COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Using Clustering Methods on High-Dimensional Genomic Data in Cancer: Two Projects

February 12, 2018

Margaret L. Hannum
MS Candidate, Department of Biostatistics, Columbia University

Presentation Overview

- Breast Cancer Subtype Classification Project
  - Background
  - Goals of project
  - Results
  - Further analysis
- Nonsense Mutation Clustering Project
  - Overview & algorithm development

Validation and Refinement of Breast Cancer Subtype Classifications using Multi-Modal Genomic Data

Mentors: Ronglai Shen, PhD, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer

&

Shuang Wang, PhD, Department of Biostatistics, Columbia University

Research motivation

- Identification of precise disease subtypes can have treatment and survival implications
- Availability of multiple genomic data types in breast cancer tumors open new options for identifying new subtypes

Background

Cancer Subtypes

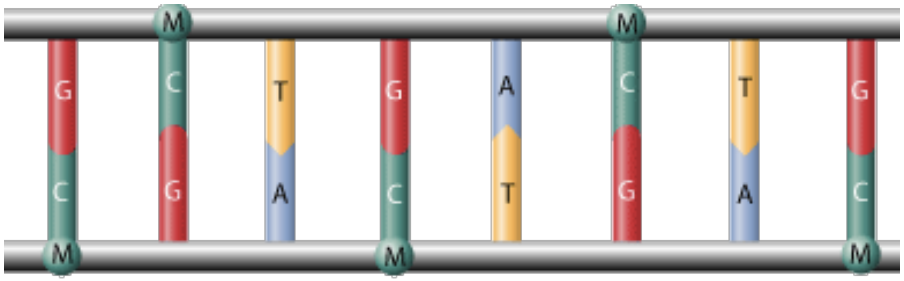
- Cancer is a heterogeneous disease
  - Example: Breast cancer (Estrogen Receptor (ER), Progesterone (PR), HER2, Basal)
- Subtypes have clinical implications (treatment and survival)
  - Luminal A: ER+, HER2-
  - Luminal B: ER+, either HER2+ or –
  - HER2 type: most are HER2+, ER-, PR-, Lymph-node+
  - Triple Negative Breast Cancer (TNBC)/Basal-like: ER-, PR-, HER2-
- Massive whole exome and whole genome sequencing efforts have changed the way we can discover new subtypes

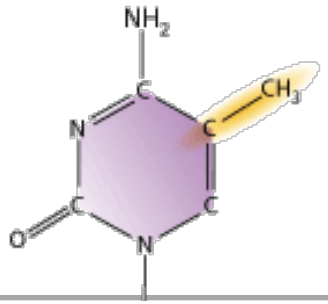
Breast Cancer Subtype Classification Background

- METABRIC Study (Curtis et al., Nature, 2012)
  - Identified 10 Breast Cancer (BRCA) subtypes by integrating copy number and expression data from 2,000 breast tumor samples
  - Identified new therapeutic targets including 11q13 copy number amplification
  - Identified new subtypes including a copy number quiescent subgroup that show favorable prognosis
  - Validated in a combined external patient cohort of over 7,500 tumor samples (Ali et al., Genome Biology, 2014)

Genomic Data Types for This Study

- Copy Number Alteration (CNA): on DNA level
- Gene Expression: on mRNA level
  - High-dimensional platforms such as Expression arrays, Next Generation sequencing (NGS)
- DNA Methylation
  - Illumina methylation arrays 27K, 450K, 850K platforms
  - Differentially Methylated Loci (DML) vs. Regions (DMRs)





DNA methylation is the addition of a methyl group (M) to the DNA base cytosine (C).

http://helicase.pbworks.com/w/page/17605615/DNA%20Methylation

Research question

- Can adding layer of DNA methylation data to copy number and gene expression data in an integrative clustering algorithm further refine breast cancer subtypes?

Our Project: Validate and Refine BRCA Subtypes using independent sample

1. *METABRIC Subtype Validation*
2. *DNA methylation dimension reduction*
3. *Subtype Refinement by incorporating DNA methylation with Copy Number and Gene Expression data*

Independent Dataset

- BRCA samples from **The Cancer Genome Atlas (TCGA)**
- **Copy number (CN) data** (segmented),
  - Affymetrix 6.0 Single Nucleotide Polymorphism (SNP) arrays (25,000+ genes, n = 1080)
- **mRNA Gene expression (GE) data**
  - mRNA-seq (16,000 genes, n = 960)
- **DNA methylation data**
  - Infinium HumanMethylation450 BeadChips (450K) (480,000+ CpG sites, n = 646)

