Investigating Features of Gene Essentiality Via CRISPR Knockout Screens

Guiding Question: What are the “essential genes” for life and what biologically meaningful facts can we say about them?

By James Bannon, Sunil Deochand, Taymor Hekal, & Hamd Bilal Tahir

Project Supervisor: Neville Sanjana

An (Essential) Outline

1. Basic Background:
    a. Essentiality In A Nutshell
    b. Brief History Knockout Screens & CRISPR
    c. The genomeCRISPR dataset & recent work

1. Methods + Results:
    a. Functional Enrichment / GSEA
    b. SNPS
    c. Cancer Mutations

Essentiality In a Nutshell

Core Essentiality vs. Conditional Essentiality

Core Essential = genes without which all cells would die/commit apoptosis.

"Complete loss of function results in complete loss of fitness."

Conditional Essential = genes without which some cell types would die.

What's **nonessential?**

Hard to say! Many confounds, like regulatory networks, exist.

Finding Essentiality Part 1: Knockdowns & Models

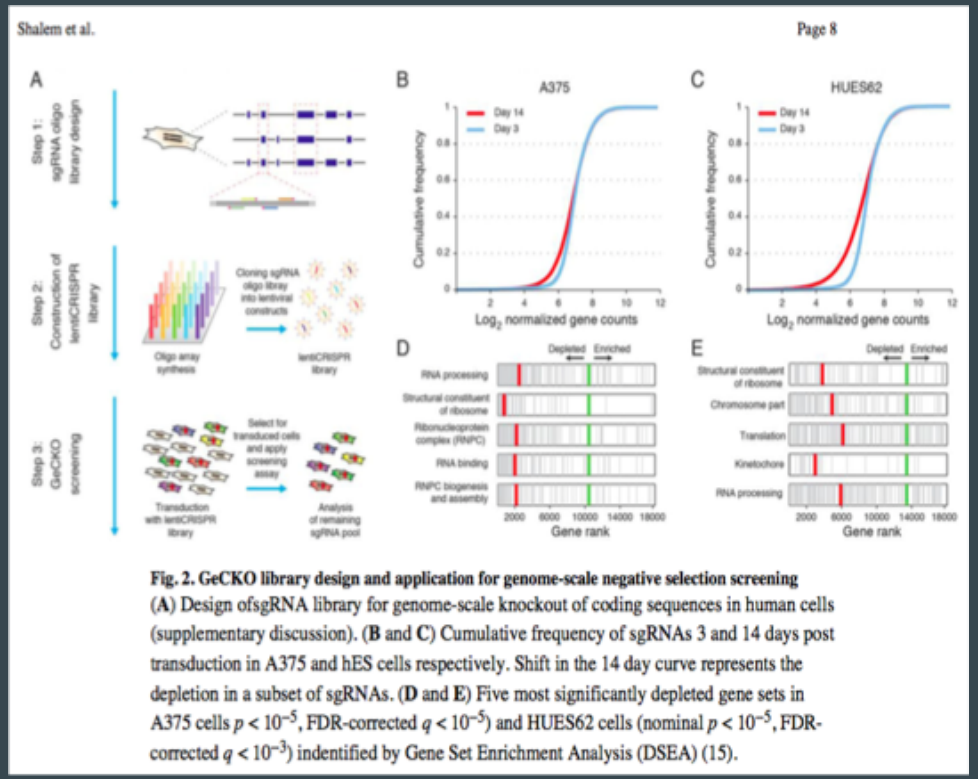


Fig. 2. GeCKO library design and application for genome-scale negative selection screening
(**A**) Design ofsgRNA library for genome-scale knockout of coding sequences in human cells (supplementary discussion). (**B** and **C**) Cumulative frequency of sgRNAs 3 and 14 days post transduction in A375 and hES cells respectively. Shift in the 14 day curve represents the depletion in a subset of sgRNAs. (**D** and **E**) Five most significantly depleted gene sets in A375 cells $p < 10^{-5}$, FDR-corrected $q < 10^{-5}$) and HUES62 cells (nominal $p < 10^{-5}$, FDR-corrected $q < 10^{-3}$) indentified by Gene Set Enrichment Analysis (DSEA) (15).

For a long time only complete gene expression inhibition was inaccurate, costly, and inefficient

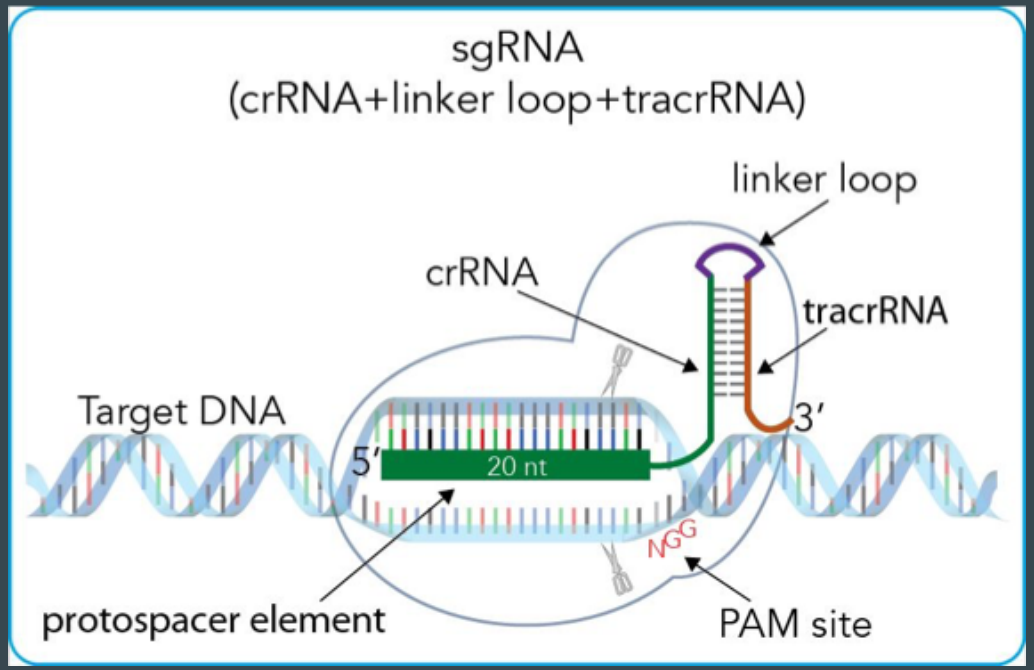
RNAi (RNA interference) was best we could do to systematically knockdown genes

Issues:

- Imprecise
- Incomplete Knockdown

=>Noisy!

Finding Essentiality Part 2: CRISPR Knockout Screens



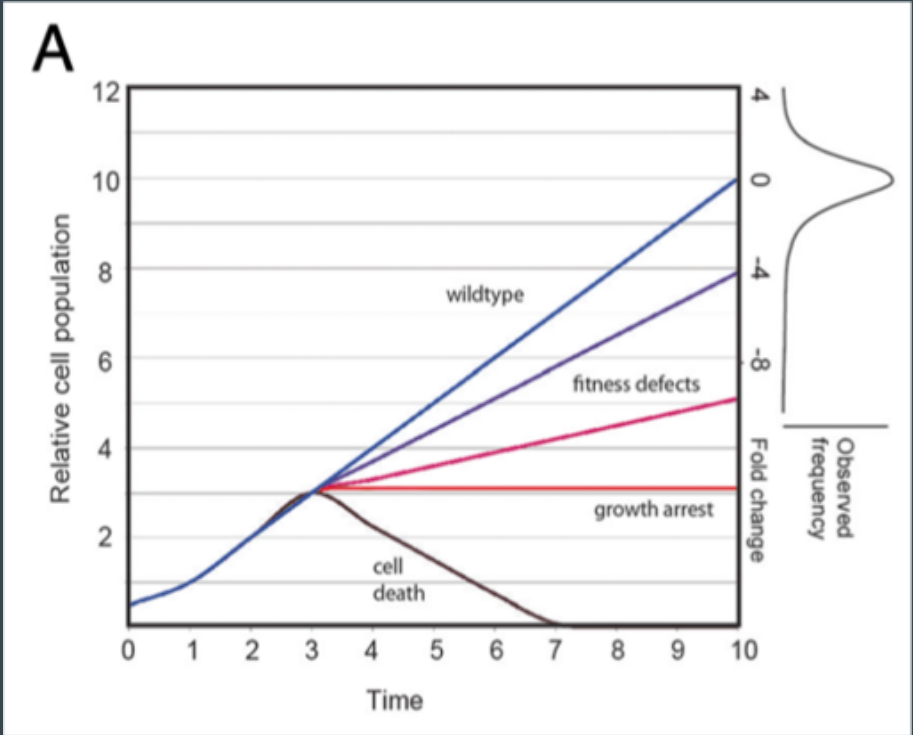
sgRNA (guide RNA) is attached to Cas9 protein and finds specific complementary sequence in genome

Binds & cleaves desired gene with high precision

Now every organism is a model organism!

http://www.idtdna.com/pages/decoded/decoded-articles/pipet-tips/decoded/2015/09/18/designing-sgrnas-for-crispr-cas9-experiments, Yang L, et al. CRISPR/Cas9-Directed Genome Editing of Cultured Cells. Curr Protoc Mol Biol, doi: 10.1002/0471142727:1–17.

