# Quantifying the Distribution of Functional Mutations

Nicholas Sinnott-Armstrong
(nasa@stanford.edu)
May 23, 2016
PCAWG 2, 5, 9, 14 Supergroup Call

# Rationale

- Tumorigenesis is a combination of many complex processes throughout a patient's life
- The tumor itself is a living, growing system
- Thus, observed mutations are a combination of predisposition to tumorigenesis (germline or pre neoplastic somatic), drivers of cancer, and passengers which don't progress disease
- Can we see any functional signatures which might “push tumors over the edge”?

# Previous work

- Just as in GWAS, somatic cancer risk could be partially attributable to small, genome wide shifts in activity of particular factors (Cowper-Sallari… Lupien 2012; Shaub… Snyder 2012; Maurano… Stamatoyannopoulos 2012; Quang… Collins 2015; Bu… Klocker 2015)

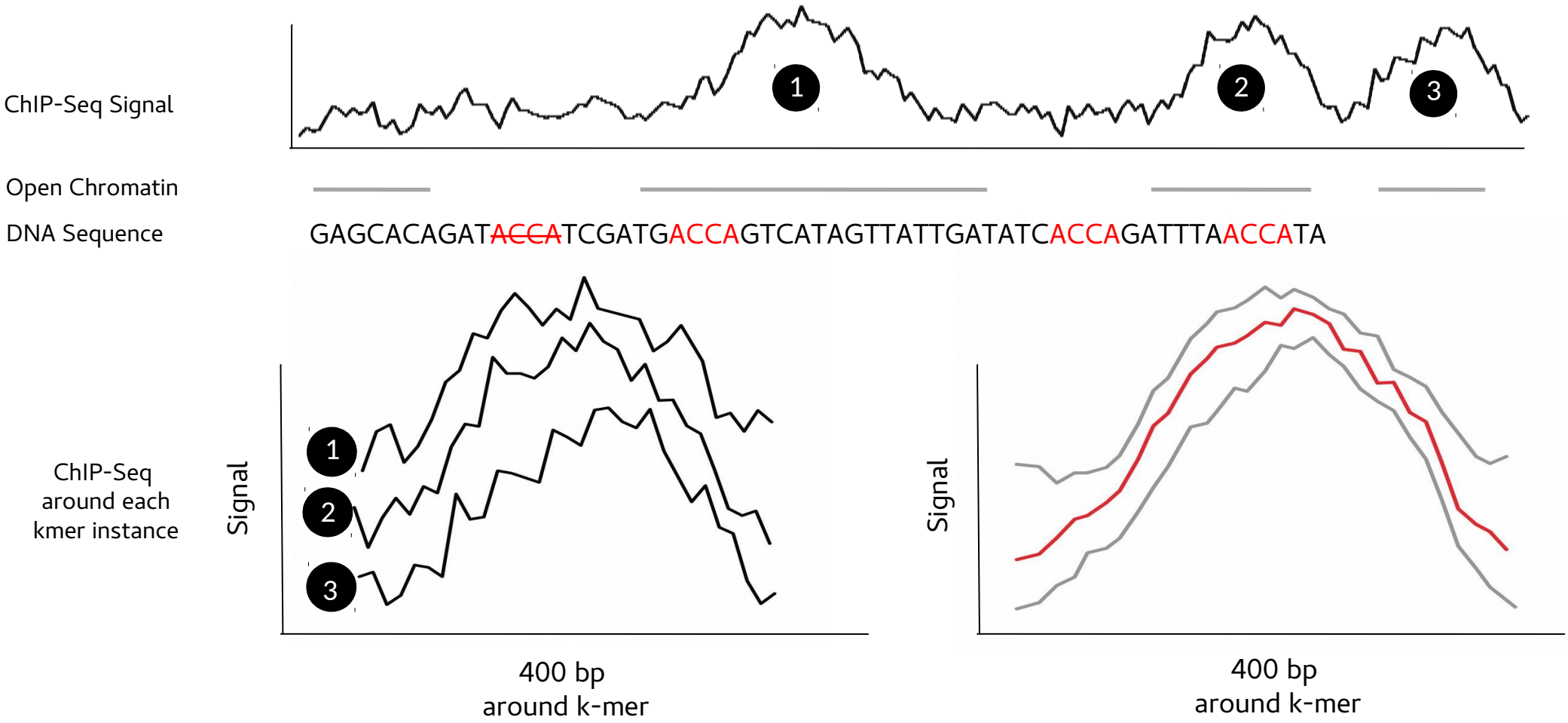
- CTCF/cohesin appears to have global enrichment in somatic TFBS alterations (Katainen… Aaltonen 2015; Flavahan… Bernstein 2016; Hnisz… Young 2016)

