CORRELATED SOMATIC & GERMLINE RARE VARIATION IN IBC COHORT

Hussein Mohsen | CBB Winter Rotation | Gerstein Lab
Mentor [Sushant Kumar] | PI [Mark Gerstein]

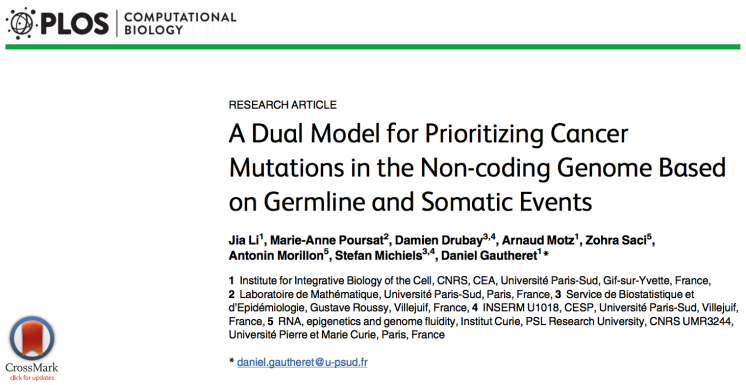
April 4, 2017
(& March 31 in CBB Rotation Talks at 300 Georges St.)

OUTLINE

- Background
- Problem Statement
- Data
- Methods
  - *Pipeline*
  - *Scripts*
- Results
- Learning Outcome & Challenges
- Near Future Work

BACKGROUND

- Two-hit hypothesis
  - Tumor suppressor genes inactivated by both germline & somatic mutations
- Recent papers
  - Focused on germline variation
  - Adopted separate model for each of germline and somatic variations



PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A Dual Model for Prioritizing Cancer Mutations in the Non-coding Genome Based on Germline and Somatic Events

Jia Li[1], Marie-Anne Poursat[2], Damien Drubay[3,4], Arnaud Motz[1], Zohra Saci[5], Antonin Morillon[5], Stefan Michiels[3,4], Daniel Gautheret[1]*

1 Institute for Integrative Biology of the Cell, CNRS, CEA, Université Paris-Sud, Gif-sur-Yvette, France, 2 Laboratoire de Mathématique, Université Paris-Sud, Paris, France, 3 Service de Biostatistique et d'Epidémiologie, Gustave Roussy, Villejuif, France, 4 INSERM U1018, CESP, Université Paris-Sud, Villejuif, France, 5 RNA, epigenetics and genome fluidity, Institut Curie, PSL Research University, CNRS UMR3244, Université Pierre et Marie Curie, Paris, France

* daniel.gautheret@u-psud.fr

Article | OPEN

Patterns and functional implications of rare germline variants across 12 cancer types

Charles Lu, Mingchao Xie […] Li Ding

Nature Communications **6**,
Article number: 10086 (2015)
doi:10.1038/ncomms10086

Received: 20 July 2015
Accepted: 02 November 2015
Published online: 22 December 2015

PROBLEM STATEMENT

- Studying patterns of correlation between rare somatic and germline cancer variations
  - Combining both signals
  - Inference of correlation patterns
  - Prediction of correlation incidence

DATA

- Inflammatory Breast Cancer (IBC) Cohort
- We used 17 of the IBC samples
- More than 4 million (rare & common) variants per sample on average

METHODS

PIPELINE

METHODS

PIPELINE

- ***PERTAIN RARE VARIANTS NOT FOUND IN DATABASES OR FOUND WITH ALLELE FREQUENCY <0.15%***
- ***ON AVERAGE, ~40-45% OF VARIANTS ARE RARE***
- ***PARTITION BOTH SAMPLES & DATABASES***

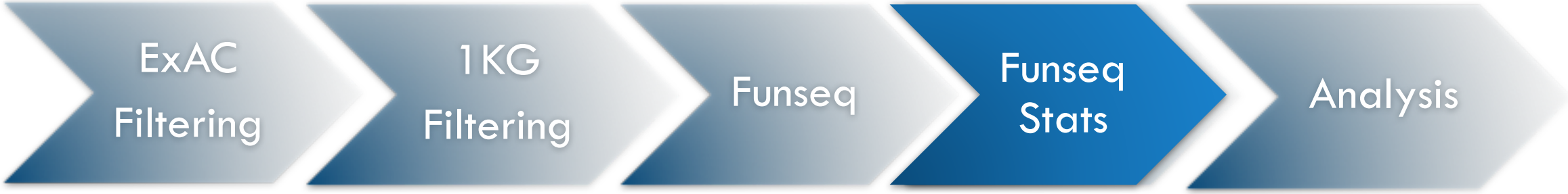
METHODS

PIPELINE

- ***GENERATE FUNSEQ ANNOTATION***

METHODS

PIPELINE



- ***GENERATE STATISTICS ON DISTRIBUTIONS W.R.T. CHROMOSOMES, SITES, NONCODING FEATURES, GENES, HUBS, AND OTHER FEATURES.***