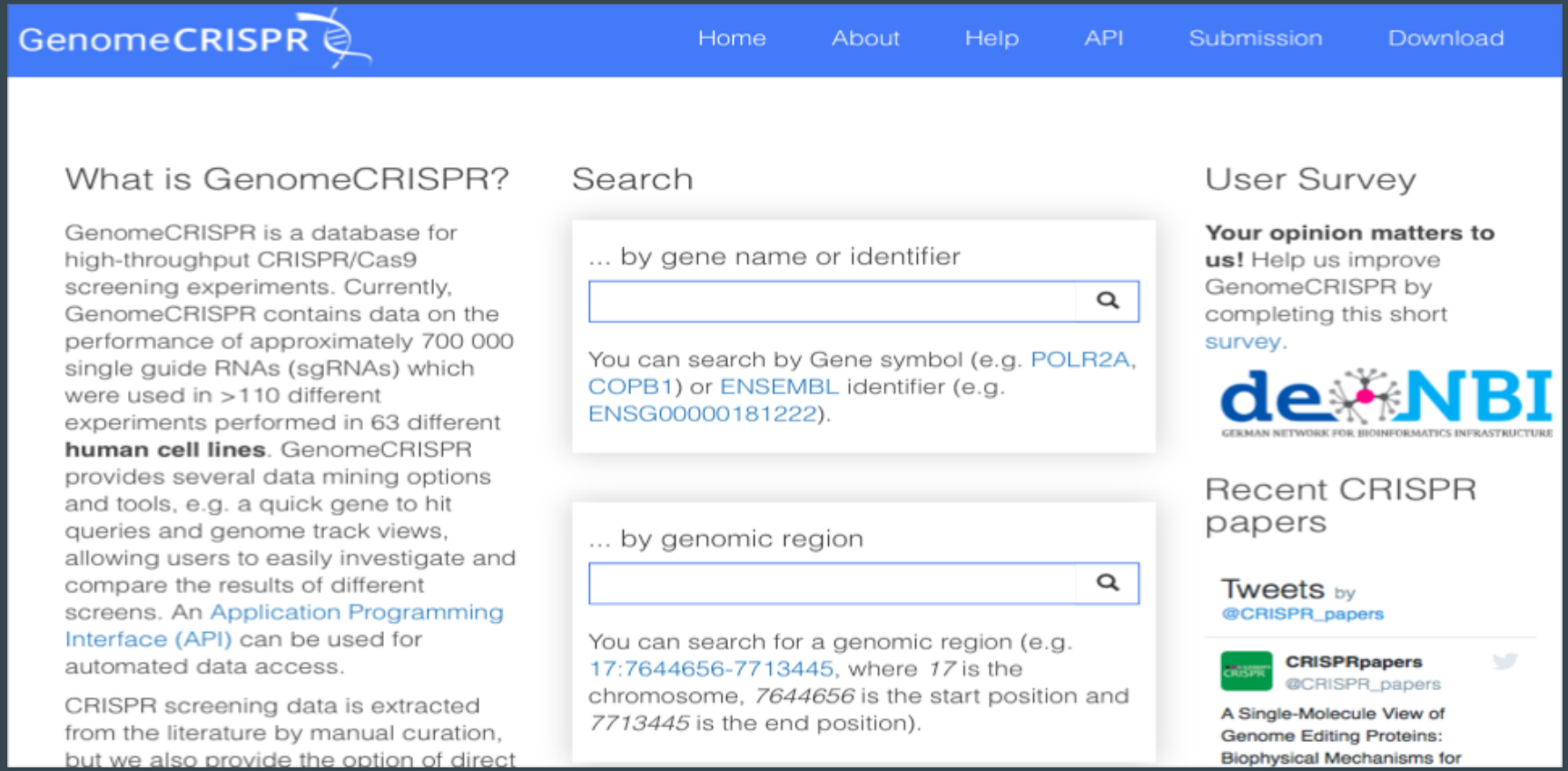
Finding Essentiality Part 2: CRISPR Knockout Screens



1. Acquire a population of cells
2. Acquire a library of sgRNAs
3. Multiple sgRNAs per gene per screen
4. Each sgRNA species only targets one gene
5. On average each cell transfected with one sgRNA
6. Sequence initial population of sgRNAs
7. After some time resequence

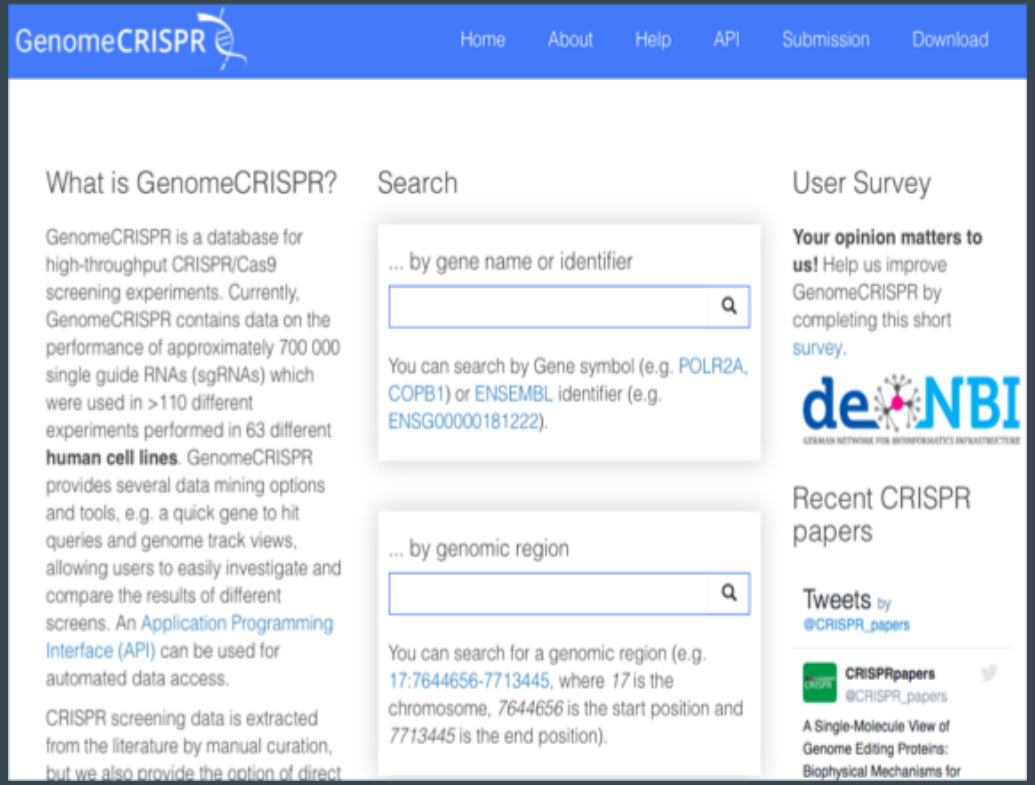
BAGEL: a computational framework for identifying essential genes from pooled library screens. Hart, Traver; Moffat; Jason *BMC* Bionformatics

Our Knockout Dataset: GenomeCRISPR



Rauscher, B., Heigwer, F., Breinig, M., Winter, J., Boutros, M. (2017). GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. Nucleic Acids Res 45:D679-D86

Our Knockout Dataset: GenomeCRISPR

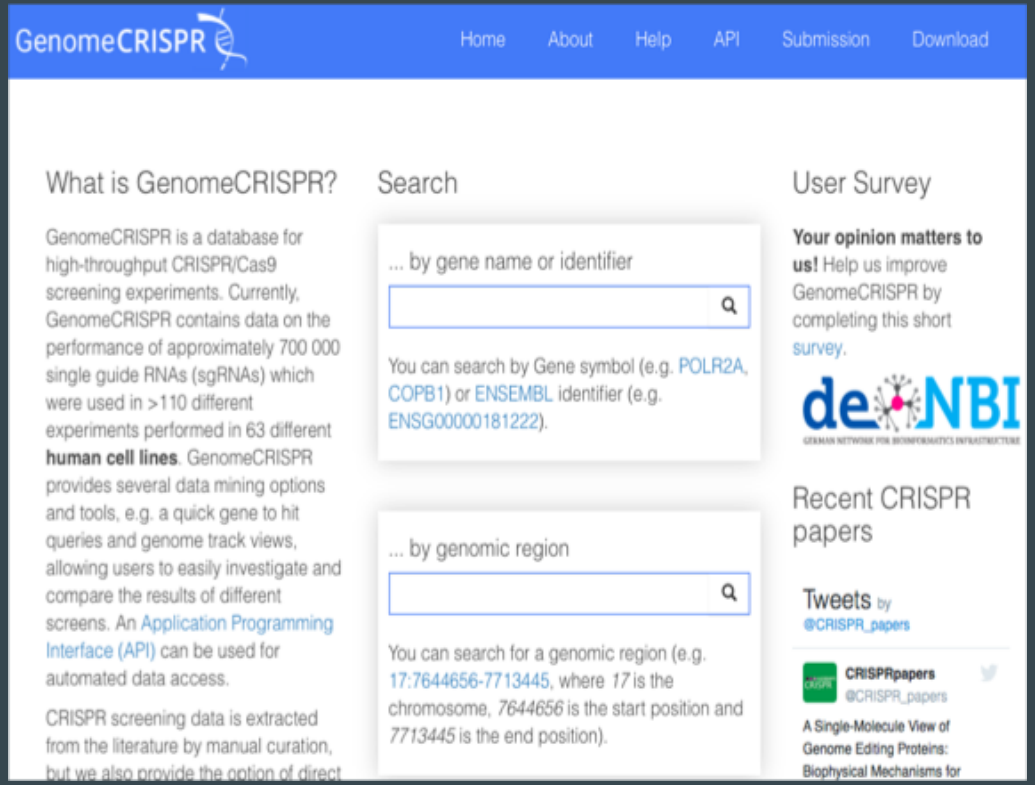


Two main types of screens:

1. Negative Selection - infuse a cell population/cell line with sgRNA library, after some time see what survives.
2. Positive Selection - Similar but in the presence of a pharmaceutical compound

Rauscher, B., Heigwer, F., Breinig, M., Winter, J., Boutros, M. (2017). GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. Nucleic Acids Res 45:D679-D86

Our Knockout Dataset: GenomeCRISPR



Two main types of screens:

1. Negative Selection - infuse a cell population/cell line with sgRNA library, after some time see what survives.
2. Positive Selection - Similar but in the presence of a pharmaceutical compound

Rauscher, B., Heigwer, F., Breinig, M., Winter, J., Boutros, M. (2017). GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. Nucleic Acids Res 45:D679-D86