# Questions

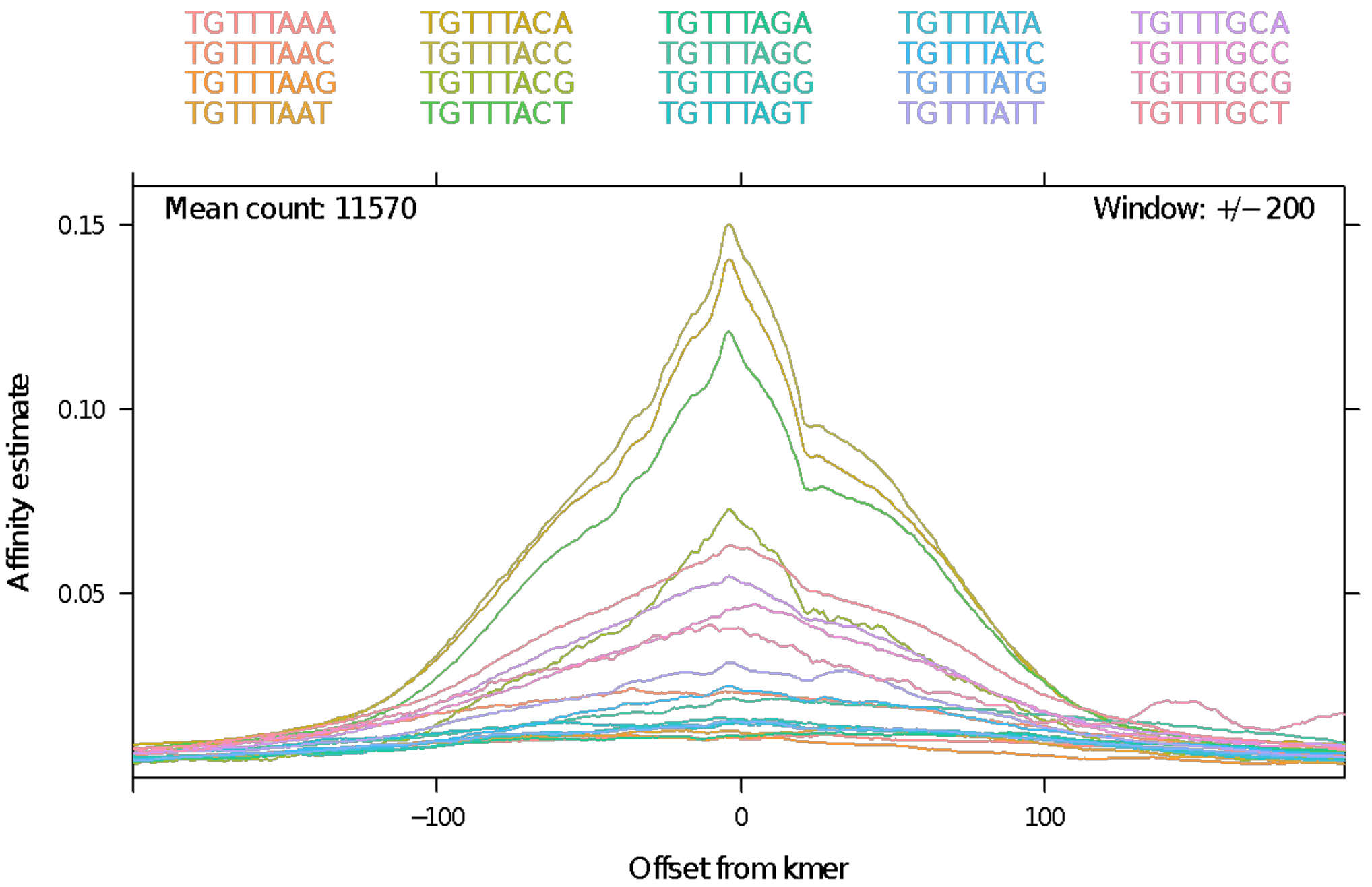
- Are there factors whose binding sites are enriched in somatic mutation count? Do these encompass specific motifs/binding modes?
- Is this mutational burden evenly distributed among cancer types?
- How much does context matter in determining which/how many mutations occur?