1. Subtype Validation

a. Use previous classifiers from METABRIC subgroups to predict subgroup assignments in TCGA sample

b. Assess goodness of assignments

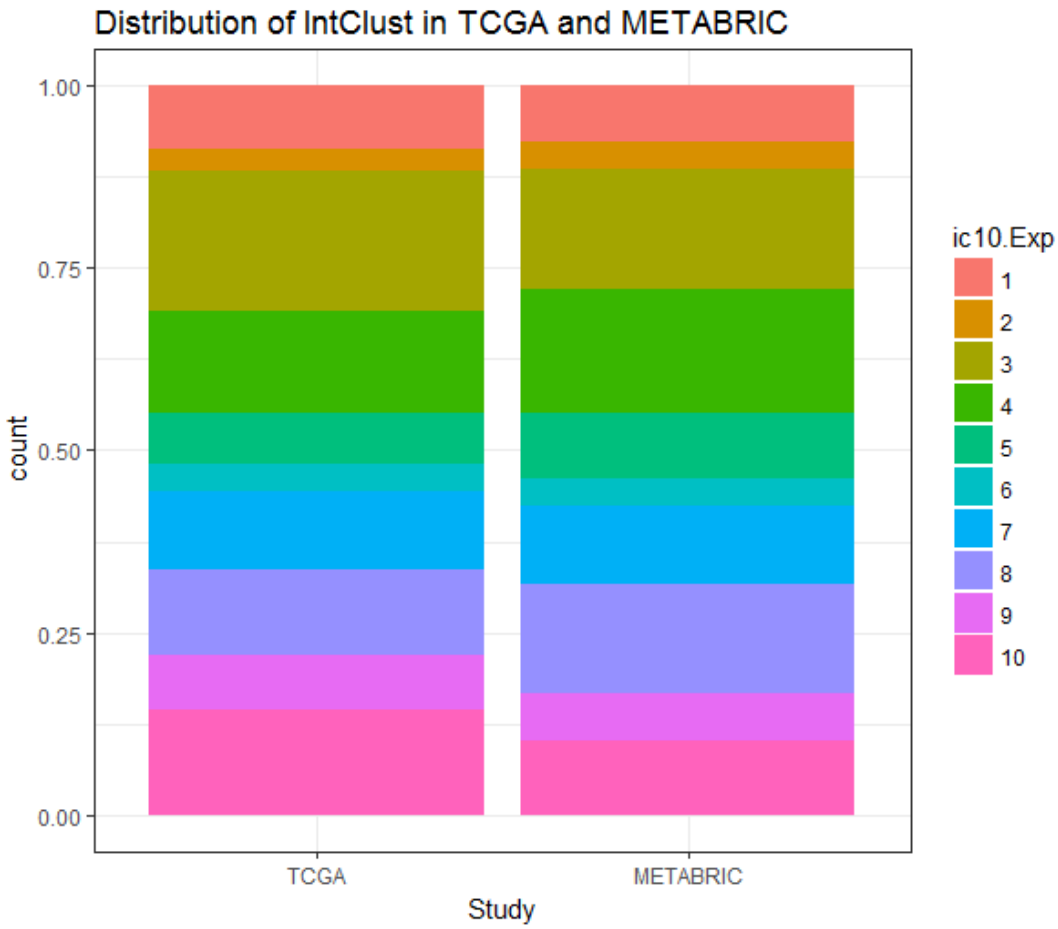
c. Train new classifiers, compare assignments

2. Dimension reduction

3. Subtype Refinement

METABRIC IntClust Validation

- Predict class assignments of input data (n = 960) using trained classifiers from METABRIC study (ic10 package) after matching expression data to 584 of 612 selected genes
- All 10 subtypes were identified in TCGA sample…but were they good assignments?



Assessing Goodness of Assignment

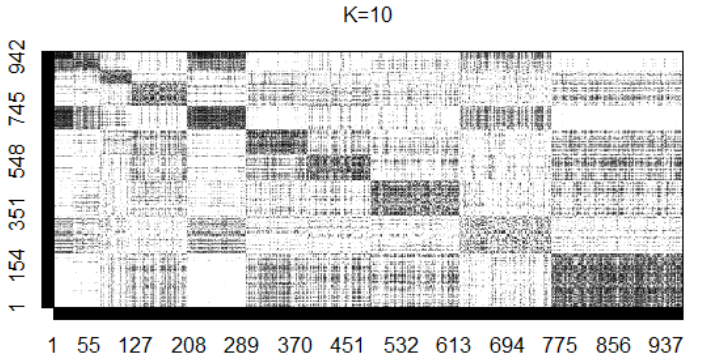
- ic10 package generated a correlation statistic between the signatures of the training dataset and of the classified features (overall correlation was 0.853, highest group ic9 = 0.989, and lowest ic4 = 0.699)
- Also ran in-group proportion analysis (Kapp & Tibshirani) which assesses if clusters in one group are found in another
  - 2000 permutations and centroids from original training set
  - 5 groups had acceptable in-group proportion based on this strict assessment