The GenomeCRISPR Dataset

	start	end	chr	strand	pubmed	cellline	condition	sequence	symbol	ensg
1	50844073	50844096	10	+	26472758	Jiyoye	viability	GCAGCATCCCAACCAGGTGGAGG	A1CF	ENSG00000148584
2	50814011	50814034	10	-	26472758	Jiyoye	viability	GCGGGAGTGAGAGGACTGGGCGG	A1CF	ENSG00000148584
3	50836111	50836134	10	+	26472758	Jiyoye	viability	ATGACTCTCATACTCCACGAAGG	A1CF	ENSG00000148584
4	50836095	50836118	10	-	26472758	Jiyoye	viability	GAGTCATCGAGCAGCTGCCATGG	A1CF	ENSG00000148584
5	50816234	50816257	10	-	26472758	Jiyoye	viability	AGTCACCCTAGCAAAACCAGTGG	A1CF	ENSG00000148584
6	50816119	50816142	10	-	26472758	Jiyoye	viability	GATCCCACCACAACCTACCTTGG	A1CF	ENSG00000148584

	log2fc	rc_initial	rc_final	effect	cas	screentype
1	0.316	[260]	[244]	2	hSpCas9	negative selection
2	2.144	[17]	[59]	9	hSpCas9	negative selection
3	1.426	[75]	[153]	8	hSpCas9	negative selection
4	1.550	[47]	[105]	8	hSpCas9	negative selection
5	0.383	[58]	[57]	3	hSpCas9	negative selection
6	0.993	[358]	[538]	6	hSpCas9	negative selection

Rauscher, B., Heigwer, F., Breinig, M., Winter, J., Boutros, M. (2017). GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. Nucleic Acids Res 45:D679-D86

Recent Work Using GenomeCRISPR

A Map Of Genetic Interactions In Cancer Cells

Benedikt Rauscher, Florian Heigwer, Luisa Henkel, Thomas Hielscher, Michael Boutros

doi: https://doi.org/10.1101/120964

Evaluation and Design of Genome-wide CRISPR/Cas9 Knockout Screens

Traver Hart, Amy Tong, Katie Chan, Jolanda van Leeuwen, Ashwin Seetharaman, Michael Aregger, Megha Chandrashekhar, Nicole Hustedt, Sahil Seth, Avery Noonan, Andrea Habsid, Olga Sizova, Lyudmilla Nedyalkova, Ryan Climie, Keith Lawson, Maria Augusta Sartori, Sabriyeh Alibai, David Tieu, Sanna Masud, Patricia Mero, Alexander Weiss, Kevin R. Brown, Matej Ušaj, Maximilian Billmann, Mahfuzur Rahman, Michael Costanzo, Chad L. Myers, Brenda Andrews, Charlie Boone, Daniel Durocher, Jason Moffat

doi: https://doi.org/10.1101/117341

BAGEL: A Bayesian Approach to Essentiality

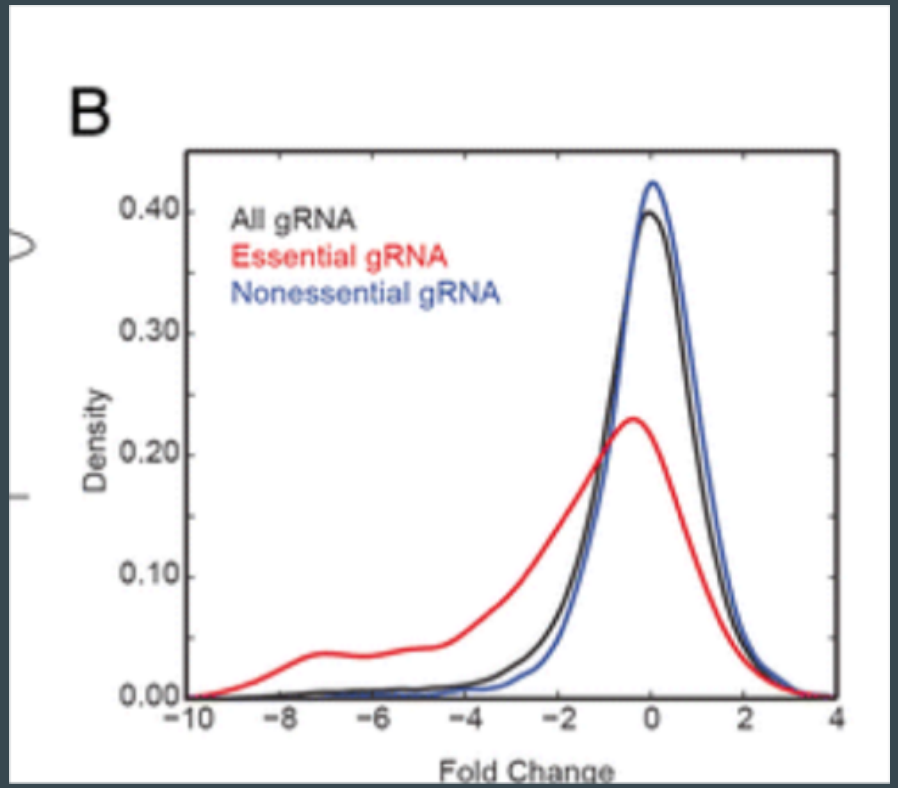
At heart an application of Bayes' Rule:

$$P(E|D) = (P(E)/P(D))P(D|E)$$

Estimate by computing: $P(D|E)$

Compute KDE estimation from reference set.

For each gene compute $P(D|E)$ via Monte-Carlo simulation.



BAGEL: a computational framework for identifying essential genes from pooled library screens. Hart, Traver; Moffat; Jason *BMC* Bionformatics

First Pass: We Attempt To Beat BAGEL

Idea was to gather and use biological information that previous tools (BAGEL) didn't account for:

- Time between RNA-seq FC measurements, tissue type, sgRNA used
- Gene Ontology terms, protein-protein interactions, hubs in networks

=> Insert into supervised learning algorithm

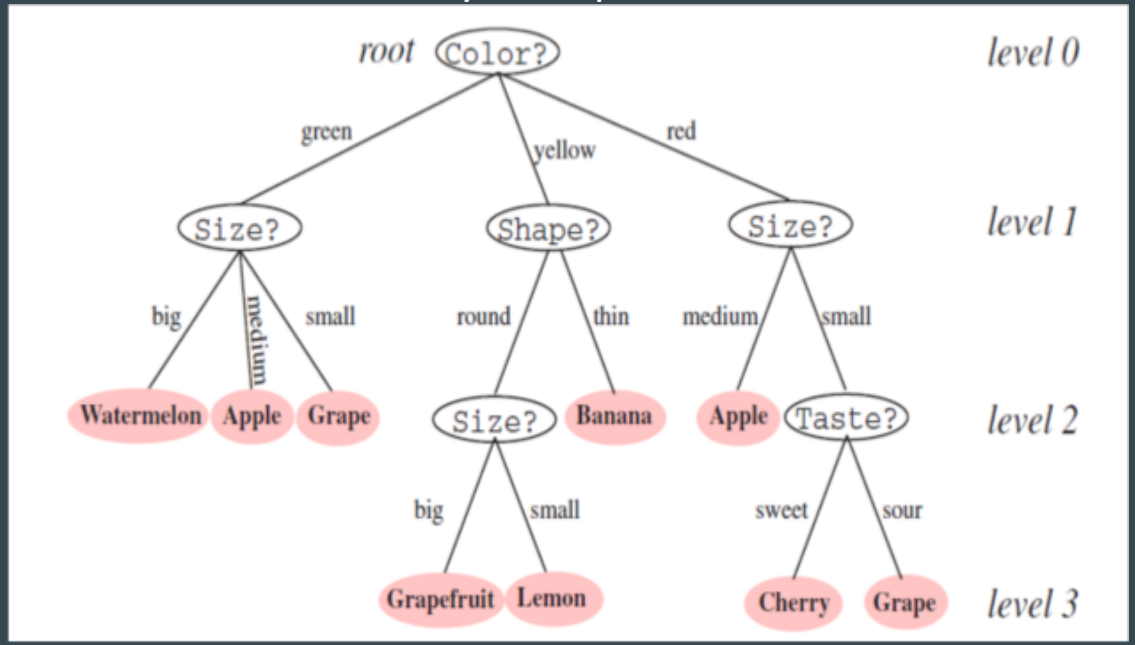
Assigning a single Gene Ontology term to each gene for the model:

- Difficult since a gene can fall under multiple parent GO terms
- Top-down BFS algorithm was inaccurate and failed to find many genes

First Pass: We Attempt To Beat BAGEL

Use Hart et. al's 'reference essential' and 'reference non-essential' sets.

Random Forest binary class prediction.



<- Basic Idea

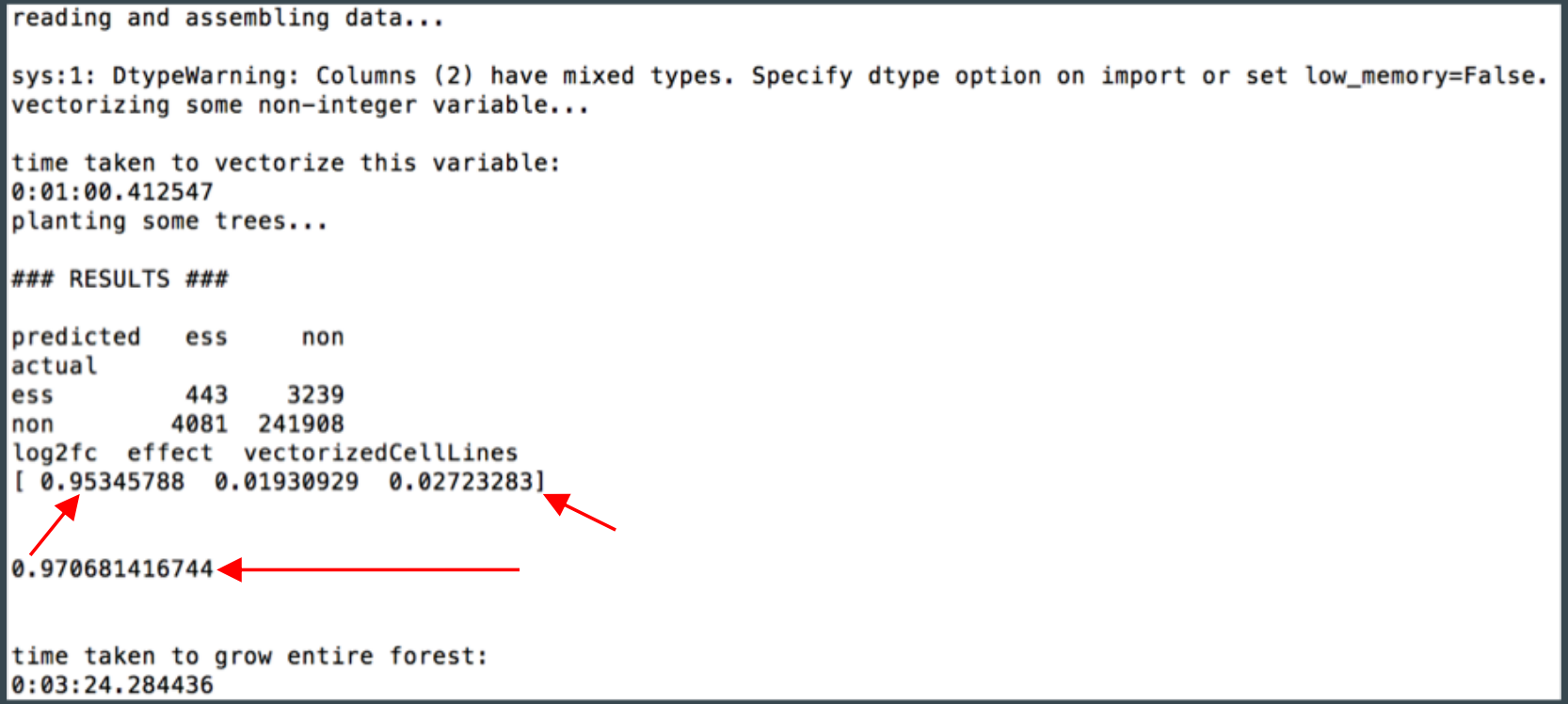
1) Build many uncorrelated decision trees.