# Estimating TF Sequence Affinity

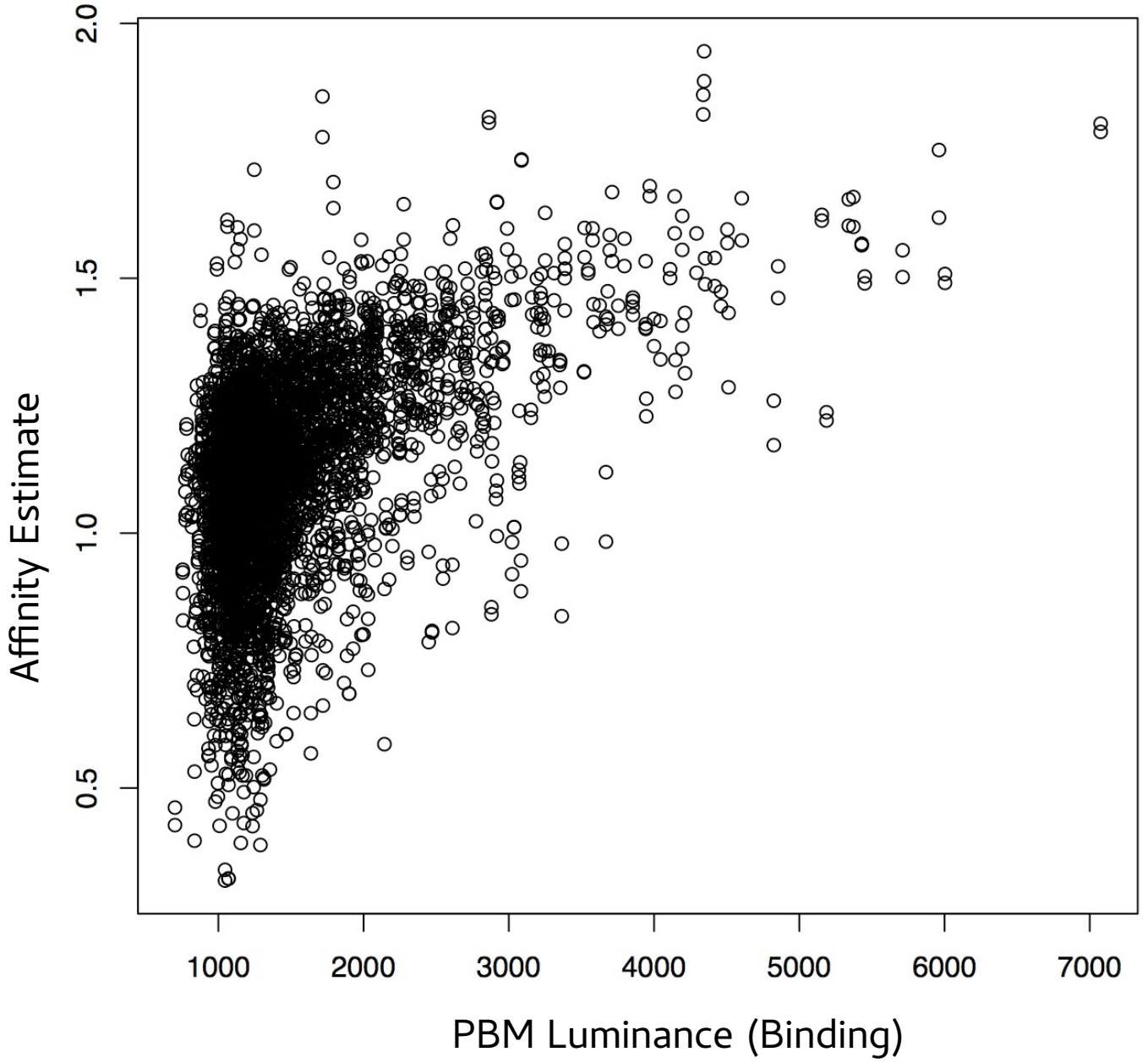
This work primarily uses a straightforward kmer model based on ChIP-seq data to estimate binding affinity of transcription factors