METHODS

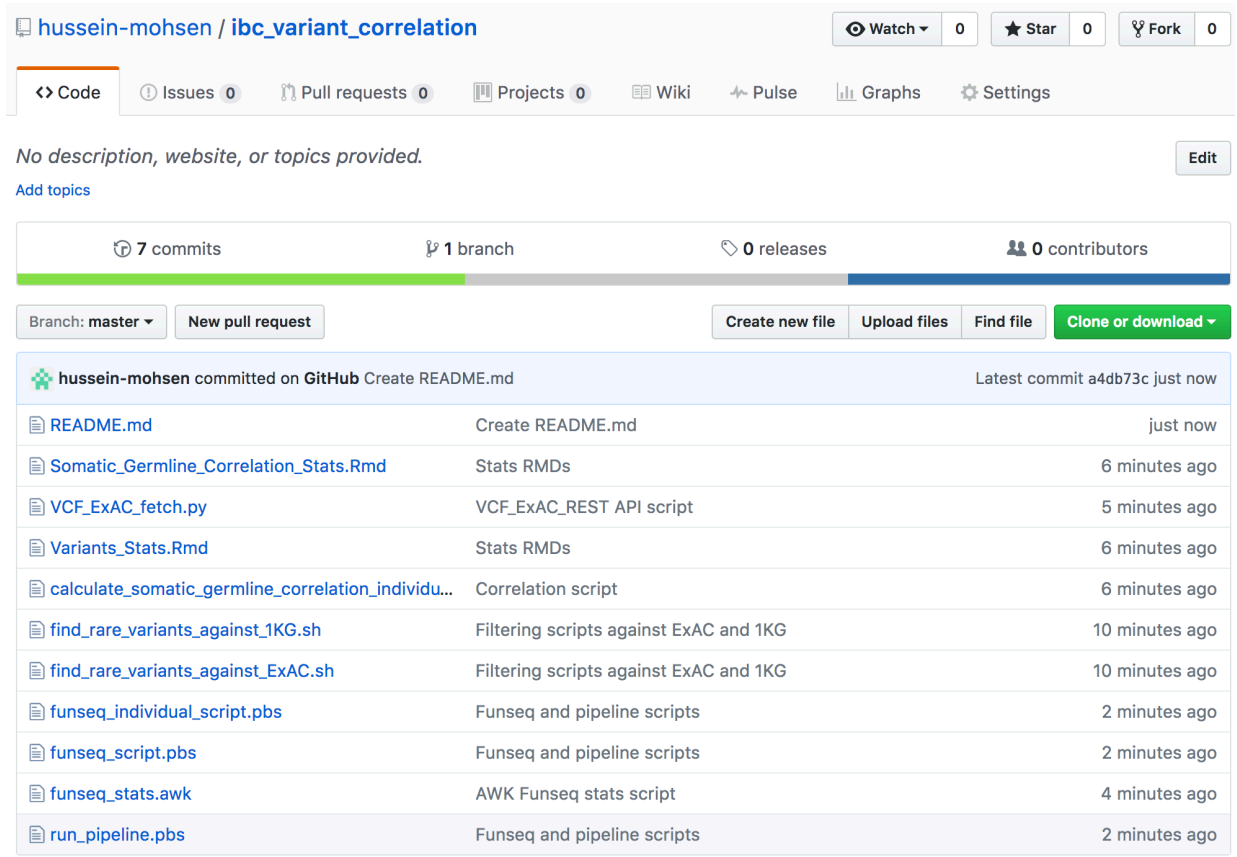
PIPELINE

- ***GENERATE RESULTS***
- ***TO BE FOR DATA GENERATION***

METHODS

SCRIPTS

```
git clone https://github.com/hussein-mohsen/ibc_variant_correlation.git
```



hussein-mohsen / ibc_variant_correlation

Watch 0 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Pulse | Graphs | Settings

No description, website, or topics provided. Edit

Add topics

7 commits | 1 branch | 0 releases | 0 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

hussein-mohsen committed on **GitHub** Create README.md — Latest commit a4db73c just now