2) Aggregate their predictions by majority vote.

First Pass: We Attempt To Beat BAGEL

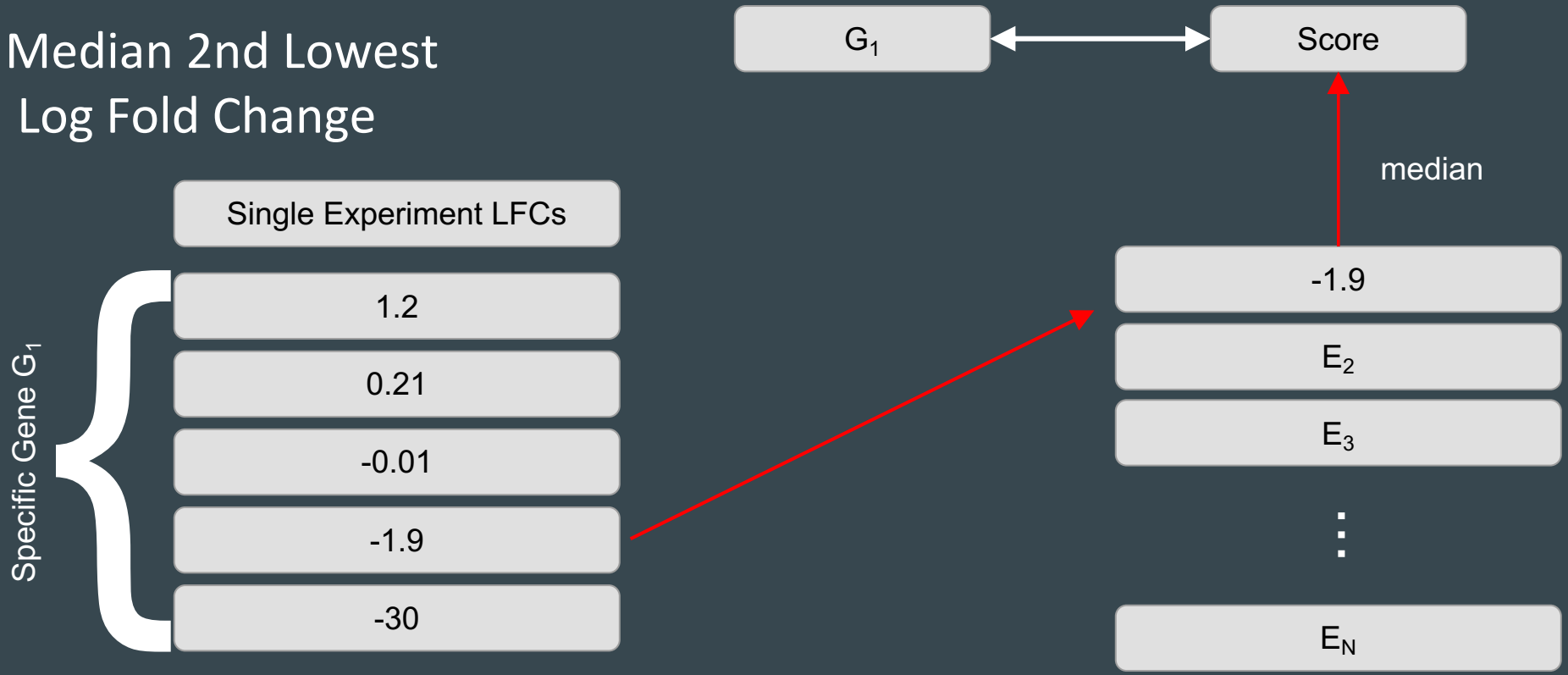
We did okay! ~97% accuracy, although quite a few false positives...

```
reading and assembling data...

sys:1: DtypeWarning: Columns (2) have mixed types. Specify dtype option on import or set low_memory=False.
vectorizing some non-integer variable...

time taken to vectorize this variable:
0:01:00.412547
planting some trees...

### RESULTS ###

predicted   ess     non
actual
ess         443    3239
non        4081  241908
log2fc  effect  vectorizedCellLines
[ 0.95345788  0.01930929  0.02723283]


0.970681416744


time taken to grow entire forest:
0:03:24.284436
```



If Not Random Forest or BAGEL, Then What? From Binary, to Continuous Essentiality

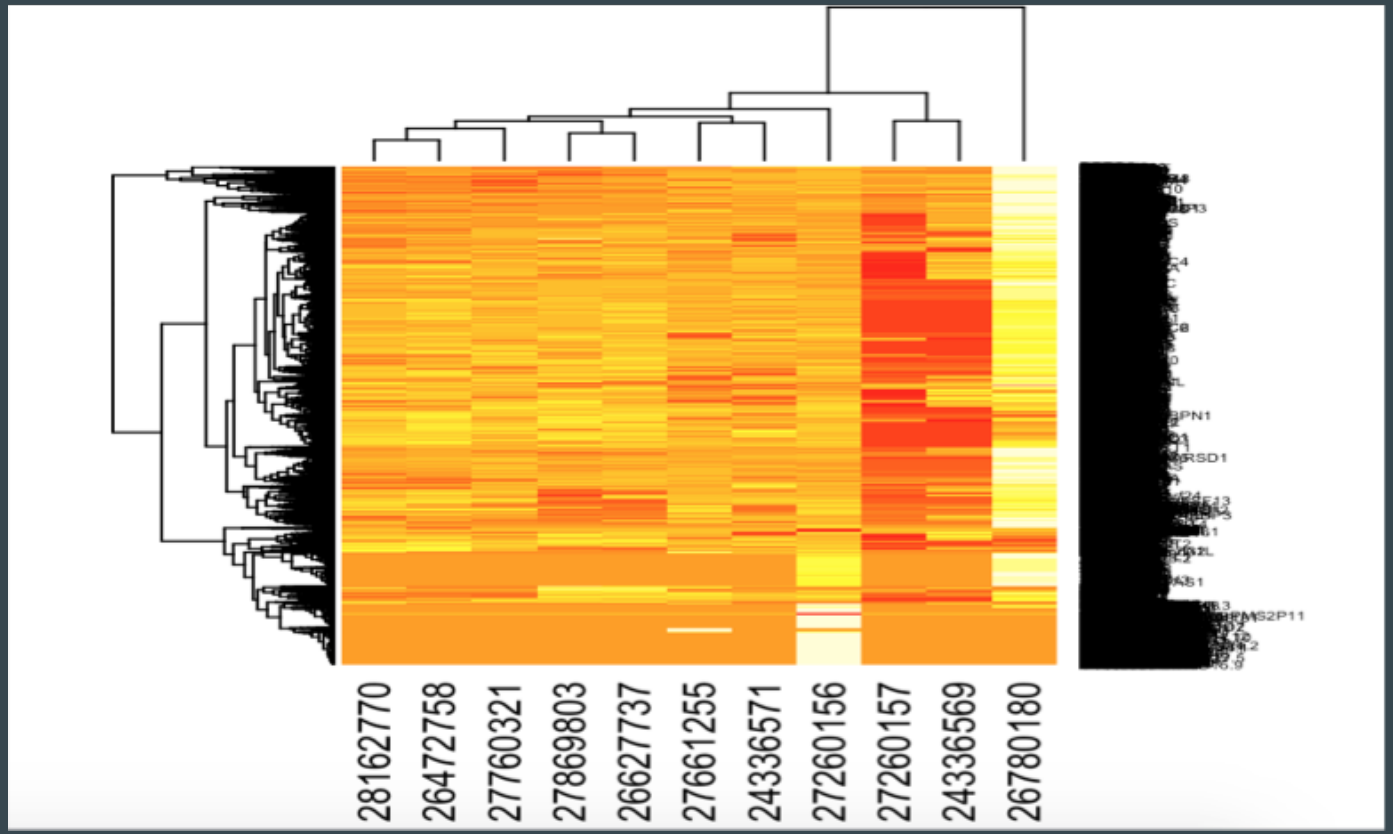
Median 2nd Lowest Log Fold Change



Exploring batch effects in the GenomeCRISPR dataset

Median LogFC per experiment (PubMed ID) across genes

Apparent inconsistency between different experiments



Is LogFC a Good Proxy for Essentiality?

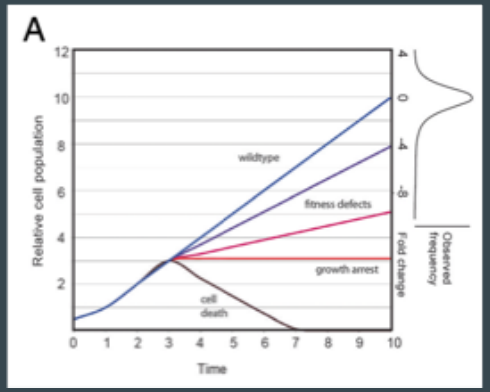
Hard to say!

Why? We have no **labeled data**

But we can make logical biological inferences

Polymerase involved in mRNA synthesis? Probably essential

Tissue-specific TF? Probably not



BAGEL: a computational framework for identifying essential genes from pooled library screens. Hart, Traver; Moffat; Jason *BMC* Bioinformatics

Is LogFC a Good Proxy for Essentiality?