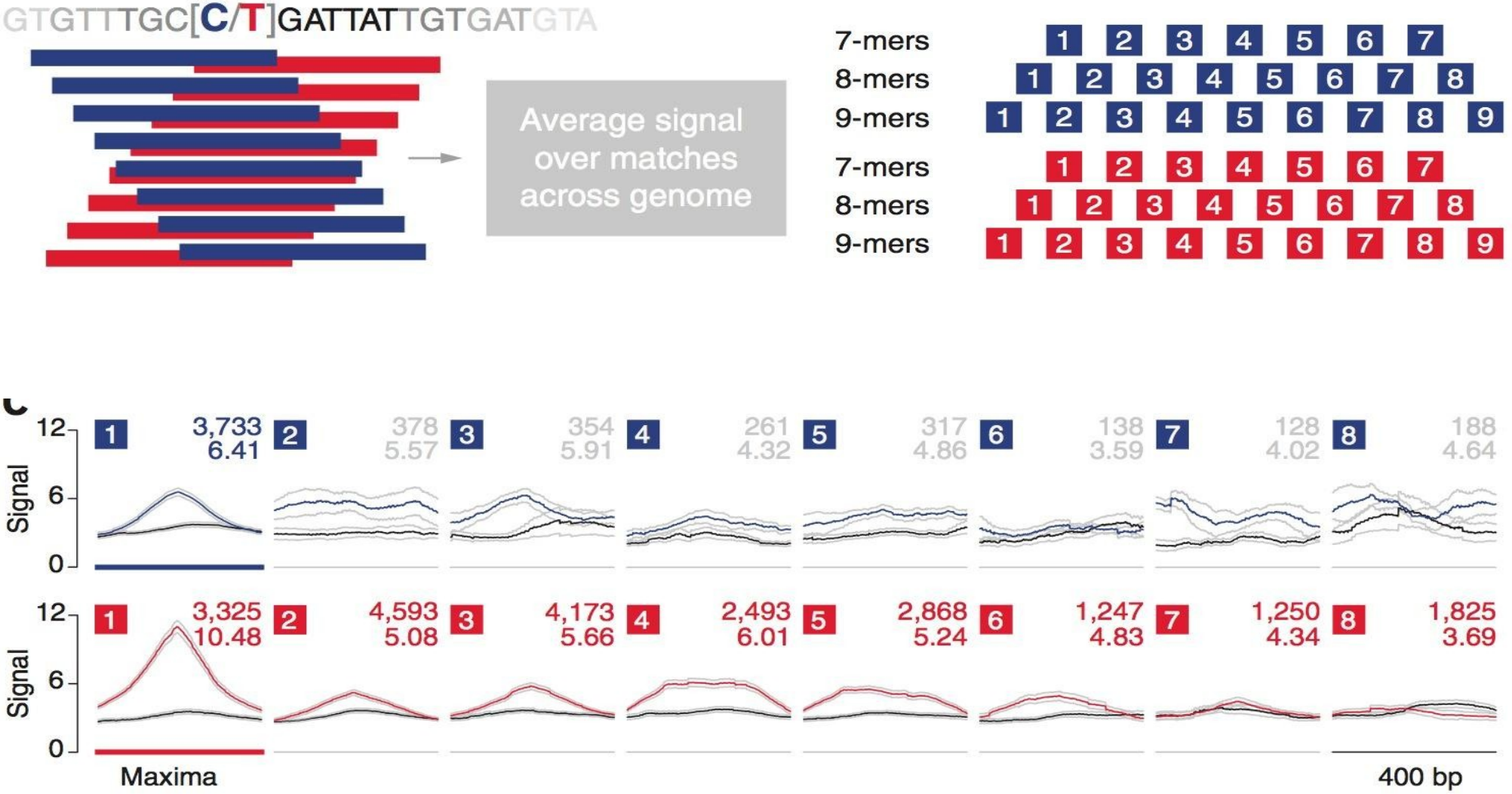
# Example: FoxA1 in T47D



# PBM Experimental validation: GATA3 (ENCODE MCF-7)



# Intragenomic Replicates (IGR)

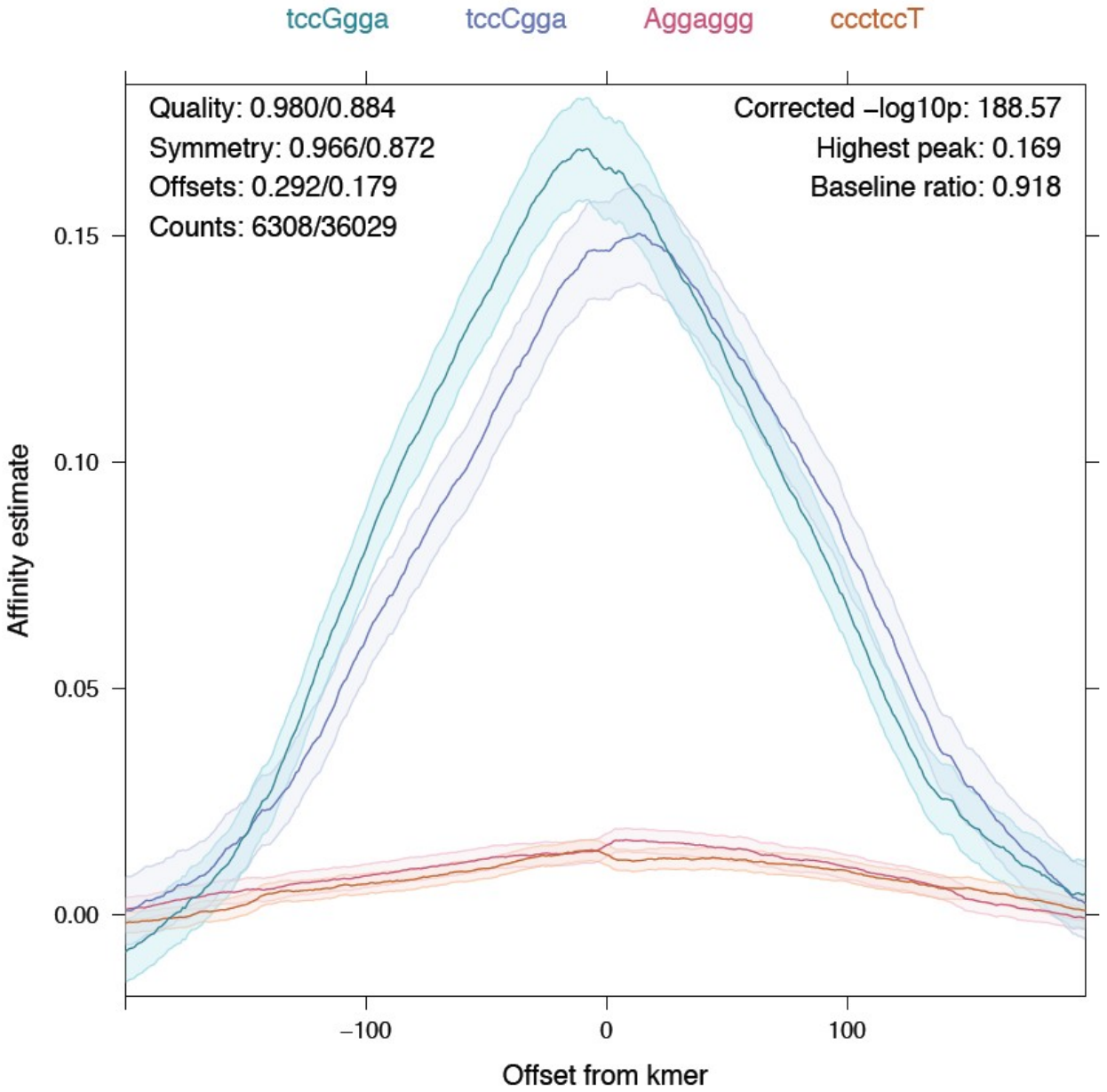


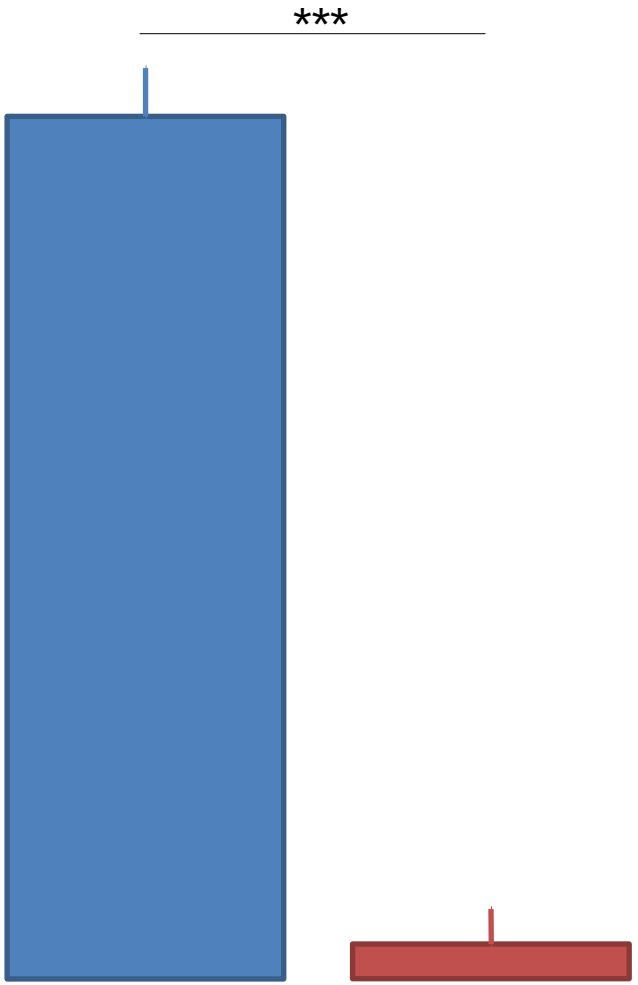

