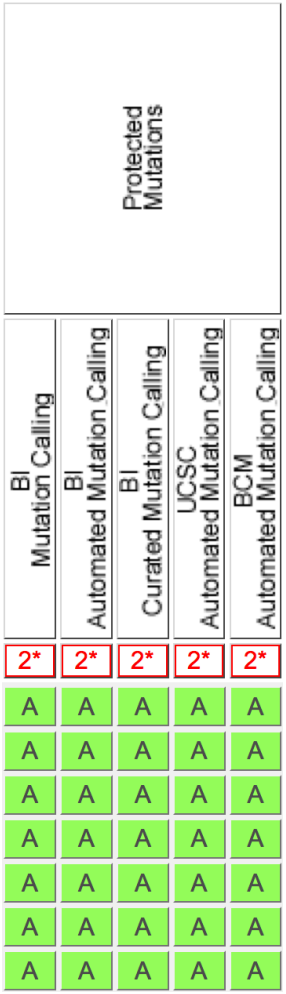
Modelling Somatic Variant Density with Whole Genome Signal Tracks

Lucas Lochovsky
Varyation subgroup
May 8, 2014

Model Target



- Somatic variant density
- Use TCGA Data Portal to download whole genome
  - Prostate cancer Level 2 protected somatic mutations for 309 samples
  - Added to TCGA Data Portal March 2014
- Protected Mutation column for prostate cancer
  - Recent addition
  - July 2013: BI Mutation Calling
  - Jan 2014: UCSC & BCM Automated Mutation Calling
  - March 2014: BI Automated Mutation Calling, BI Curated Mutation Calling
- Used the curated mutation calls

Model Target

- PRAD Level 2 protected somatic mutations from BI Curated Mutation Calling
  - SNP VCFs
  - Extract high quality somatic variants
  - Bin into the same 100,000-bp regions as the other tracks
    - Each bin represents the number of variants found in this region across the 309 samples
- **Model:** For each 100,000-bp region *r*:

$$\text{PRAD\_somatic\_variant\_density}(r) = w_1\text{DNA\_repl\_timing}(r) + w_2\text{H3K4me1}(r) + w_3\text{H3K4me3}(r) + w_4\text{RNA\_seq}(r) + w_5\text{GC\_percent}(r) + w_6\text{DNA\_recomb\_rate}(r) + w_7\text{1KG\_SNV\_density}(r)$$

Pairwise Correlation Matrix

- Correlation matrix
  - Blue: Correlation with significant p-value
  - Red: Correlation > 0.30 with significant p-value
- P-value matrix
  - Red: P-value < 0.05

Correlation matrix

	1KG SNV density	H3K4me1 marks	H3K4me3 marks	RNA-seq	DNA replication timing	GC bias	DNA recombination rate	PRAD somatic variant density
1KG SNV density	1							
H3K4me1 marks	-0.0291	1						
H3K4me3 marks	-0.1595	0.8062	1					
RNA-seq	0.1313	0.0085	-0.0139	1				
DNA replication timing	-0.421	-0.0652	0.0447	-0.1075	1			
GC bias	0.6253	-0.3238	-0.4055	0.0742	-0.2226	1		
DNA recombination rate	0.3904	-0.118	-0.156	0.028	-0.0997	0.4188	1	
PRAD somatic variant density	0.2629	0.0079	-0.0205	0.0708	-0.1804	0.1819	0.0368	1