Hard to say!

Why? We have no **labeled data**

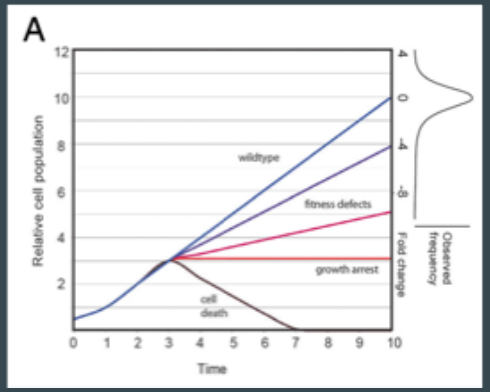
But we can make logical biological inferences

Polymerase involved in mRNA synthesis? Probably essential

Tissue-specific TF? Probably not

Can we do this systematically for all our GenomeCRISPR genes?

Yes, **Functional Enrichment!**



BAGEL: a computational framework for identifying essential genes from pooled library screens.
Hart, Traver; Moffat; Jason *BMC* B

Gene Functional Enrichment (or GSEA) : What is it, and how to quantify enrichment?

Testing for significant enrichments in a gene set

H_0: No enrichment in given gene set; genes are randomly distributed throughout functional terms

P-value returned using a binomial test:

What are the chances of rolling a die and recording '5' 2 times or less out of 100 n die

```
> pbinom(2, 100, 1/6)
[1] 0.00000264
```

Use the pmf for the BD, and keep a running su

$$p-value = \sum \binom{K}{k} p(c)^k (1-p(c))^{K-k}$$

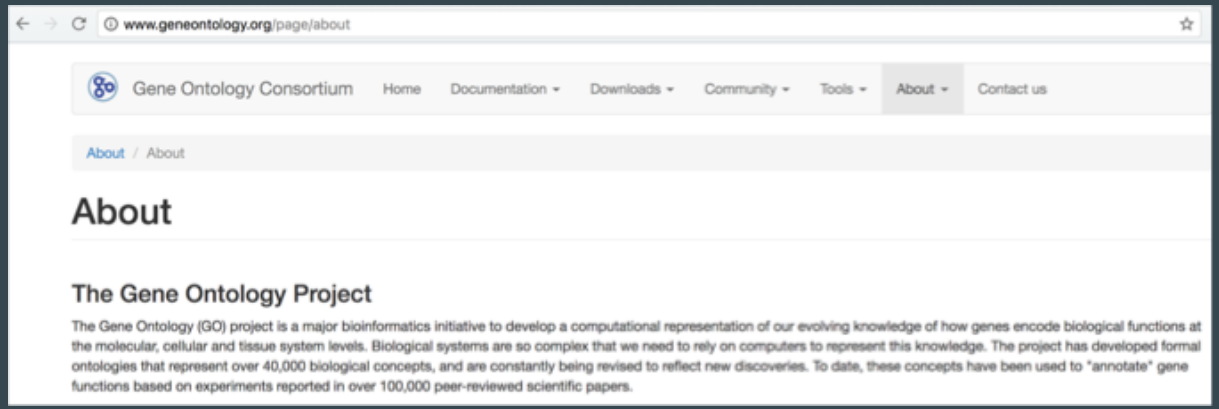
In R:

Available Tools - Gene Ontology Consortium Database and Panther

Consortium supported by a P41 grant from the National Human Genome Research Institute (NHGRI)

Consists of a tree with 6 parent terms: 2 each for ‘Biological Process,’ ‘Molecular Function,’ and ‘Cellular Component’

A single gene can have multiple GO terms in this tree



Gene Functional Enrichment (or GSEA): Results using lowest 1% LogFC genes

GO biological process complete	#	#	expected	Fold Enrichment	+/-	▲ P value
Unclassified	3706	67	52.84	1.27	+	0.00E00
translational initiation	146	32	2.08	15.37	+	1.46E-23
viral gene expression	124	29	1.77	16.40	+	7.53E-22
SRP-dependent cotranslational protein targeting to membrane	94	26	1.34	19.40	+	4.27E-21
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	119	28	1.70	16.50	+	4.57E-21
nuclear-transcribed mRNA catabolic process	195	33	2.78	11.87	+	6.33E-21
ribonucleoprotein complex biogenesis	448	45	6.39	7.05	+	1.77E-20
cotranslational protein targeting to membrane	100	26	1.43	18.24	+	1.98E-20
viral transcription	113	27	1.61	16.76	+	2.20E-20
protein targeting to ER	103	26	1.47	17.71	+	4.11E-20
multi-organism metabolic process	144	29	2.05	14.13	+	4.48E-20
mRNA catabolic process	208	33	2.97	11.13	+	4.53E-20
establishment of protein localization to endoplasmic reticulum	107	26	1.53	17.04	+	1.05E-19
RNA catabolic process	233	34	3.32	10.24	+	1.23E-19
ribosome biogenesis	317	38	4.52	8.41	+	2.27E-19
cellular macromolecule catabolic process	813	57	11.59	4.92	+	2.63E-19
intracellular protein transport	846	58	12.06	4.81	+	3.01E-19
translation	389	41	5.55	7.39	+	4.11E-19
mRNA metabolic process	660	51	9.41	5.42	+	1.13E-18
establishment of protein localization to organelle	384	40	5.47	7.31	+	2.19E-18
peptide biosynthetic process	412	41	5.87	6.98	+	3.27E-18
protein localization to endoplasmic reticulum	127	26	1.81	14.36	+	7.04E-18
RNA processing	868	56	12.38	4.53	+	3.14E-17
protein targeting	425	40	6.06	6.60	+	7.67E-17
rRNA processing	258	32	3.68	8.70	+	2.97E-16
macromolecule catabolic process	948	57	13.52	4.22	+	3.40E-16
protein transport	1386	69	19.76	3.49	+	4.86E-16

So is LogFC a Good Proxy for Essentiality?

Simple answer:

Yes

More complex:

Genes that accelerate uncontrolled proliferation? (e.g. cancer)

Conditional essentiality

GO biological process complete	#	#	expected	Fold Enrichment	+/-	Δ P value
Unclassified	3706	67	52.84	1.27	+	0.00E00
translational initiation	146	32	2.08	15.37	+	1.46E-23
viral gene expression	124	29	1.77	16.40	+	7.53E-22
SRP-dependent cotranslational protein targeting to membrane	94	26	1.34	19.40	+	4.27E-21
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	119	28	1.70	16.50	+	4.57E-21
nuclear-transcribed mRNA catabolic process	195	33	2.78	11.87	+	6.33E-21
ribonucleoprotein complex biogenesis	448	45	6.39	7.05	+	1.77E-20
cotranslational protein targeting to membrane	100	26	1.43	18.24	+	1.98E-20
viral transcription	113	27	1.61	16.76	+	2.20E-20
protein targeting to ER	103	26	1.47	17.71	+	4.11E-20
multi-organism metabolic process	144	29	2.05	14.13	+	4.48E-20
mRNA catabolic process	208	33	2.97	11.13	+	4.53E-20
establishment of protein localization to endoplasmic reticulum	107	26	1.53	17.04	+	1.05E-19
RNA catabolic process	233	34	3.32	10.24	+	1.23E-19
ribosome biogenesis	317	38	4.52	8.41	+	2.27E-19
cellular macromolecule catabolic process	813	57	11.59	4.92	+	2.63E-19
intracellular protein transport	846	58	12.06	4.81	+	3.01E-19
translation	389	41	5.55	7.39	+	4.11E-19
mRNA metabolic process	660	51	9.41	5.42	+	1.13E-18
establishment of protein localization to organelle	384	40	5.47	7.31	+	2.19E-18
peptide biosynthetic process	412	41	5.87	6.98	+	3.27E-18
protein localization to endoplasmic reticulum	127	26	1.81	14.36	+	7.04E-18
RNA processing	868	56	12.38	4.53	+	3.14E-17
protein targeting	425	40	6.06	6.60	+	7.67E-17
rRNA processing	258	32	3.68	8.70	+	2.97E-16
macromolecule catabolic process	948	57	13.52	4.22	+	3.40E-16
protein transport	1386	69	19.76	3.49	+	4.86E-16

Association between SNPs and Gene Essentiality

Hypothesis$_1$: SNPs will be less frequent in more essential genes

Single nucleotide polymorphisms are defined as single base variations that are present at a frequency greater than 1% in the population

In coding regions there are synonymous and non-synonymous SNPs

Non-synonymous SNPs affect the protein sequence

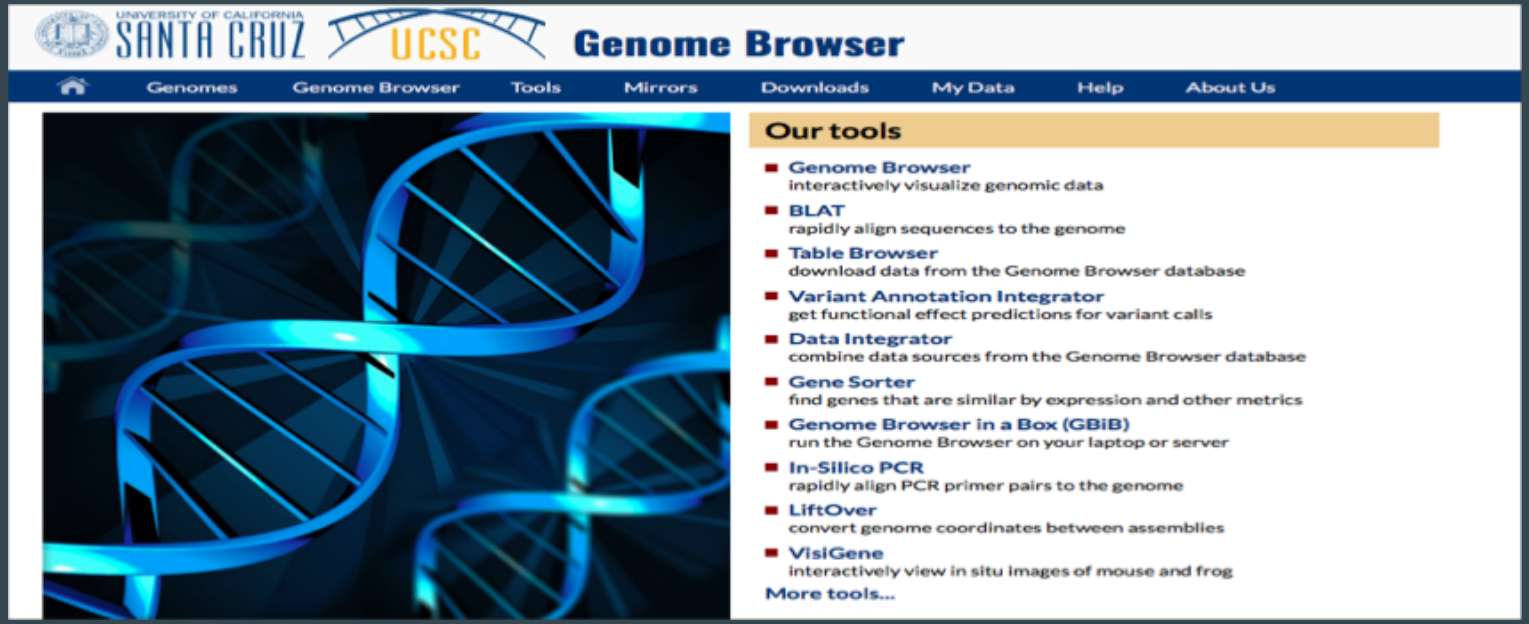
SNPs in non-coding regions can affect splicing , TF binding, and other functions

