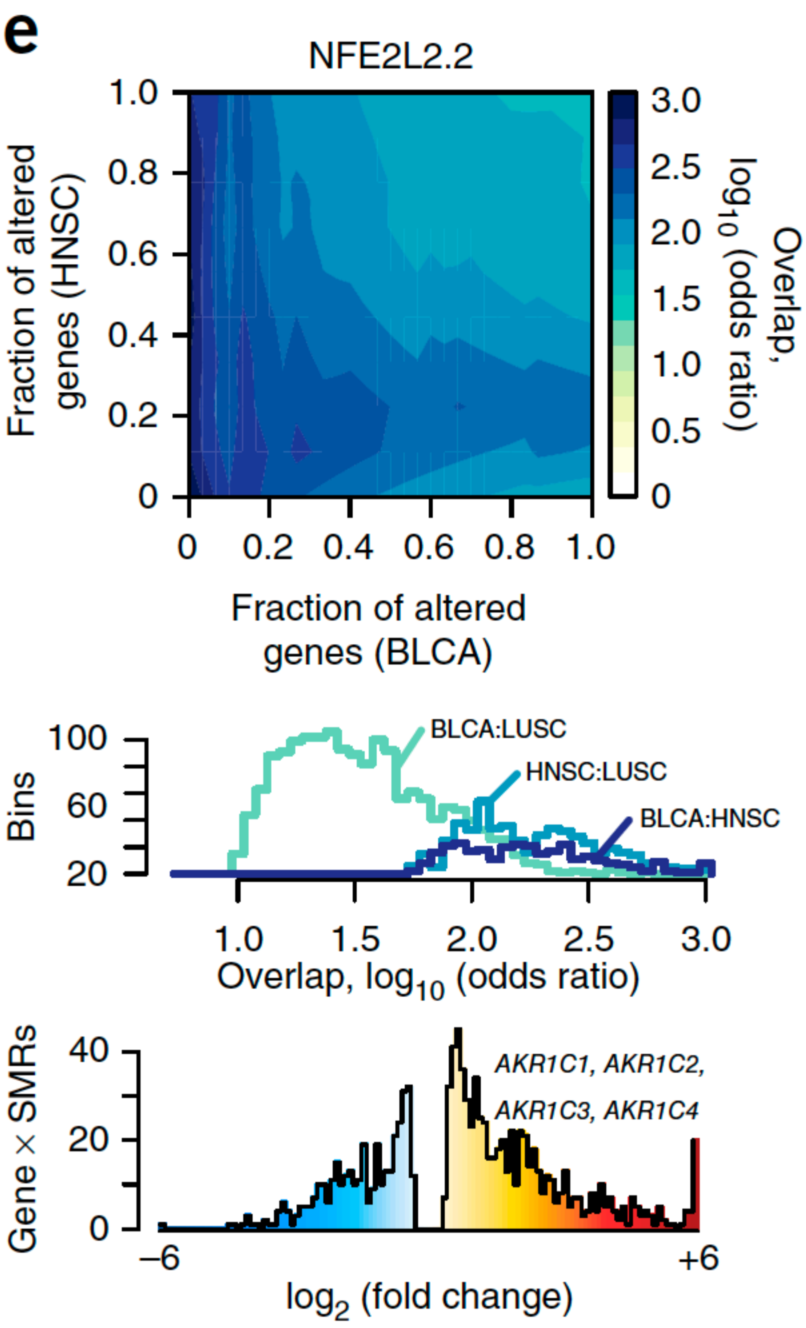


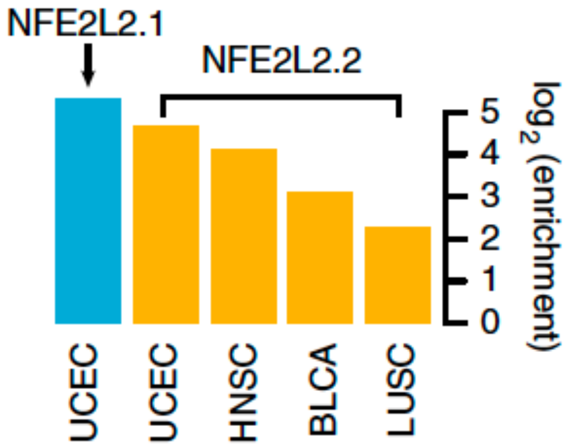
- Matched RNA-seq data: association between mutations in 30 SMRs with >=10 differentially expressed genes (FDR<5%)
  - ✓ highlight recurrent GSK3 pathway alterations in endometrial cancer
  - ✓ recurrent mTOR as well as EIF4 and epidermal growth factor (EGF) pathway alterations in glioblastoma
- Synonymous point mutations in a bladder cancer SMR in *SNX19* were associated with significant increases in the protein expression levels of RAB25
  - ✓ a RAS family GTPase that promotes ovarian and breast cancer progression
  - ✓ These increases are consistent with RNA expression differences in *RAB25*