README.md	Create README.md	just now
Somatic_Germline_Correlation_Stats.Rmd	Stats RMDs	6 minutes ago
VCF_ExAC_fetch.py	VCF_ExAC_REST API script	5 minutes ago
Variants_Stats.Rmd	Stats RMDs	6 minutes ago
calculate_somatic_germline_correlation_individu...	Correlation script	6 minutes ago
find_rare_variants_against_1KG.sh	Filtering scripts against ExAC and 1KG	10 minutes ago
find_rare_variants_against_ExAC.sh	Filtering scripts against ExAC and 1KG	10 minutes ago
funseq_individual_script.pbs	Funseq and pipeline scripts	2 minutes ago
funseq_script.pbs	Funseq and pipeline scripts	2 minutes ago
funseq_stats.awk	AWK Funseq stats script	4 minutes ago
run_pipeline.pbs	Funseq and pipeline scripts	2 minutes ago

METHODS

SCRIPTS

- Funseq stats script easily generalizable (will talk to Shaoke)
- Chromosome, location, HUB, GENE, PPI, Coding & Noncoding features, and others.
- Flexible, accepts custom regex and new features

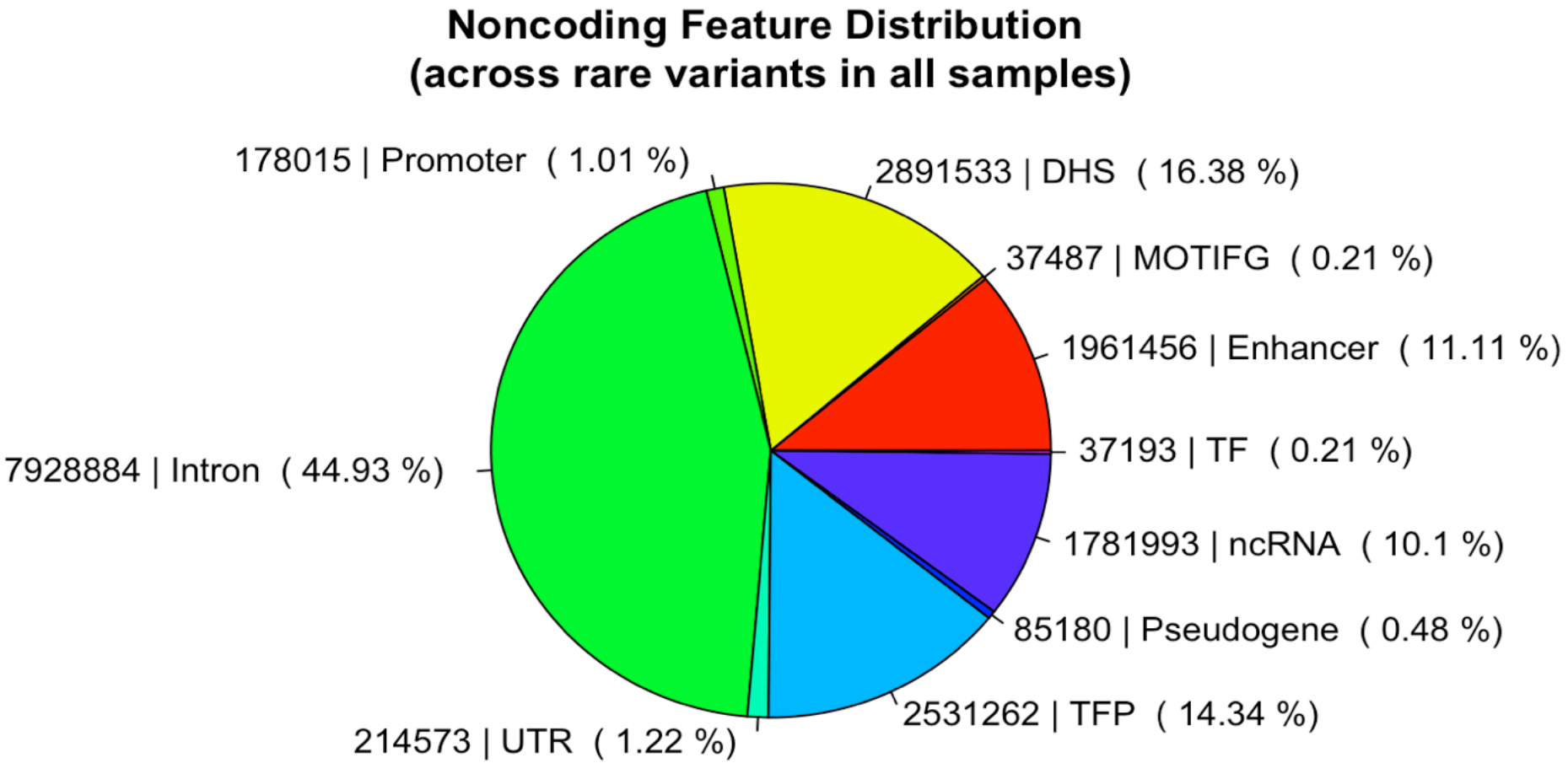
```
$ awk -f funseq_stats.awk -v regex_args="Promoter" funseq2/out_I/Output.vcf
```