P-value matrix

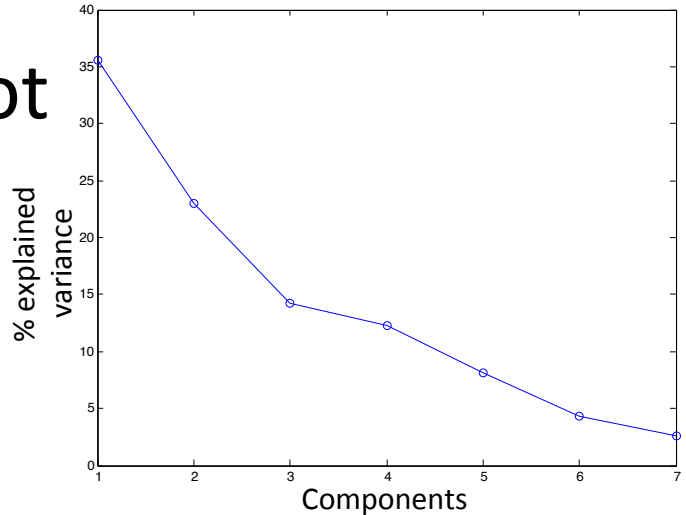
	1KG SNV density	H3K4me1 marks	H3K4me3 marks	RNA-seq	DNA replication timing	GC bias	DNA recombination rate	PRAD somatic variant density
1KG SNV density	1							
H3K4me1 marks	0	1						
H3K4me3 marks	0	0	1					
RNA-seq	0	0.1512	0.0183	1				
DNA replication timing	0	0	0	0	1			
GC bias	0	0	0	0	0	1		
DNA recombination rate	0	0	0	0	0	0	1	
PRAD somatic variant density	0	0.182	0.0005	0	0	0	0	1

Principal Components Analysis

- Coefficient matrix

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
1KG SNV density	0.4422	0.4204	-0.0673	-0.0553	-0.3753	0.6835	0.1106
H3K4me1 marks	-0.3791	0.5705	-0.1205	0.08	-0.1205	-0.0638	-0.701
H3K4me3 marks	-0.4408	0.4841	-0.1314	0.1479	-0.0873	-0.1752	0.7028
RNA-seq	0.0905	0.1853	0.8732	0.4389	0.0394	-0.029	-0.0023
DNA replication timing	-0.2306	-0.4229	-0.2036	0.6856	-0.4646	0.1968	-0.0443
GC bias	0.5269	0.1155	-0.1211	0.0816	-0.4784	-0.6772	-0.0171
DNA recombination rate	0.3571	0.1923	-0.381	0.5471	0.6251	0.0034	-0.0164

- Scree plot



Looks like there's an "elbow" at Comp 3

Linear fit (all variables)

- Linear regression model:
  - y ~ 1 + 1KG_SNV_density + GC_percent + H3K4me1 + H3K4me3 + DNA_recomb_rate + DNA_repl_timing + RNA-seq

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-2.79E-16	0.0056422	-4.95E-14	1
1KG SNV density	0.21789	0.0081259	26.815	1.83E-156
GC percent	0.074696	0.0081397	9.1767	4.73E-20
H3K4me1	-0.016129	0.0098185	-1.6427	0.10046
H3K4me3	0.048703	0.0099399	4.8998	9.65E-07
DNA recomb rate	-0.082698	0.0063345	-13.055	7.66E-39
DNA repl timing	-0.080174	0.006288	-12.75	3.91E-37
RNA-seq	0.031164	0.0057025	5.4649	4.67E-08

<-- Not significant

Number of observations: 28801, Error degrees of freedom: 28793

Root Mean Squared Error: 0.958

R-squared: 0.0834, Adjusted R-Squared 0.0831

F-statistic vs. constant model: 374, p-value = 0

Linear fit (excluding H3K4me1)

- Linear regression model:
  - y ~ 1 + 1KG_SNV_density + GC_percent + H3K4me3 + DNA_recomb_rate + DNA_repl_timing + RNA-seq

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-1.81E-16	0.0056423	-3.21E-14	1
1KG SNV density	0.21572	0.0080178	26.905	1.70E-157
GC percent	0.076268	0.0080835	9.435	4.19E-21
H3K4me3	0.035976	0.0062271	5.7773	7.67E-09
DNA recomb rate	-0.082483	0.0063334	-13.024	1.16E-38
DNA repl timing	-0.079101	0.0062541	-12.648	1.44E-36
RNA-seq	0.031128	0.0057027	5.4586	4.84E-08