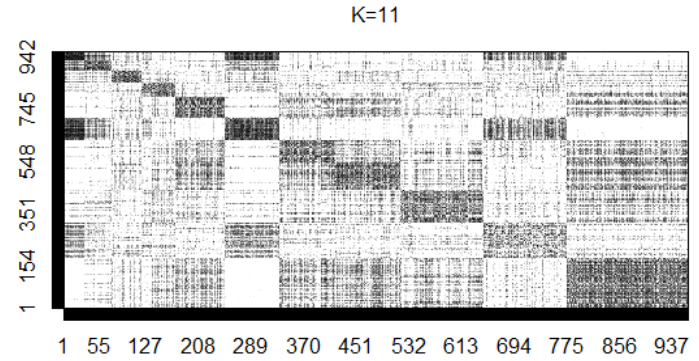
Train new classifiers with TCGA

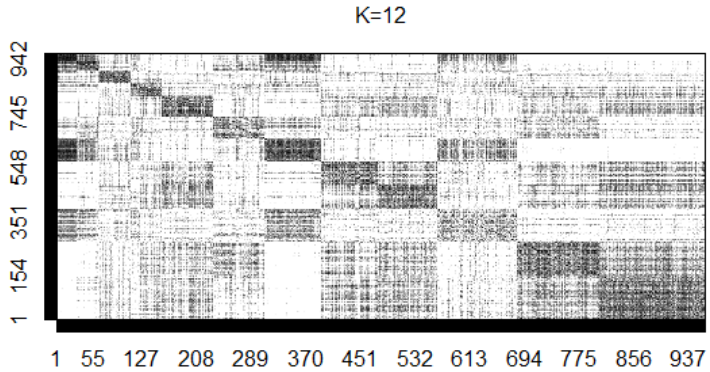
- **Goal:** Use CN and GE data from TCGA to train new cluster centroids, then compare to METABRIC assignments
  - Matched CN features highly correlated with GE (**542 genes, n = 960)**
- **iCluster** is an algorithm for integrative clustering developed by Shen et al (2012)
  - Joint latent variable model

CN and GE Cluster results

- **Integrative clustering on CN and GE data (**iClusterPlus package) with following settings: K=3 through 20, and lasso shrinkage method
  - Decided on K=11 or 12 clusters (Rand Index 0.818 concordance with METABRIC predicted assignments)
  - Use this and predicted assignments from METABRIC to compare after we add the methylation layer to clustering