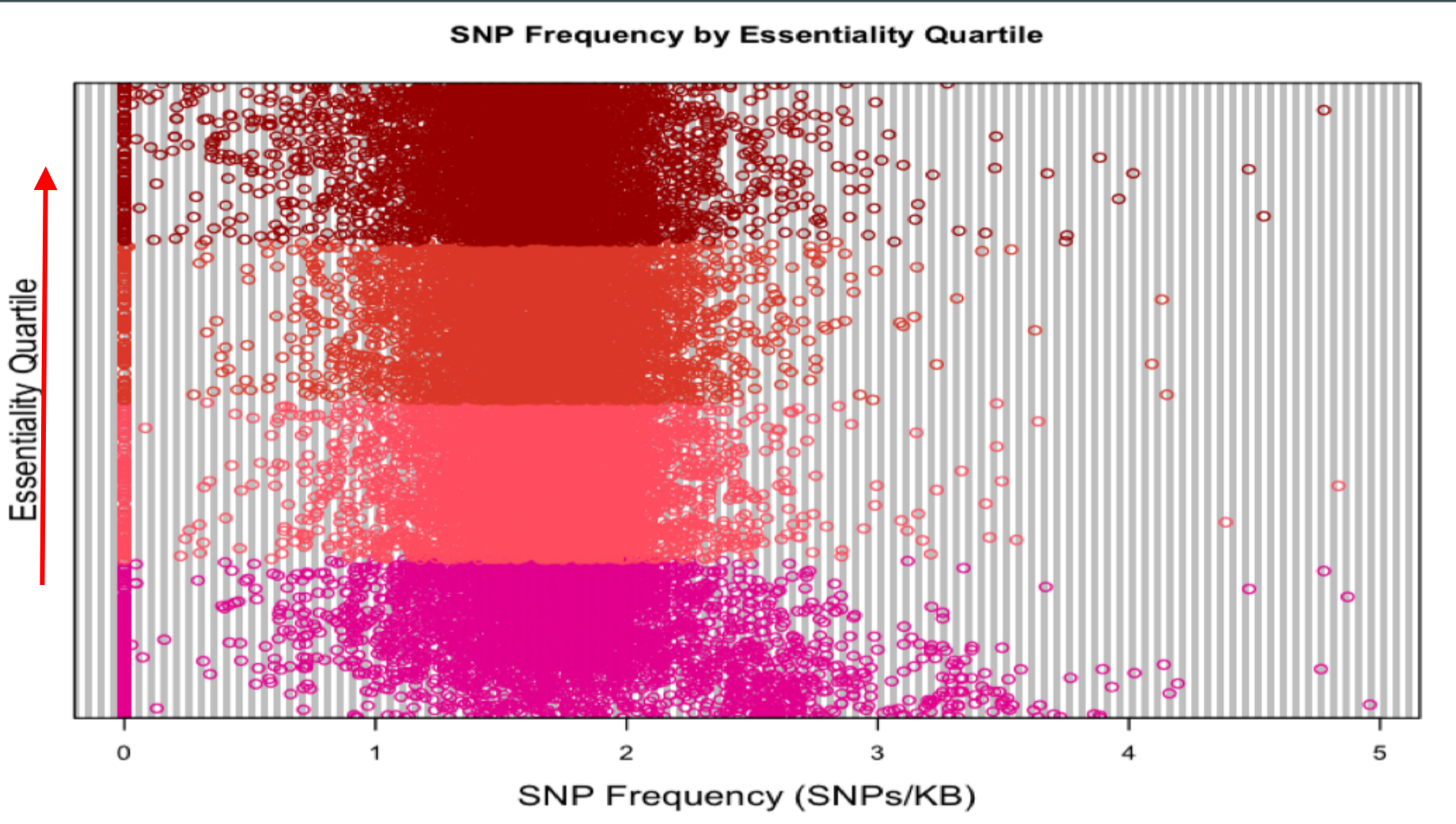
SNP Data from UCSC Genome Browser

SNP data pulled from ‘snpcommon147’ - UCSC database

Intersected SNP information with Log2FCs from GenomeCRISPR data

Genes arranged in quartiles based on LogFC

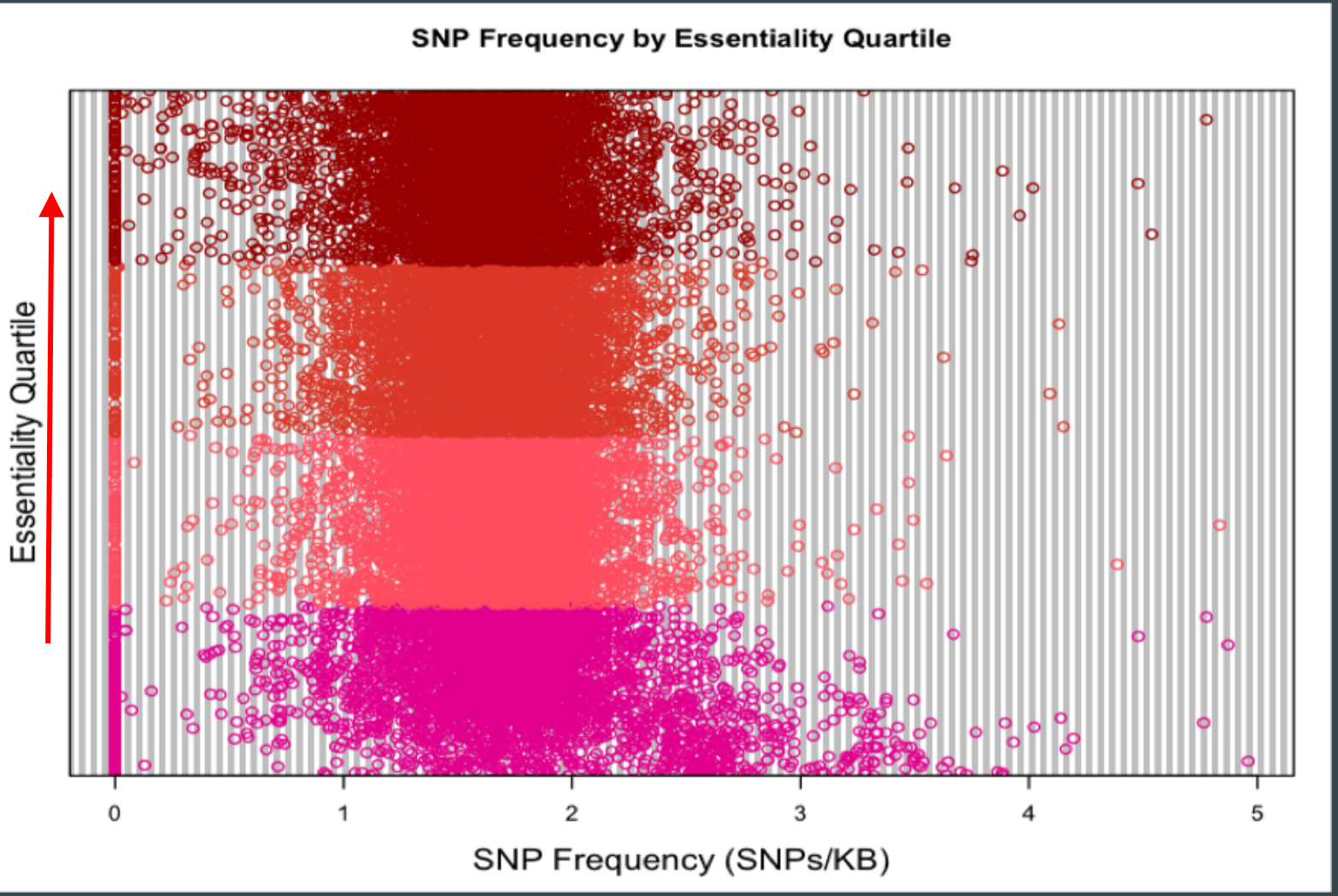



SNP Frequency by Essentiality Quartile
Essentiality Quartile
0
1
2
3
4
5
SNP Frequency (SNPs/KB)

Evaluating the Significance Between Quartiles - An Intro to the Wilcoxon Rank Sum Test

Combined Data	23	31	37	46	49	55	57
Rank	1	2	3	4	5	6	7
Q1 SNP Freq	37	49	55	57			
Rank	3	5	6	7			
Q2 SNP Freq	23	31	46				
Rank	1	2	4				