Association of each SMR pair

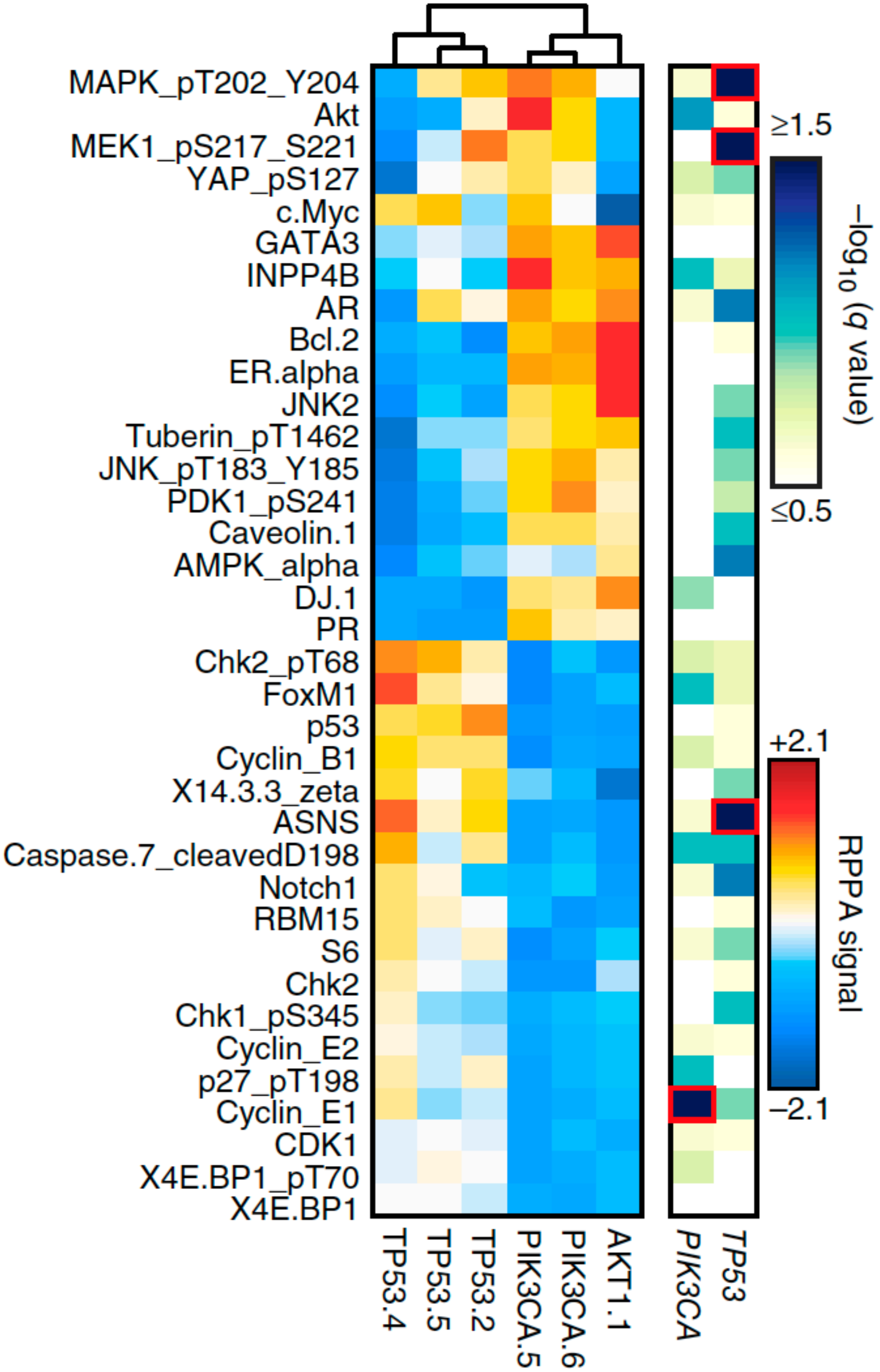


- 23 SMRs from 17 genes
- Similarity between differentially expressed gene sets associated with mutations in each SMR pair
- Concordant changes in gene expression for SMR pairs, suggesting potential functional relationships
  - ✓ Well-established relationship between PIK3CA and AKT1
  - ✓ mutations in the same SMR in different cancers can elicit similar molecular profiles in distinct cancers