1. Subtype Validation

2. Dimension reduction

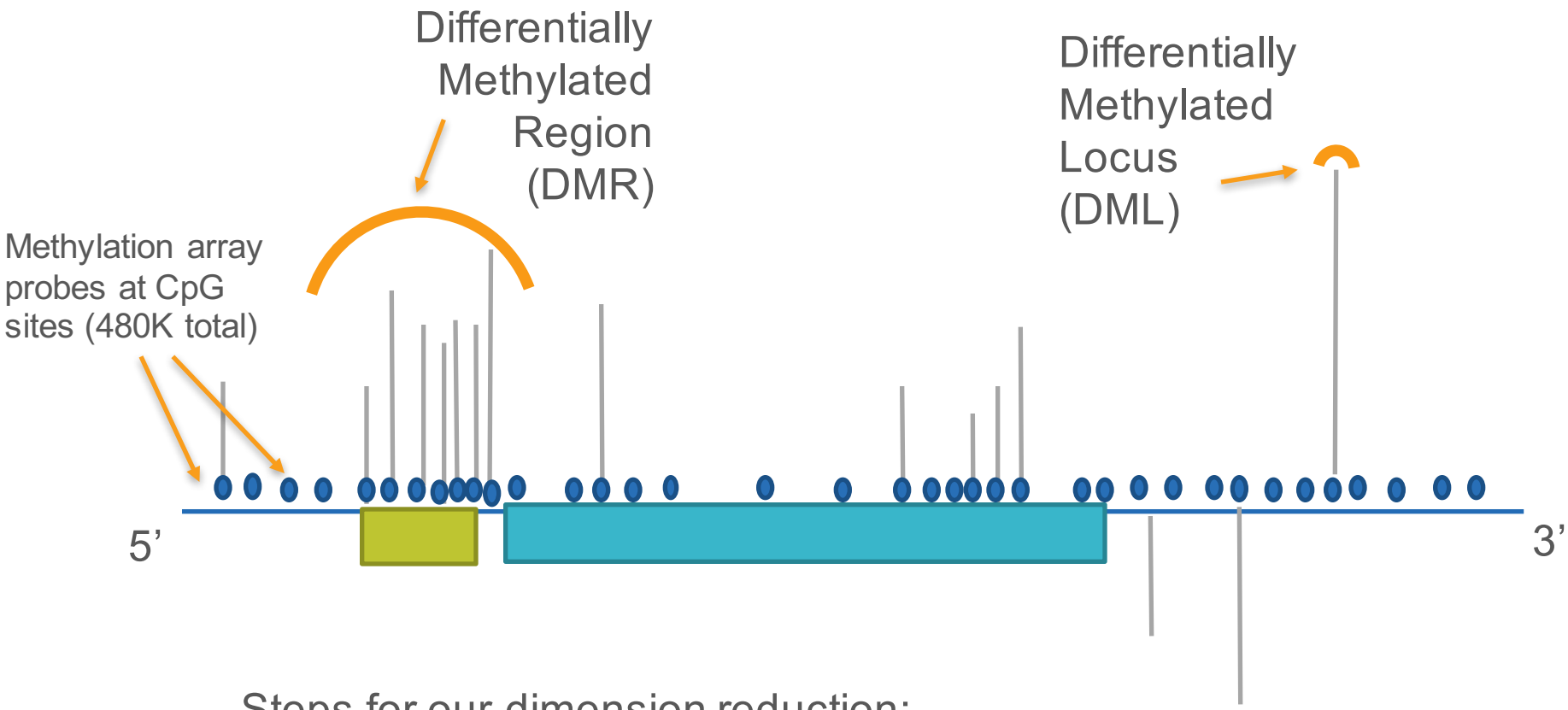
a. Clean data

b. Identify differentially methylated regions

c. Select probes highly negatively correlated with gene expression

1. Subtype Refinement

A little more about methylation data



Steps for our dimension reduction:

1. Clean data (480K → 370K)
2. Identify differentially methylated regions, and probes within those regions (370K → 70K)
3. Select top probes correlated with gene expression (70K → 1.2K)

Methylation Data Preparation

- 480K Probes, 646 total tumor samples, 90 tumor-matched normal
- Data were cleaned using R packages ChAMP (Morris) and minfi (Aryee) with following measures:
  - CpG sites on sex chromosome/missing chromosome information or overlapping with known SNPs were removed; CpG sites with <70% coverage across samples or samples with <95% coverage across CpGs were removed (6 pairs removed)
  - Type II probe bias was corrected using BMIQ (Teschendorff)
  - Performed kNN imputation (with k = 10) to address missing data
  - Final dataset 326,105 autosomal probes

DMR Identification

- 3 possible approaches in the literature
  - Bumphunter
  - Probe Lasso
  - DMRCate

Selection of Methylation Probes

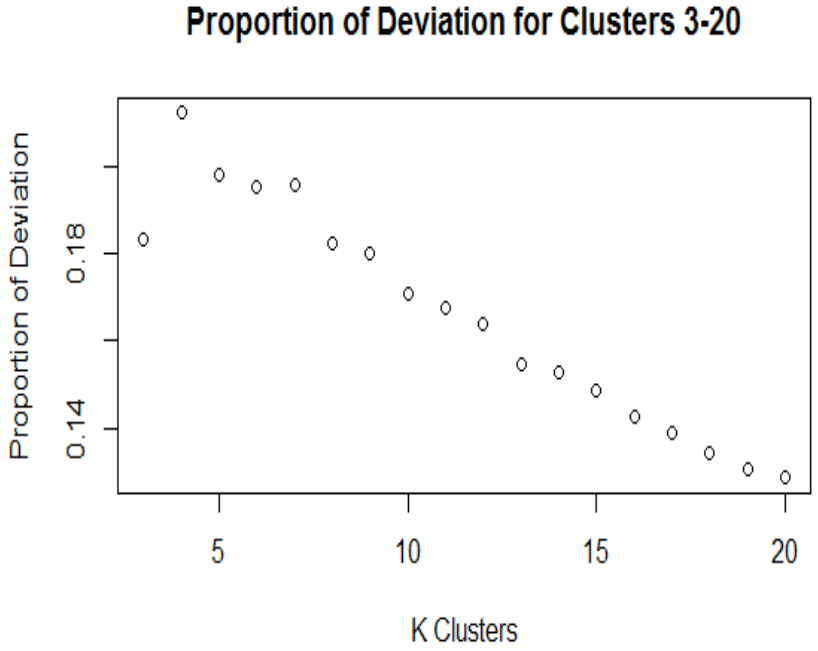
- Identified regions using DMRcate algorithm (Peters, et al. 2015)
  - 320K probes, 90 tumor-matched normal-adjacent pairs
  - Extracts DMRs via Gaussian kernel smoothing on pre-specified bandwidth (7 CpGs within 1kb)
  - Identified 6030 significant DMRs (FDR <0.05 using BH correction)
    - 4733 hypermethylated, 1297 hypomethylated.
    - Mean no. CpGs per DMR = 11 (Range 7 to 112)
  - 69,491 total unique CpG probes in the 6030 DMRs: Still need to reduce dimension further!
- Selected 1,192 probes highly negatively correlated (Pearson < -0.5) with gene expression
- In summary: Went from 480K → 1.2K probes