= 7



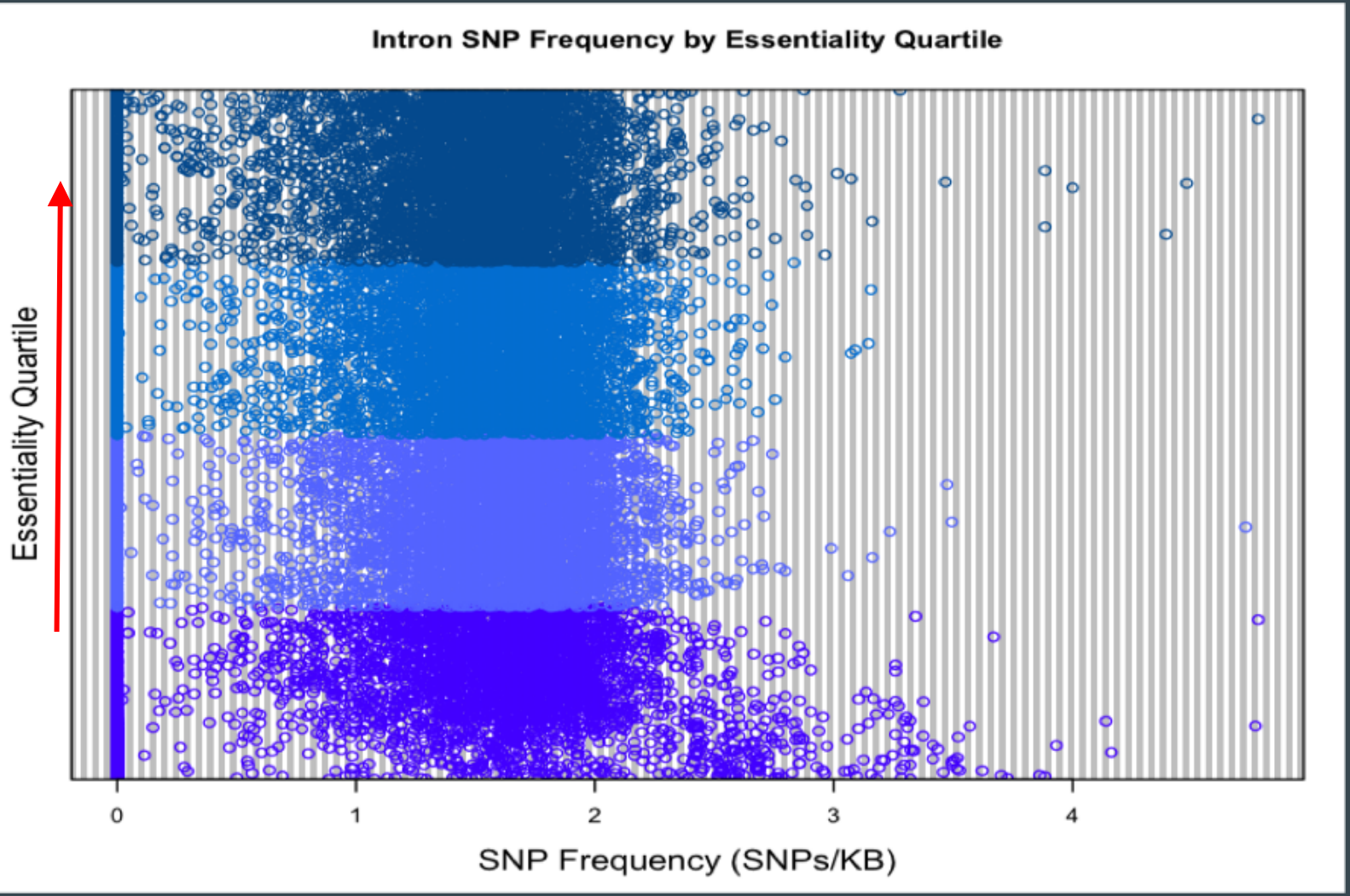
Wilcoxon Rank Sum

Q1 rank sum	Q2 rank sum
14402000	11516000
Q2 rank sum	Q3 rank sum
12578000	13340000
Q3 rank sum	Q4 rank sum
11966000	13947000

Q1 rank sum	Q2 rank sum
14402000	11516000
Q1 rank sum	Q3 rank sum
14146000	11772000
Q1 rank sum	Q4 rank sum
13469000	12444000

	p-value
Q1-Q2	2.20E-16
Q2-Q3	0.006723
Q3-Q4	1.10E-10

	p-value
Q1-Q2	2.20E-16
Q1-Q3	2.20E-16
Q1-Q4	2.20E-16



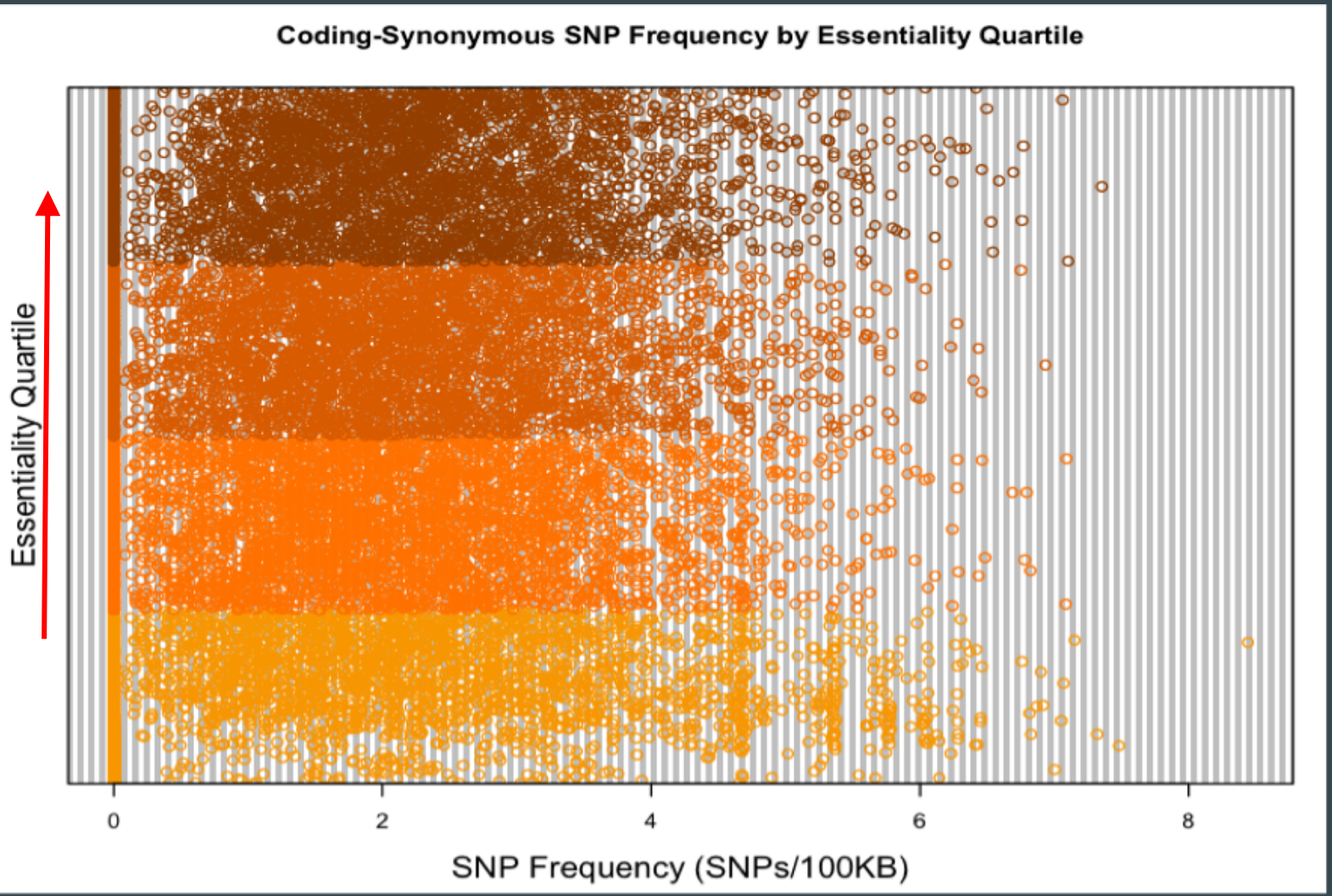
Wilcoxon Rank Sum

Q1 rank sum	Q2 rank sum
15438000	10480000
Q2 rank sum	Q3 rank sum
12557000	13361000
Q3 rank sum	Q4 rank sum
12700000	13913000

Q1 rank sum	Q2 rank sum
15438000	10480000
Q1 rank sum	Q3 rank sum
15185000	10733000
Q1 rank sum	Q4 rank sum
14530000	11383000

	p-value
Q1-Q2	2.20E-16
Q2-Q3	0.006723
Q3-Q4	1.10E-10

	p-value
Q1-Q2	2.20E-16
Q1-Q3	2.20E-16
Q1-Q4	2.20E-16



Wilcoxon Rank Sum

Q1 rank sum	Q2 rank sum
15475000	10443000
Q2 rank sum	Q3 rank sum
13340000	12578000
Q3 rank sum	Q4 rank sum
12580000	13333000

Q1 rank sum	Q2 rank sum
15475000	10443000
Q1 rank sum	Q3 rank sum
15801000	10118000
Q1 rank sum	Q4 rank sum
15329000	10584000

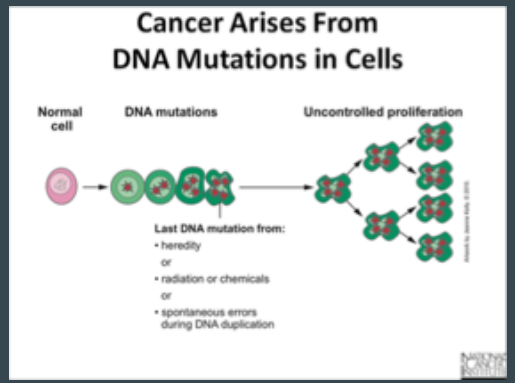
	p-value
Q1-Q2	2.20E-16
Q2-Q3	0.009302
Q3-Q4	1.01E-02

	p-value
Q1-Q2	2.20E-16
Q1-Q3	2.20E-16
Q1-Q4	2.20E-16

Gene Essentiality, SNPs & Cancer

Hypothesis$_1$: Cancer tissues should show a stronger inverse relationship between SNP frequency and gene essentiality

Cancers evolve rapidly and develop mutations, but they still need the cellular machinery essential for life (e.g. DNA/RNA polymerases)



The Cancer SNP Data

- Mutation Annotation Format (MAF) files contain information about mutations in somatic tissue in the general population
- Source: The Cancer Genome Atlas (TCGA)
- Intersected SNP information from MAF files for *various cancers* with Log2FCs from GenomeCRISPR data