Breast cancer risk-associated SNPs
are enriched in affinity modulating variants.

Cowper-Sallari ... Lupien *Nature Genetics* 2012

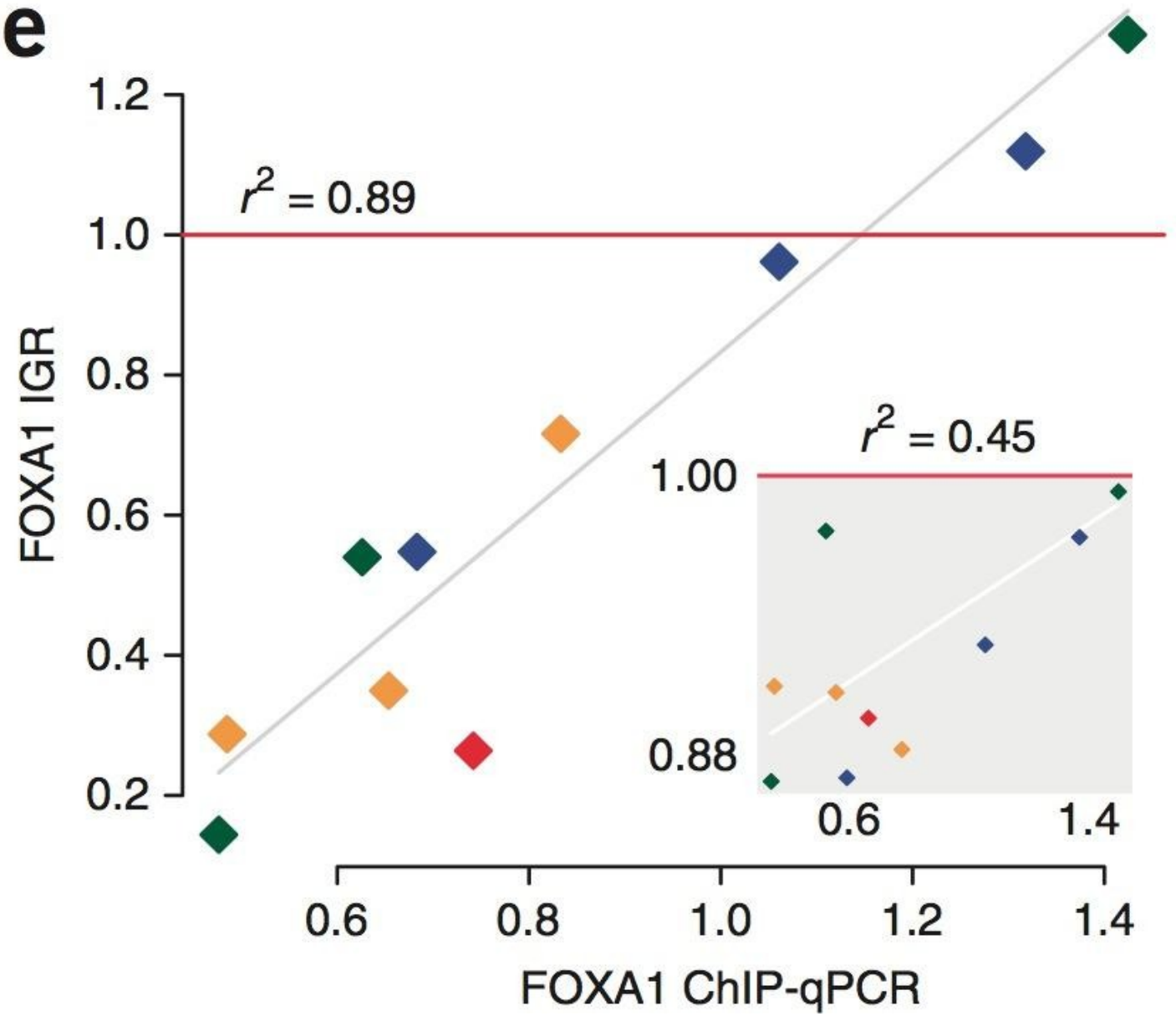
# Intragenomic Replicates (IGR)