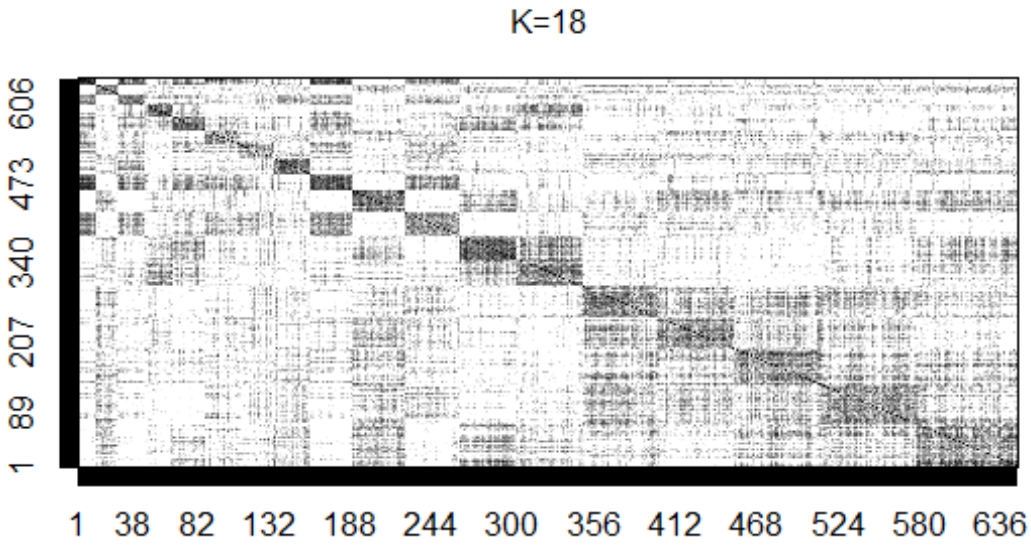
1. Subtype Validation
2. Dimension reduction
3. Subtype Refinement

a. Incorporate all three data types finally!

Integrative Clustering with multiple genomic datatypes

- Used iCluster method again to perform integrative clustering on copy number, gene expression, and methylation data with the following settings: K=3 through 20, lasso shrinkage method for all three data types.





Conclusions (thus far)

- **10 to 12 BRCA subgroups** found using copy number and gene expression data with integrative clustering algorithm
- **18 to 19 BRCA subgroups** when DNA methylation data incorporated
- As with any clustering method, determination of “best” K is not immediately clear, follow-up to find evidence of clinically meaningful “new” subgroups is necessary.

Next Steps

- Check quality of 18 clusters (In-group proportion, silhouette analysis)
- Cross-tabulation (using 2 vectors of cluster assignments) to see which, if any, of the 12 CN/GE clusters split the most with the 18 CN/GE/Methylation cluster assignments
  - Examine if there are survival differences and/or differential methylation features
  - See if we can show concordant expression changes

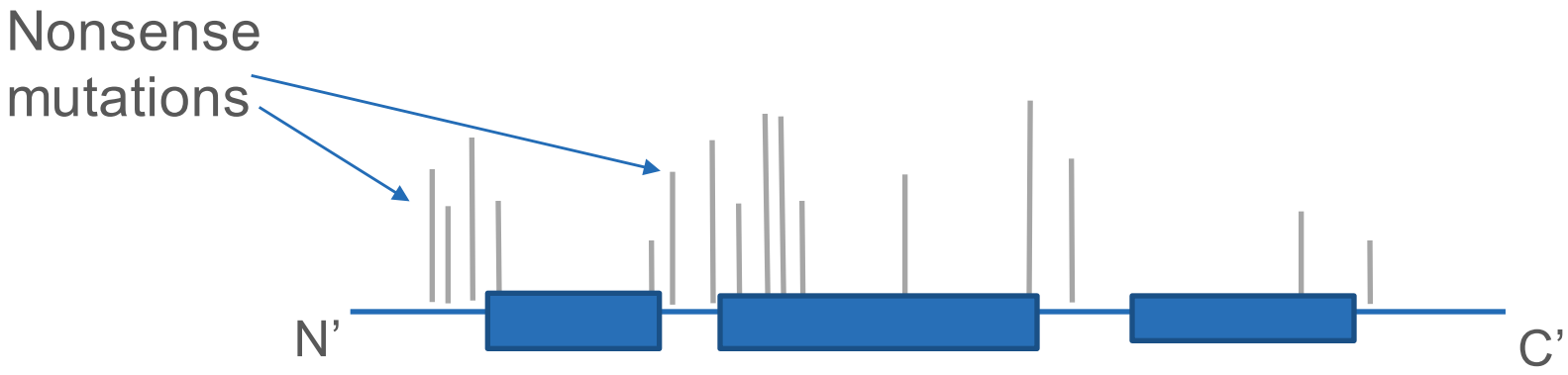
Clustering of Nonsense Mutations in Cancer