- Results interpretable with R

```
NCDS 1.01027063155683 3 2.4983e-06
NCDS 0.482556148956296 1 8.32766e-07
NCDS 0.800092844000001 8 6.66213e-06
NCDS 0.800092844000002 10 8.32766e-06
NoncodingFeatures Enhancer 108027 0.111836
NoncodingFeatures DHS 158652 0.164246
NoncodingFeatures Promoter 9853 0.0102004
NoncodingFeatures Intron 433008 0.448275
NoncodingFeatures UTR 11544 0.011951
NoncodingFeatures MOTIFG 2098 0.00217197
NoncodingFeatures TFP 138879 0.143776
NoncodingFeatures Pseudogene 4528 0.00468765
NoncodingFeatures ncRNA 97205 0.100632
NoncodingFeatures TF 2148 0.00222374
OtherFeatures nonsynonymous 330 0.0198592
OtherFeatures synonymous 226 0.0136005
OtherFeatures cancer 15496 0.932539
OtherFeatures prematureStop 8 0.000481435
OtherFeatures + 292 0.0175724
OtherFeatures - 265 0.0159475
```

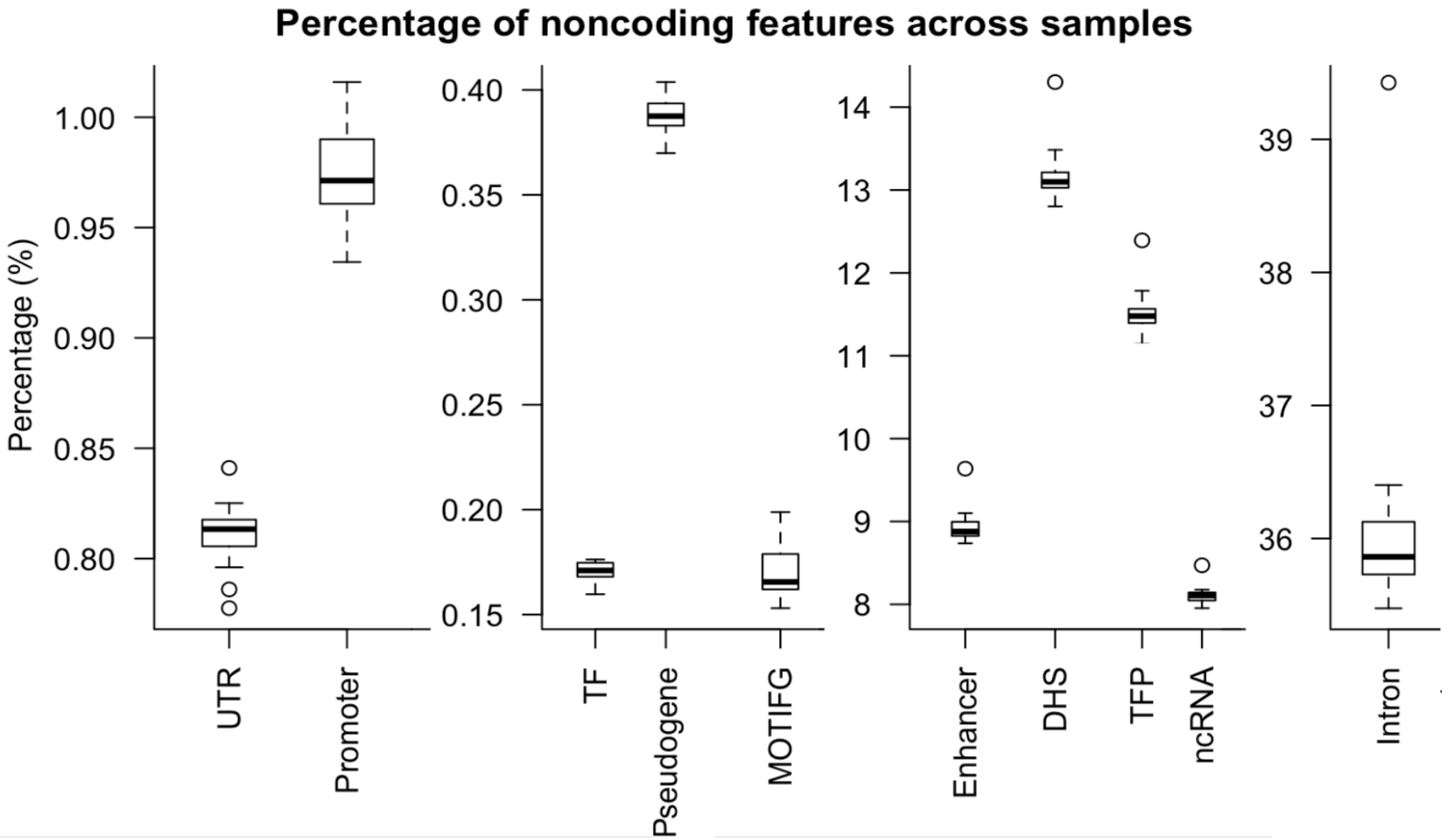
RESULTS

NONCODING FEATURE DISTRIBUTION



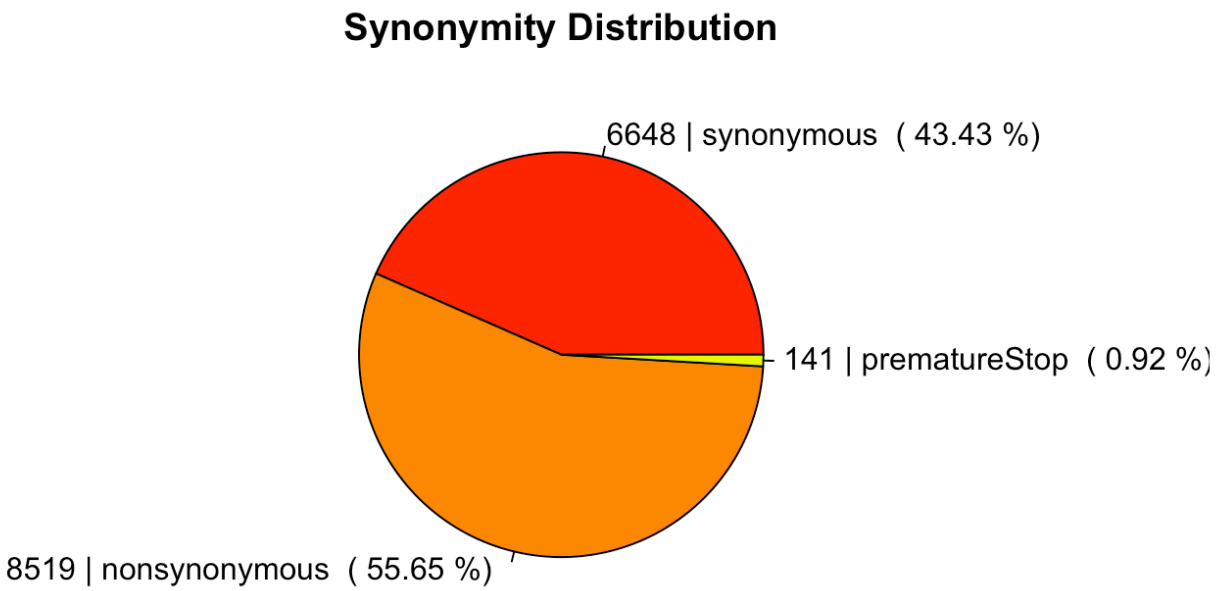
RESULTS

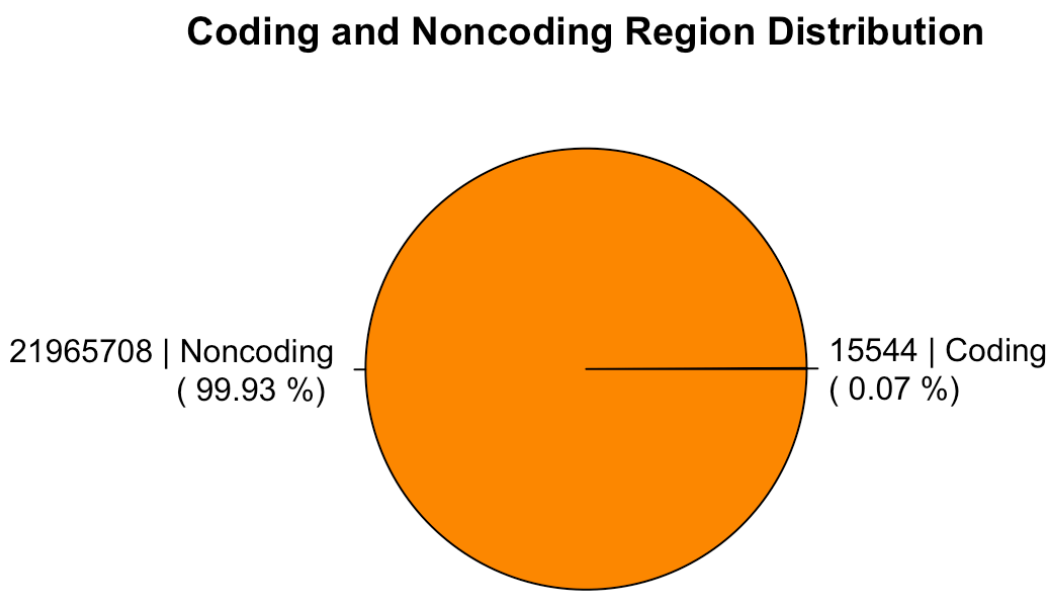
NONCODING FEATURE DISTRIBUTION



RESULTS

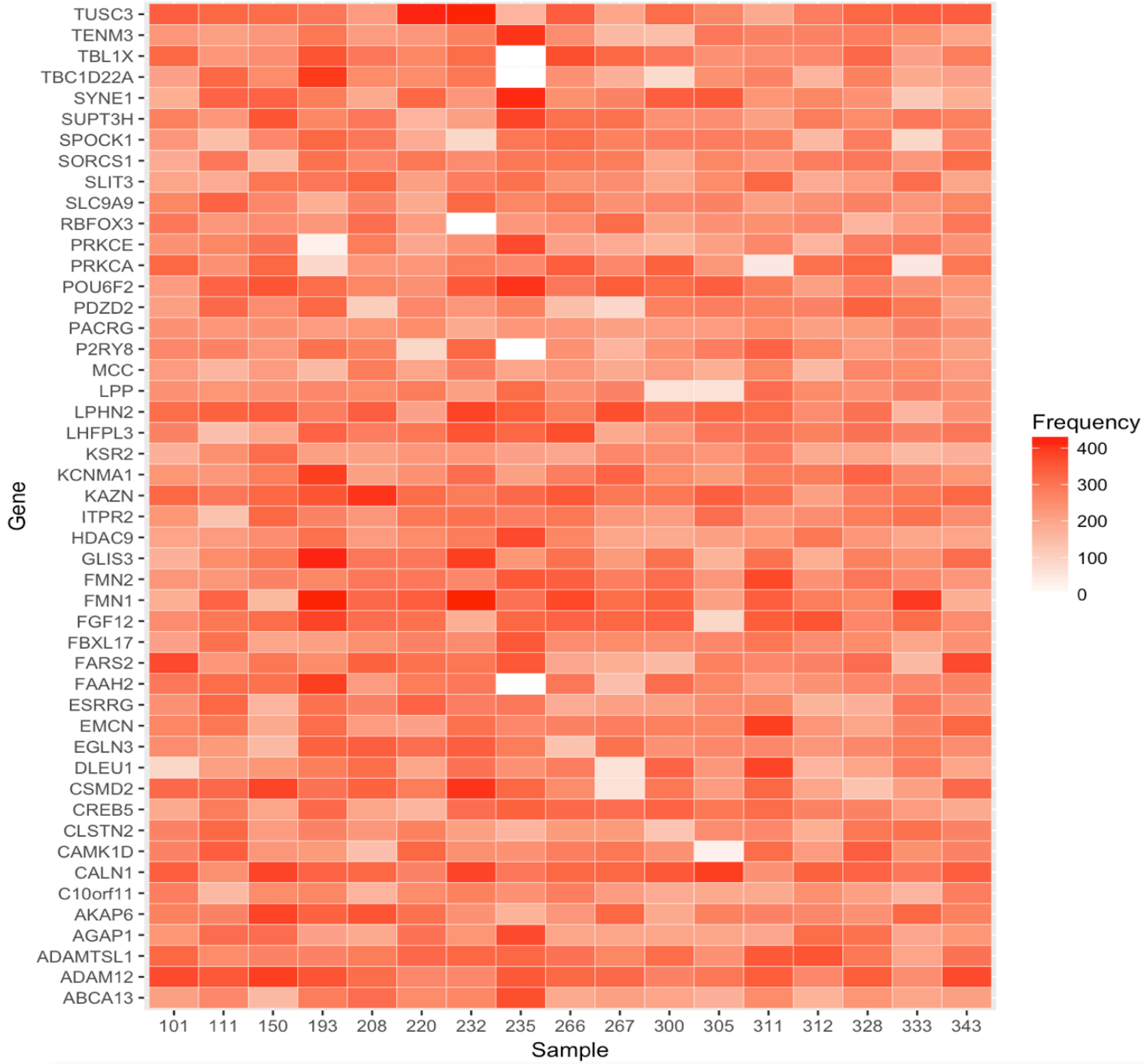
SYNONYMITY & CODING REGION DISTRIBUTION





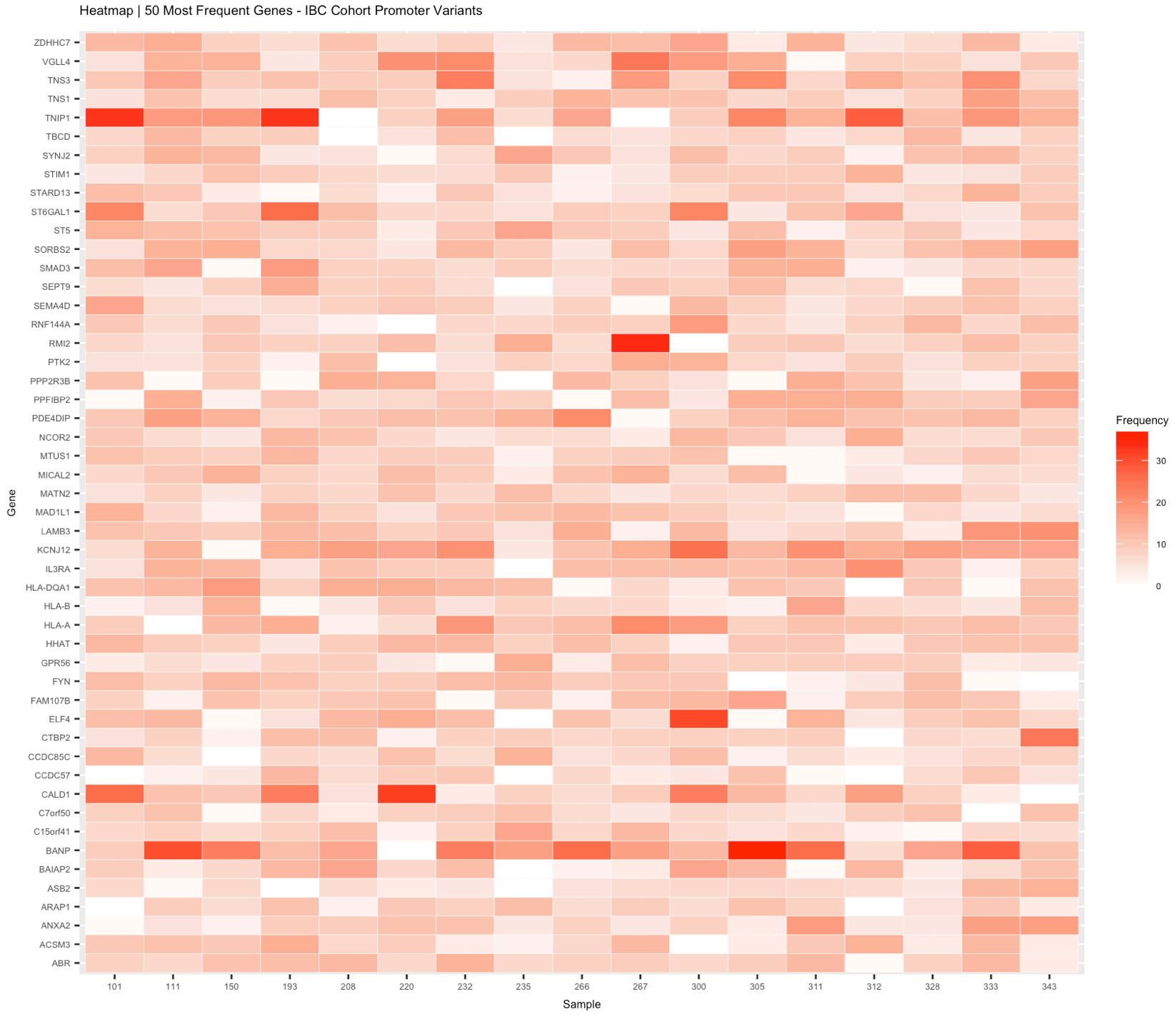
RESULTS