PCa mutation (Baca... Garraway 2012) evaluated with LNCaP ETV1 ChIP-seq (Cistrome) IGR model



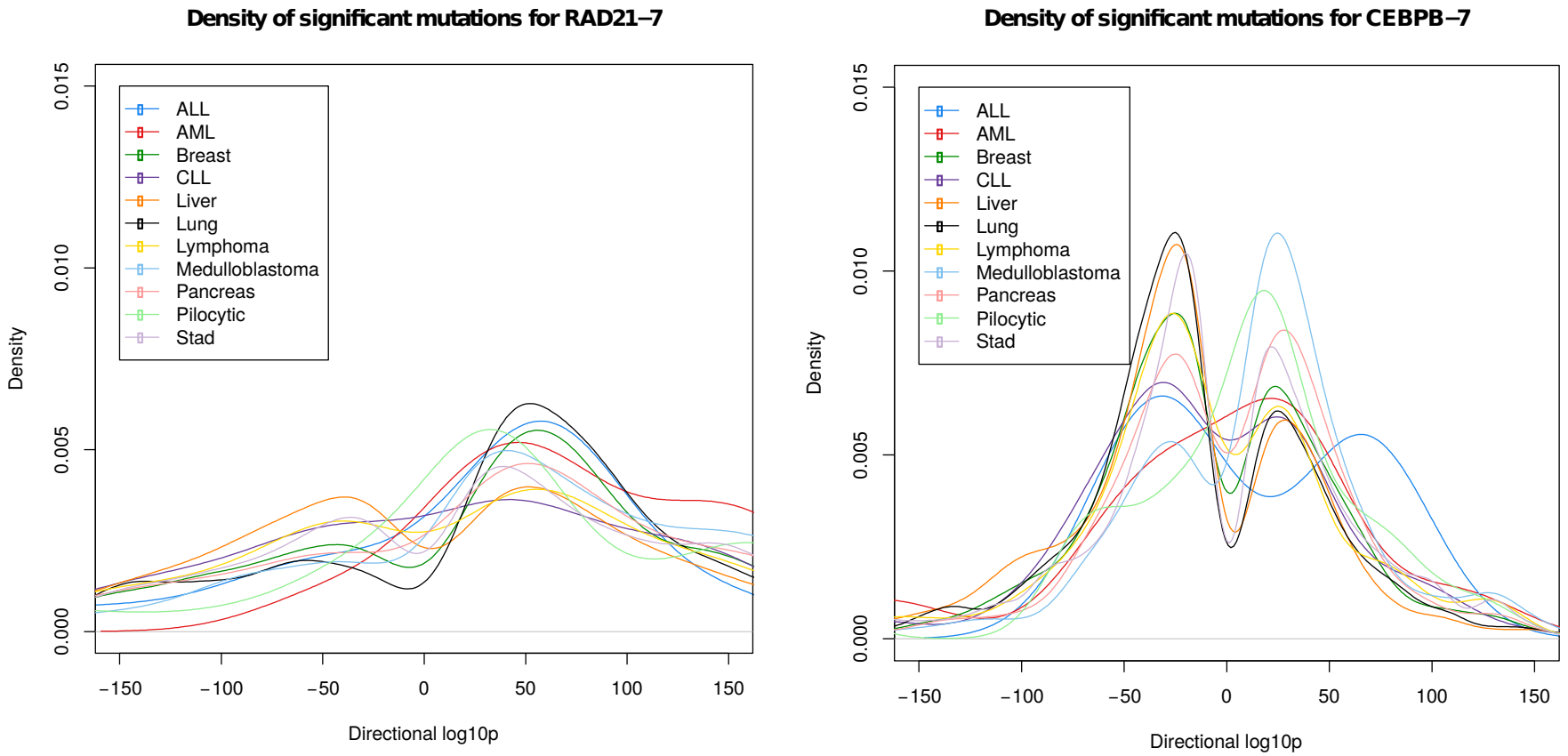
# ChIP–qPCR Experimental validation: IGR >> PWMs