Mentor: Yufeng Shen, PhD, Department of Systems Biology, Columbia University

Collaborators: Amr Al-Zain, Rachel Madley, Katherine Croce, Olga Lyudovyk

Goal of Project

- Identify genes with clustering patterns of nonsense mutations in cancer tumors



Background

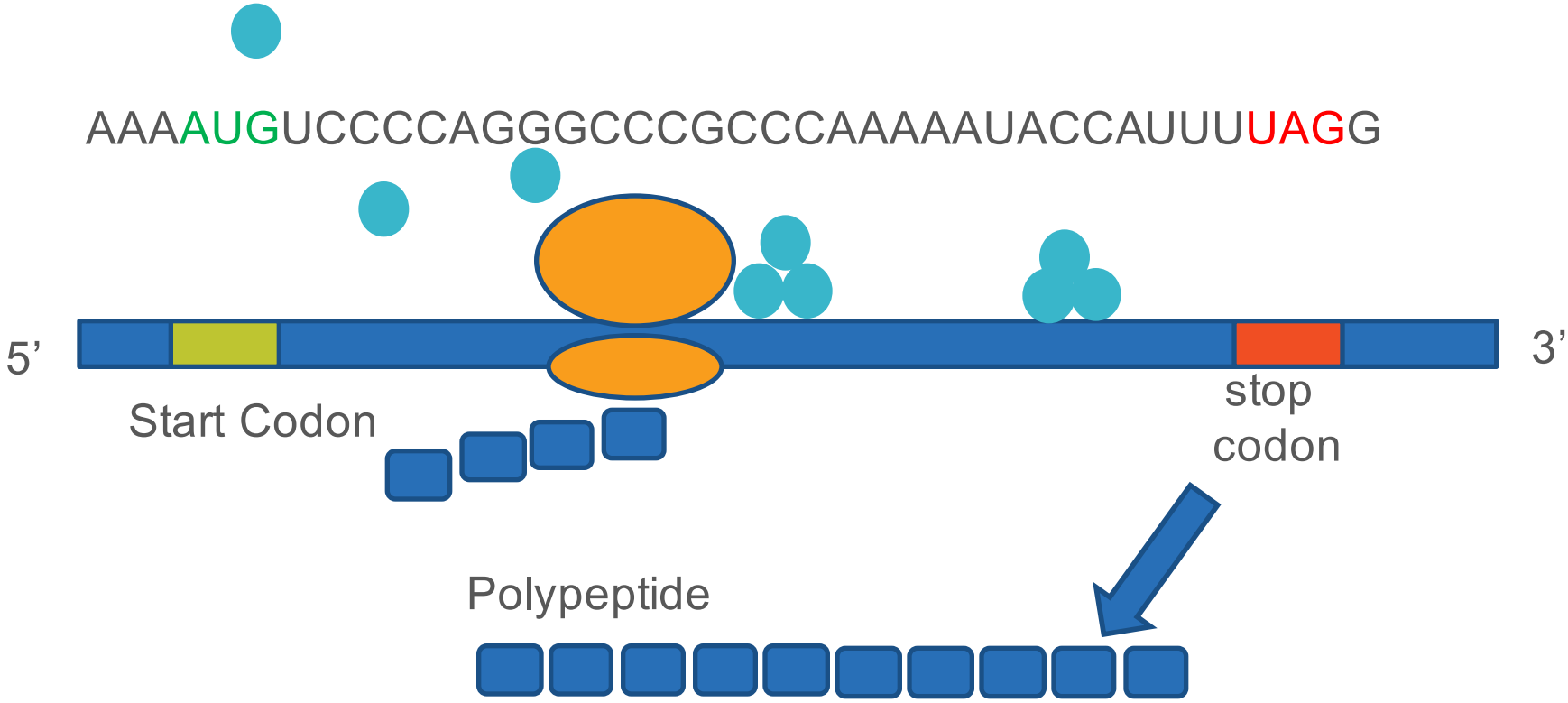
Background: mutations in cancer

- Driver mutations
  - Tumor suppressor genes (TSG) control cell cycle regulation, apoptosis, DNA damage response
  - Oncogenes are negative regulators of cell cycle control, DNA repair etc
- Passenger mutations: do not provide a selective advantage