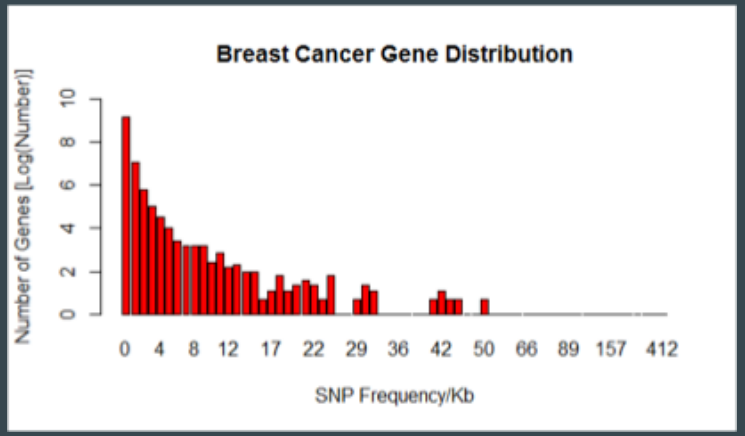
```
> genomeCrispr_genes_with_valid_entrez_ids[1:10,]
   Gene Name Entrez Gene ID 2nd Lowest lg2fc Mean 2nd Lowest lg2fc Median 2nd Lowest lg2fc
1    DPY19L2          283417       3.001230305         3.001230305           3.001230305
2      SSC4D          136853       2.583966520         2.583966520           2.583966520
3      ZNF20            7568       0.649695954         0.649695954           0.649695954
4      OR2A2          442361       0.442805438         0.442805438           0.442805438
5      ZNF57          126295       0.403666385         0.403666385           0.403666385
6      OOSP1          255649       0.390533486         0.390533486           0.390533486
7      PRR13           54458       0.380756527         0.380756527           0.380756527
8       FEN1            2237       0.380241775         0.380241775           0.380241775
9      HOXC6            3223       0.356927933         0.356927933           0.356927933
10   MIR548W       100422923       0.305147176         0.588347317           0.588347317
```

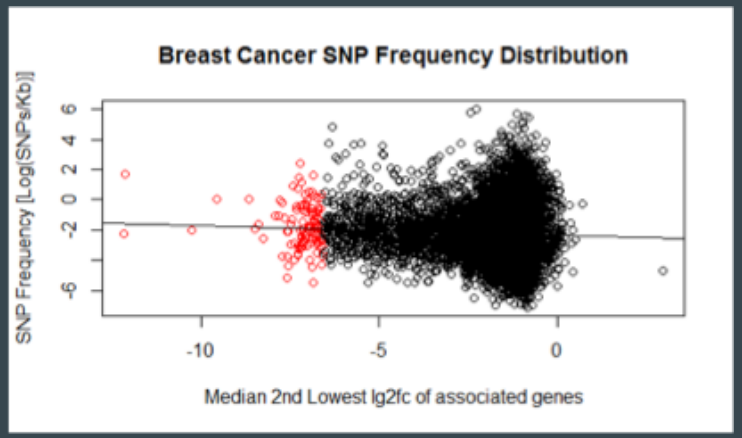
```
> maf_file_breast_cancer[1:10,]
   Entrez_Gene_Id Chromosome Start_position Tumor_Seq_Allele1 Tumor_Seq_Allele2
1            7173          2        1418203                C                G
2            1472         20       23669592                C                T
3           58499          9      109687060                G                A
4          441108          5      131796262                C                T
5            8330          6       27805735                T                C
6           55205         18       56586060                G                C
7           83481          8      144945188                C                G
8          374308         10       27687621                C                T
9            4519         MT          15323                G                A
10          25903          1      161967680                C                T
```

MAF File: A Tab-delimited file containing somatic and/or germline mutation annotations

Breast Cancer SNP to Essentiality Analysis - BRCA

- Normalized SNP counts according to gene length (Megabases)
- Bar chart displays distribution of SNP frequency/Kb
- Scatterplot shows LogFC of each gene vs normalized SNP Frequency/Mb
  - Red points indicate lowest logFC (i.e. most essential)






```
Coefficients:
                                  Estimate Std. Error  t value   Pr(>|t|)
(Intercept)                     -2.3211952  0.0256142 -90.6215 < 2.22e-16 ***
data_brca_asc$Median.2nd.Lowest.Log.Fold.Change -0.0593024  0.0126956  -4.6711 3.0285e-06 ***
```

Potential Biases in SNP Data

Not adjusted for allele frequency

No correction for ‘Essentiality Bias’

- Essential genes are likely to be well-studied, and SNPs and/or mutations of these genes would be better documented

Projecting LogFC data from GenomeCRISPR onto specific types of cancers disregards insights on conditional essentiality

- Only patterns about core essential genes can be inferred

Things We'd Do Differently

Avoid a false start on Random Forests

Go deeper into nonparametric hypothesis testing.

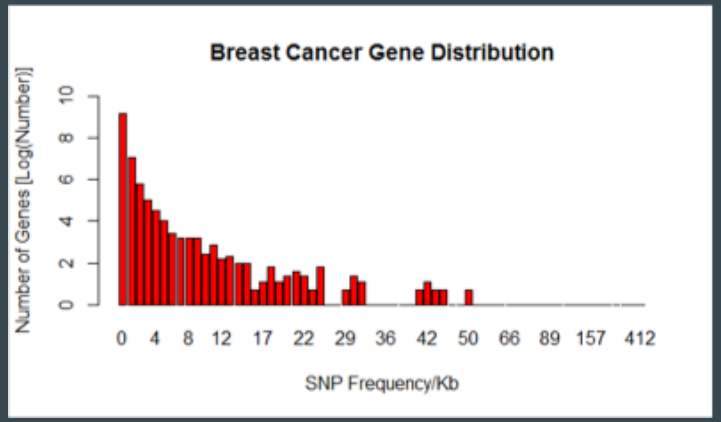
Explore other genomic features based on essentiality

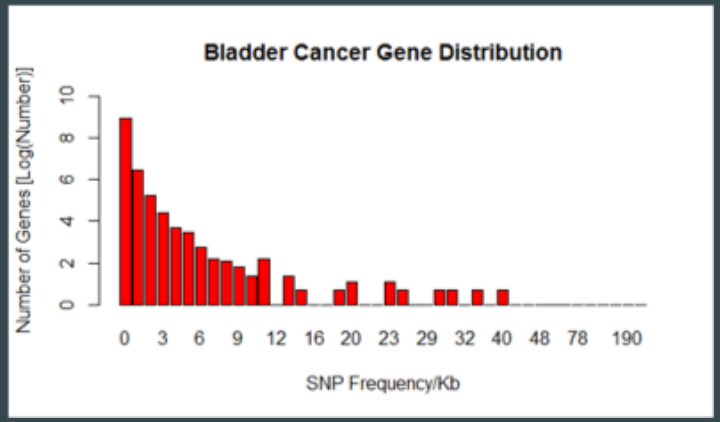
- E.g. Protein-protein interactions, network-based essentiality, etc

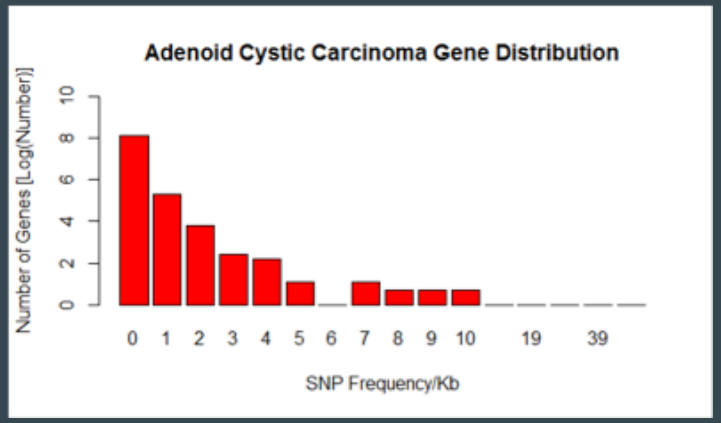
Thank You!

Appendix Slides:
Things we did but ran out of time to present on

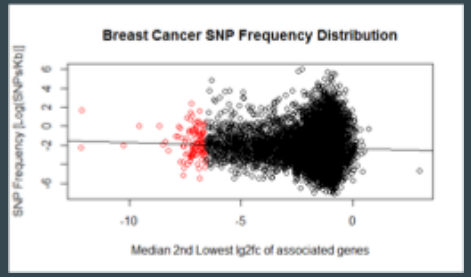
SNP vs LogFC Charts For Other Cancers Show The Same Trend



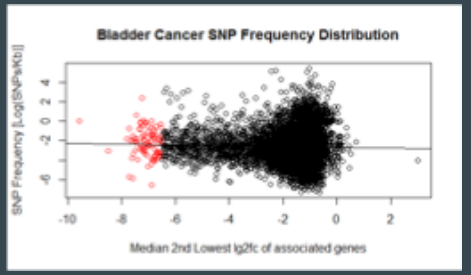




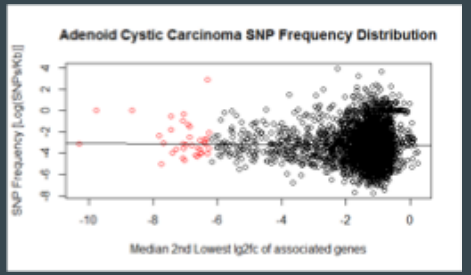
SNP vs LogFC Charts For Other Cancers Show The Same Trend




```
Coefficients:
                                                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)                                    -2.3211952  0.0256142 -90.6215 < 2.22e-16 ***
data_brca_asc$Median.2nd.Lowest.Log.Fold.Change -0.0593024  0.0126956  -4.6711 3.0285e-06 ***
```




```
Coefficients:
                                                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)                                    -2.7051992  0.0297600 -90.90063 < 2.22e-16 ***
data_blca_asc$Median.2nd.Lowest.Log.Fold.Change -0.0397544  0.0147732  -2.69098  0.0071379 **
```




```
Coefficients:
                                                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                    -3.2706365  0.0468892 -69.75247  < 2e-16 ***
data_acc_asc$Median.2nd.Lowest.Log.Fold.Change -0.0193308  0.0256091  -0.75484  0.45039
```