IGR r^2=0.89 with ChIP. PWMs r^2=0.45

Cowper-Sallari ... Lupien *Nature Genetics* 2012

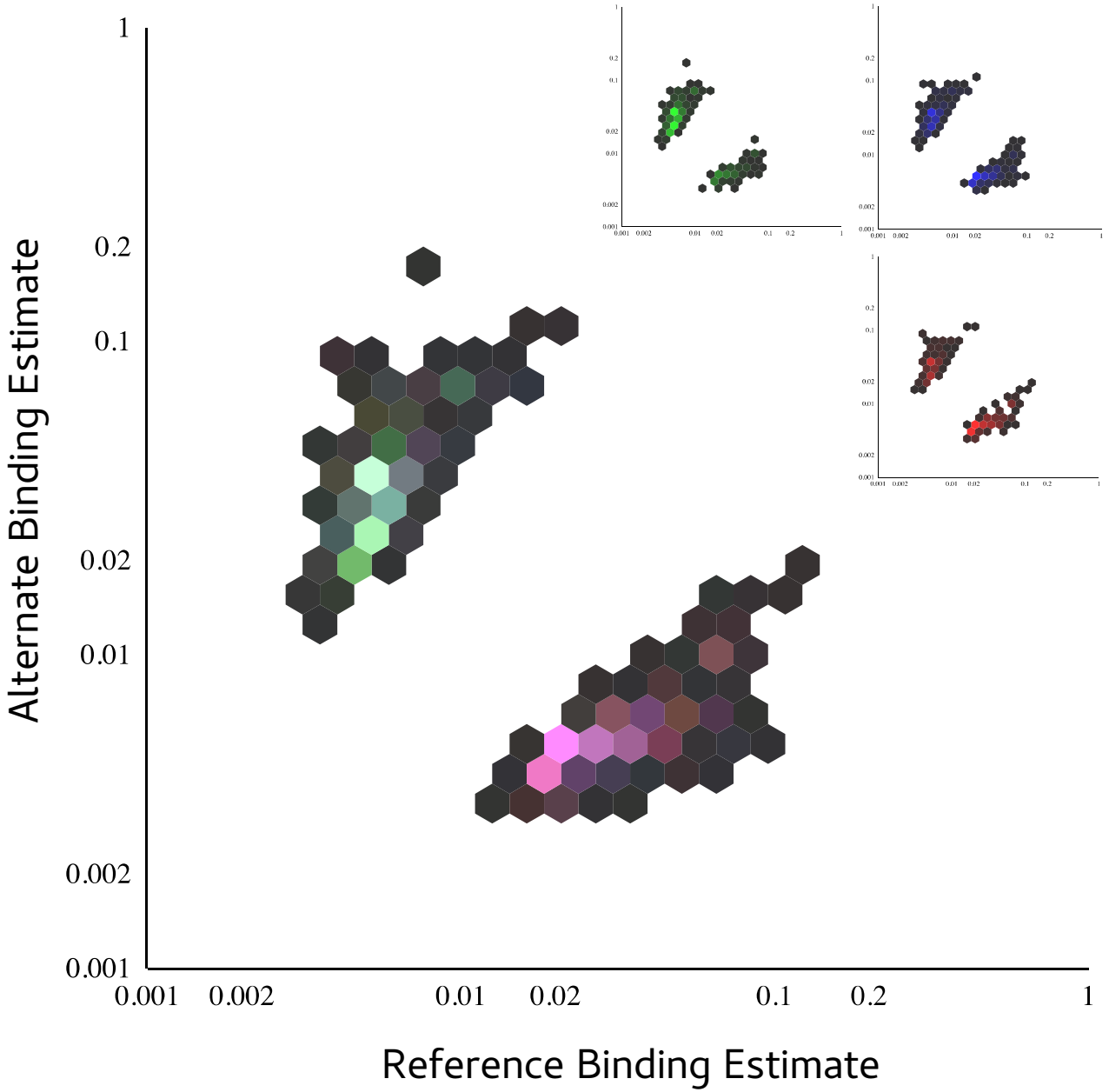
# Functional Annotation of Variants Reveals Cancer-Specific Factors




Individual tumor types from Alexandrov et al. 2013 exhibit **distinct directional biases in affinity modulation** for any individual transcription factor. Some transcription factors, such as RAD21 (left panel), are estimated to consistently decrease their binding in mutated sites across tumor types, while others, such as CEBPB, have different directional biases in different tumor types (right panel).

Caveat: Will be rerunning on the PCAWG May 2016 release, this analysis uses pan-cancer Alexandrov 2013 data.

# Atf3 mutations in DHS vary between cancer types in their predicted increase or decrease in binding relative to the reference




Models were built from ENCODE and all significant mutations which lie in DNaseI regions for which both replicates pass genome-wide Bonferroni significance are reported.