Genetic Mutations: Substitutions

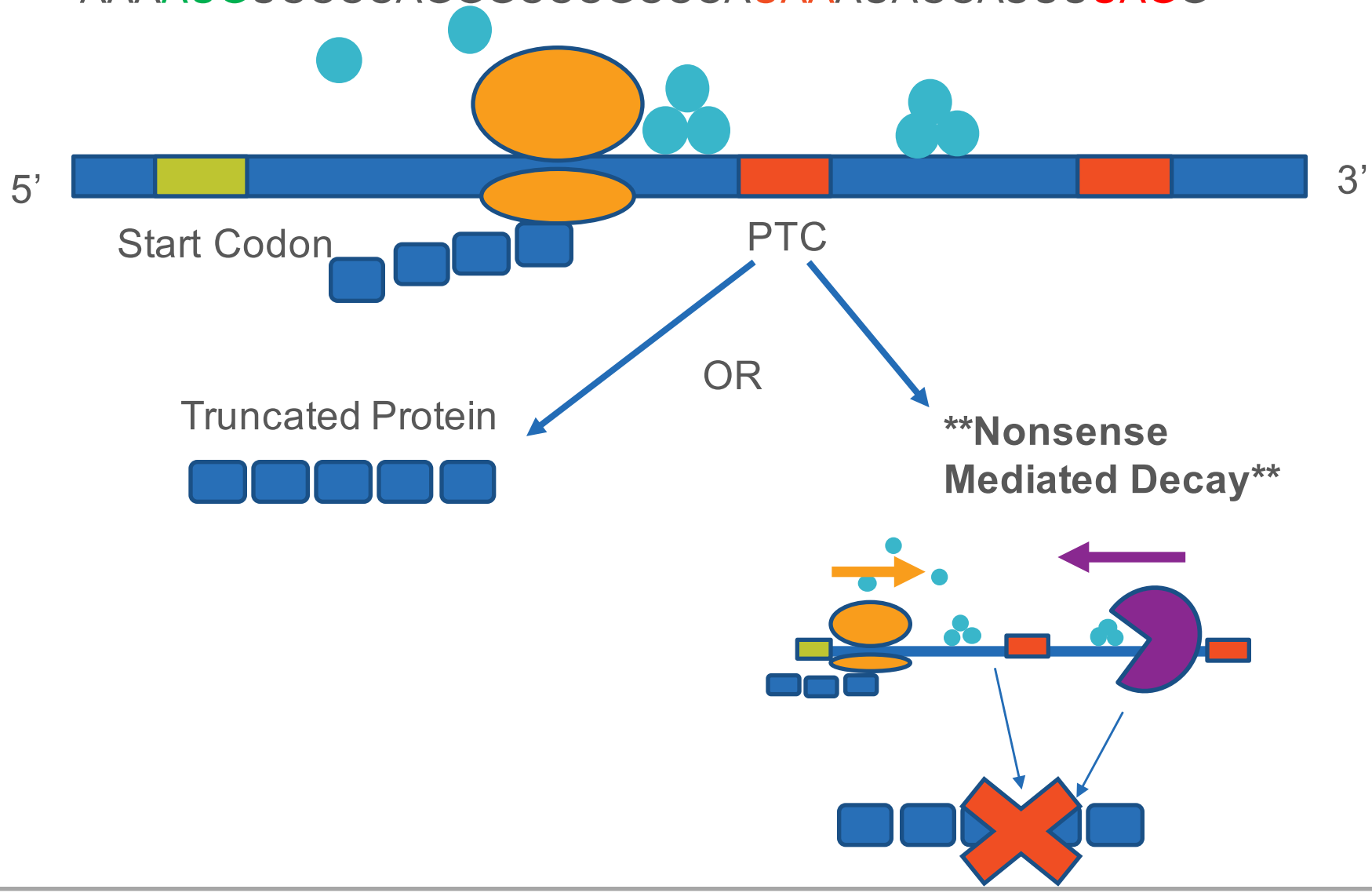
- 4 bp, 64 codons (4^3)
  - 1 “start” codon (AUG)
  - 3 ”stop” codons (UAA, UAG, UGA)
  - 60 codons for 20 amino acids (redundant codons per AA)
- One base pair substitution: CUAAAA → CUA?AA
  - Possible outcomes:
    - “Silent” mutation: creates same amino acid
    - “Missense” mutation: creates different amino acid
    - “Nonsense” mutation: creates premature termination codon (PTC)

Translation



Nonsense Mediated Decay

AAAAUGUCCCCAGGGCCCGCCCAUAAAUACCAUUUUAGG



Nonsense Mediated Decay (NMD)

- NMD efficiency is variable, especially in the disease setting
  - Cancer context: studying nonsense mutations informative since these mutations can result in either NMD (protein degradation) or truncated proteins, which may have biological effects (especially in tumor suppressors or oncogenes)

Research Question

- Is there a significant clustering of nonsense mutations in certain genes in cancer tumors, after taking into account background mutation rate?
  - Clustering in hotspots
  - Clustering positionally