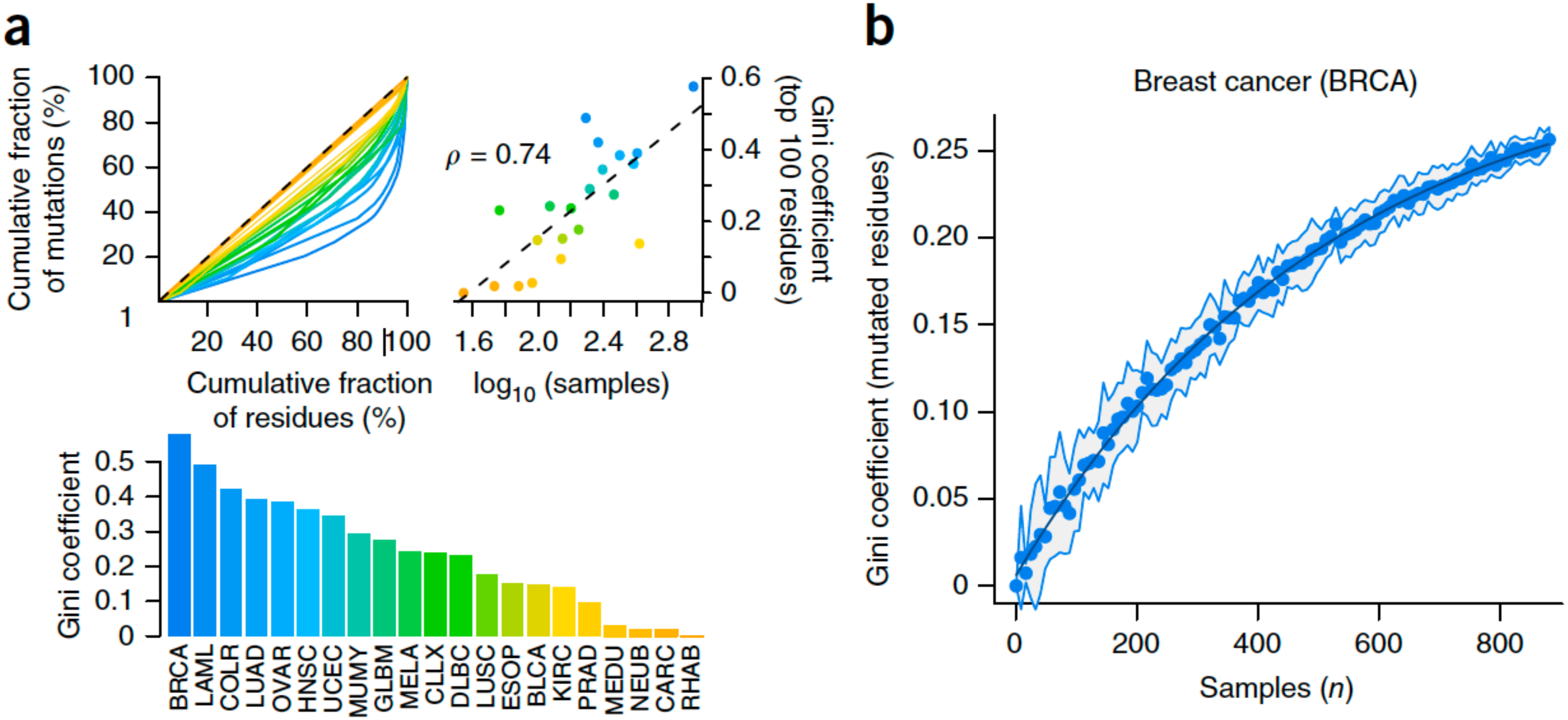


- The overlap between differentially expressed genes associated with alteration of the NFE2L2.2 SMR in bladder cancer and head and neck carcinoma
  - ✓ The distribution of odds ratios of similarity is summarized for three comparisons
  - ✓ Samples with NFE2L2.2 mutations exhibit highly increased expression of aldo-keto reductase enzymes
- Relative enrichment for oxidoreductase activity (GO:0016616) in specific cancer types
  - ✓ mutations in *NFE2L2* SMRs were highly enriched



- The patients with breast cancer were **grouped** by mutations in six SMRs in *PIK3CA*, *AKT1* and TP53
    - ✓ alterations in distinct SMRs within *TP53* were associated with highly similar changes in protein levels
- Differential expression between SMRs from *TP53* or *PIK3CA*
    - ✓ observed SMR-specific differences in ASNS levels and MAPK and MEK1 phosphorylation among samples with altered *TP53* SMRs
- Established differences in the molecular signatures associated with alterations of SMRs in the same gene

Structure in the distribution of cancer mutations remains largely uncharacterized



- Sought an alternative metric to assess structure in the distribution of the somatic coding mutations
  - measuring the Gini coefficient of amino acid substitutions per residue in each cancer type
  - Gini coefficients of dispersion were well correlated with sample numbers
- Subsampling demonstrated that, even with sample numbers >850, a large proportion of the structure of protein-altering mutations in breast cancer remains unseen