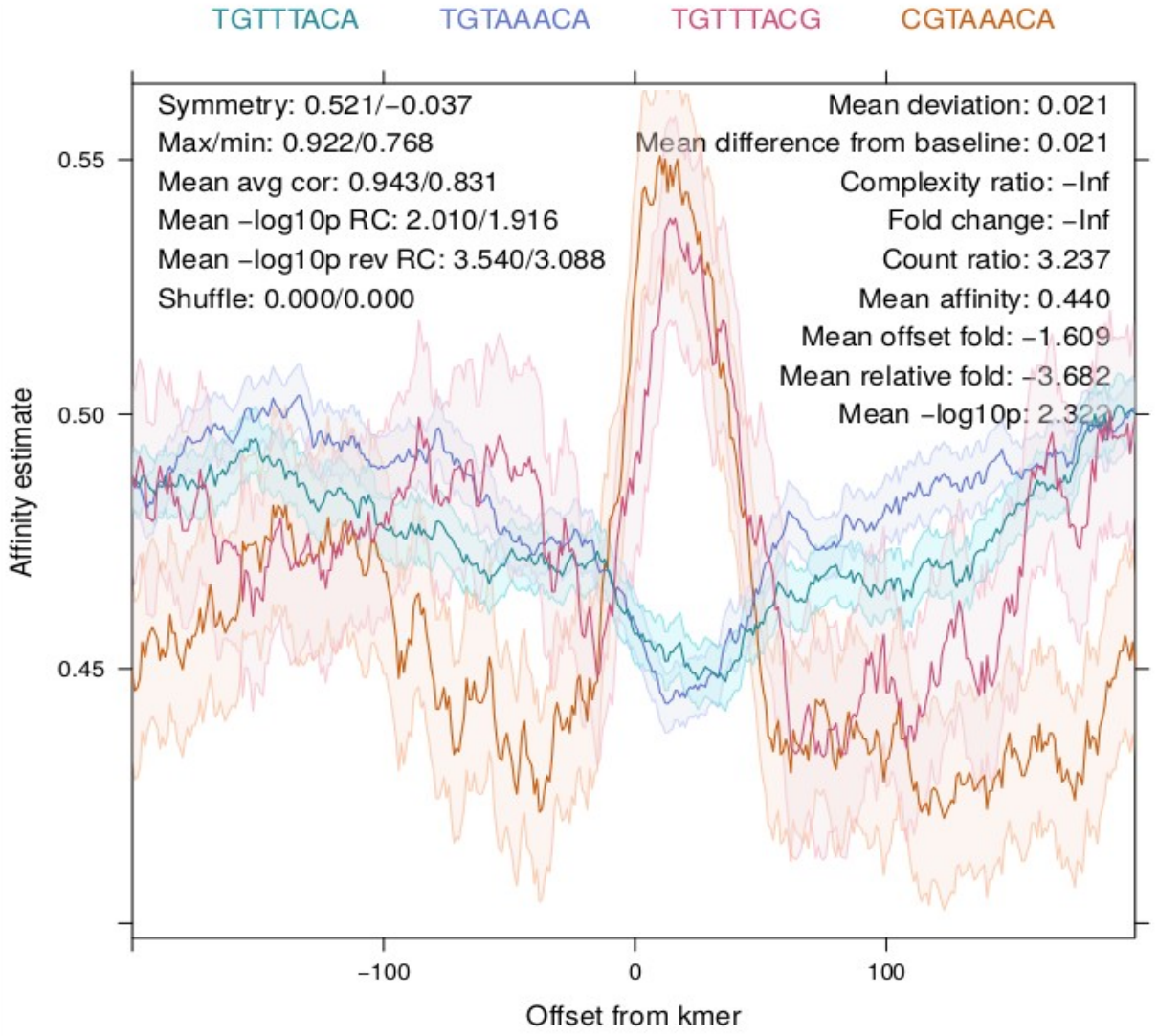
Binding affinity was assessed using a model employed by IGR (Cowper-Sallari et al 2012), trained against all 7mers within ENCODE DHS Master peaks.

Variants for which the binding was not significantly different than shuffles or for which the fold change in binding was less than 4 are excluded from visualization.

Mutations were visualized using a hexbin with each individual color channel normalized to total mutations. Lung (blue) has no significant difference between increase and decrease for fold change > 4, breast (red) has a significant decrease (binomial Bonferroni corrected $p < 9.8e-4$, ~ 0.602 less binding), and liver (green) has a significant increase ($p < 1.76e-26$, ~0.706 more binding).

Caveat: Will be rerunning on the PCAWG May 2016 release, this analysis uses pan-cancer Alexandrov 2013 data.

# Mutational enrichment possibly associated with TFBS motif kmers

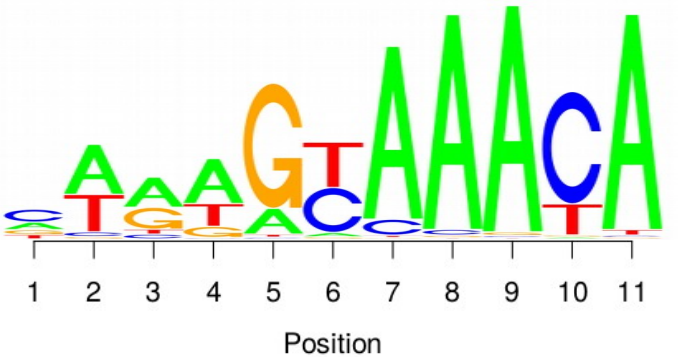



Caveat: Currently rerunning w/gaussian smoothing kernel in place of fixed 25bp window, and this uses pan-cancer Santa Cruz data now.

Hypothesis: looking at which kmers mutations are likely to hit will reveal particular transcription factors shaping tumor progression.

Divide genome into 25bp tiles. For each ENCODE DHS kmer, average the total number of mutations present in the overlapping tiles across all kmer instances to get an average mutation.

FoxA1 motif appears to have a kmer dependent enrichment of mutations



# Summary

- Appears that there are some bound kmers (e.g. for FoxA1) which are enriched in somatic mutation counts.
- Mutational burden of “passenger” mutations is not evenly distributed among cancer types
- Some factors (e.g. Rad21) are consistently altered between different cancer types, might reflect pan-cancer mechanisms of tumorigenesis