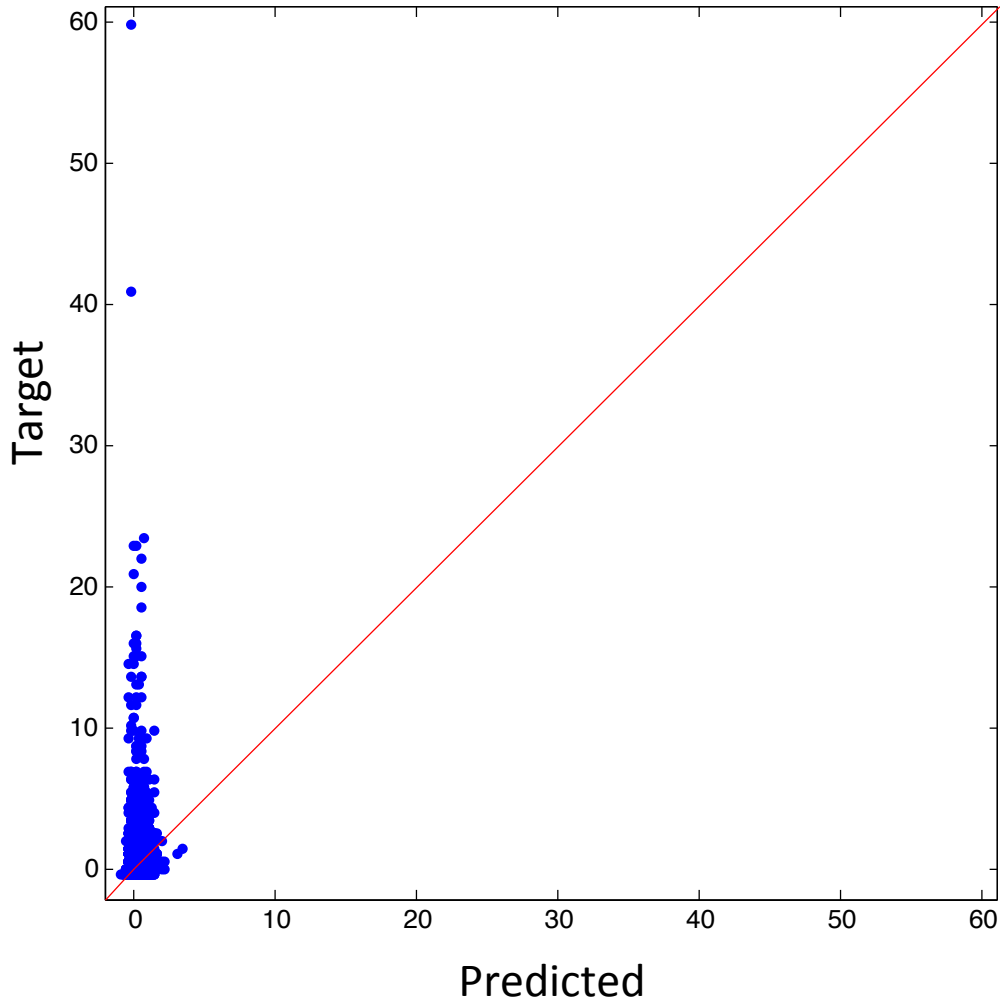
Number of observations: 28801, Error degrees of freedom: 28794

Root Mean Squared Error: 0.958

R-squared: 0.0833, Adjusted R-Squared 0.0831

F-statistic vs. constant model: 436, p-value = 0

Scatterplot of Target vs. Predicted



A few regions are predicted below the target → Higher somatic density than expected

Target vs. Predicted

- Are these higher density regions cancer-related, or part of the background?
- Investigated the regions with the greatest difference between target and predicted
  - 33 regions with difference >10
- Find genes in those regions, and find recurrent variants in those genes (LARVA-Core)

Target vs. Predicted

Gene	# rec samples	# rec variants
FRG1B	83	24
TTN	48	6
TTN-AS1	45	6
MLLT3	36	7
NBPF10	32	5
MUC16	32	4
SPOP	30	8
NBPF1	29	5

Also observed the neuroblastoma-related NBPF pseudogene family in these regions of elevated somatic variant density

- Excerpt from the top of the list
- Includes legit prostate cancer genes (SPOP)
- Also spurious genes (MUC16)
- **Followup:** Calibrate null model based on recurrences that appear across multiple cancers
  - **Common genes:** Background mutation (in effect across all cancers)
  - **Unique genes:** Cancer-specific