50 MOST FREQUENT GENES
RARE VARIANTS



RESULTS

50 MOST FREQUENT GENES
RARE PROMOTER VARIANTS

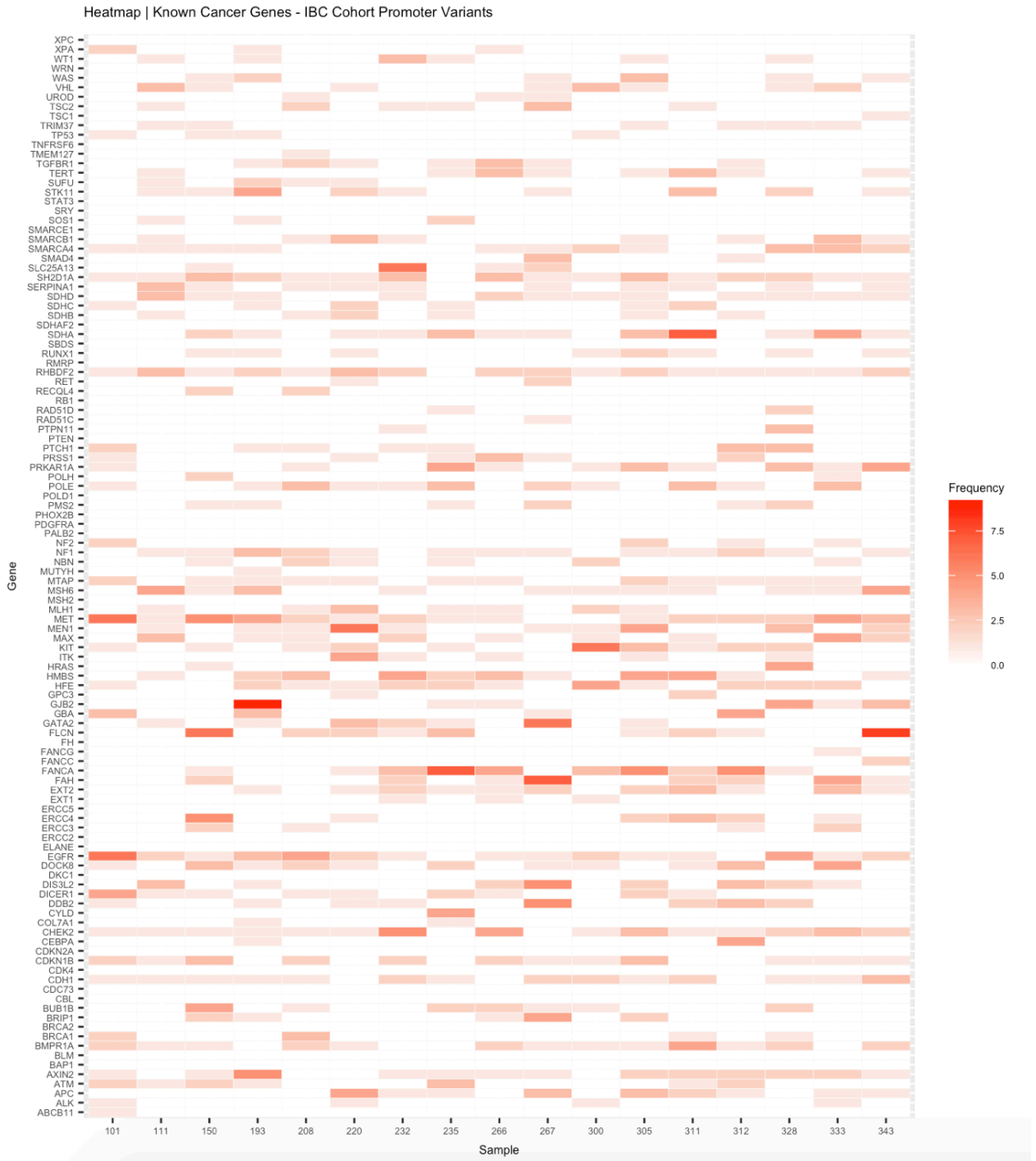


RESULTS

KNOWN CANCER GENES | IBC RARE VARIANTS

RESULTS

KNOWN CANCER GENES | IBC RARE PROMOTER VARIANTS



SOMATIC-GERMLINE CORRELATION

- A somatic-germline correlation is a measure of the co-occurrence of both kinds of mutations
- At the gene level, frequency of co-occurrence
- Might be at other levels (hubs, coding/noncoding regions, etc.)
- Might be in terms of ratios instead of frequencies

RESULTS

SOMATIC-GERMLINE VARIATION
50 MOST FREQUENT GENES

RESULTS

SOMATIC-GERMLINE VARIATION
KNOWN CANCER GENES | IBC

LEARNING OUTCOME & CHALLENGES

- AWK, REGEX, text processing on UNIX Systems
- Bedtools, VCF & BED, and REST APIs
- Better understanding of cancer genomics

- Louise > Grace > Farnam transitions
- Memory limits and disk over-quotas
  - Solved through experimentation with DB and sample partitions

NEAR FUTURE WORK

- Further analysis
  - Genes, hubs, location, chromosome
- Data Generation
  - 17,000+ x 17 samples x variation type (somatic vs germline)
- Machine Learning problem(s) definition
  - Which gene/variant features should be considered?
  - Classification vs regression problem
    - Will correlation occur vs to what extent it would?
  - How many models?
  - Which models?
    - Parametric? Nonparametric? Deep Learning (with Dropout to combine both somatic and germline)?
    - Frey *et al.* (2015) only attempted DL so far
  - Prediction Task
    - How likely is a given sequence/gene going to have somatic-germline variation correlation?
    - How likely is a given sample expected to have variation correlation?
    - How malignant would the effect of correlation be?

END