About the Dataset

- Individual somatic data compiled by Chang et al. from TCGA, International Genome Consortium, and other published studies
- 10,668 human cancer tumor samples, 41 cancer types
- Removed indels since we didn't want to examine frameshift stop codons
- Classified genes using COSMIC census

	Nonsense Mutations	Silent Mutations	Number of Genes
TSG	4312	8978	159
Oncogene	983	6363	94
Other	95,132	539,755	16,240
Total	102,244	562,481	16,493

- Filtered to look at 1092 highly mutated genes (>15 nonsense mutations)

Our Project: Examine clustering patterns of nonsense mutations in cancer

1. *Used existing algorithms and examined results*
2. *Developed new algorithm*

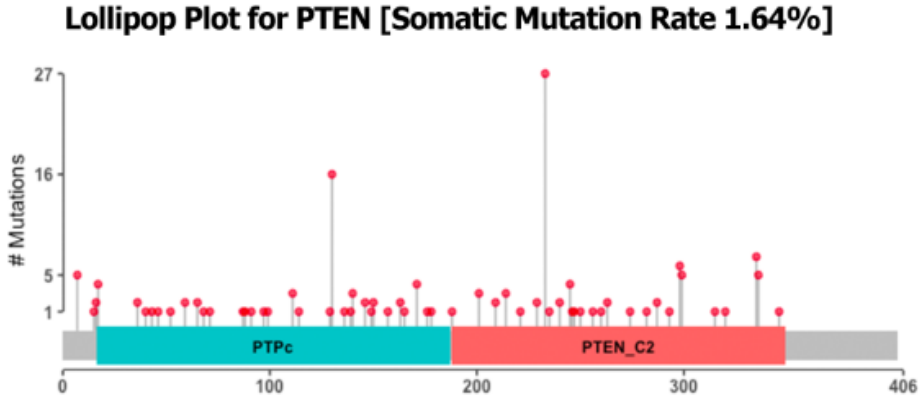
1. Existing clustering algorithms

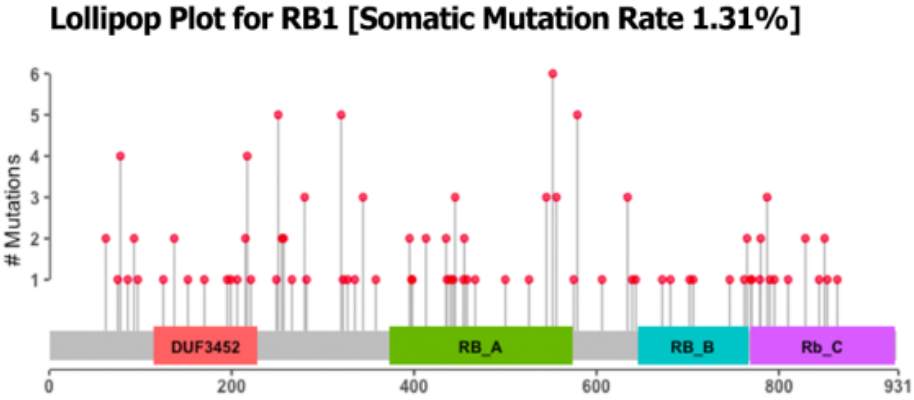
2. New algorithm development

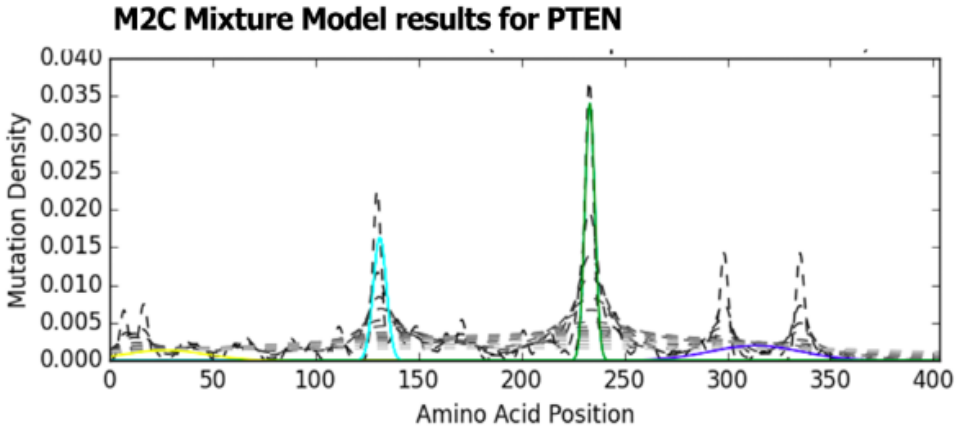
Selected two promising algorithms

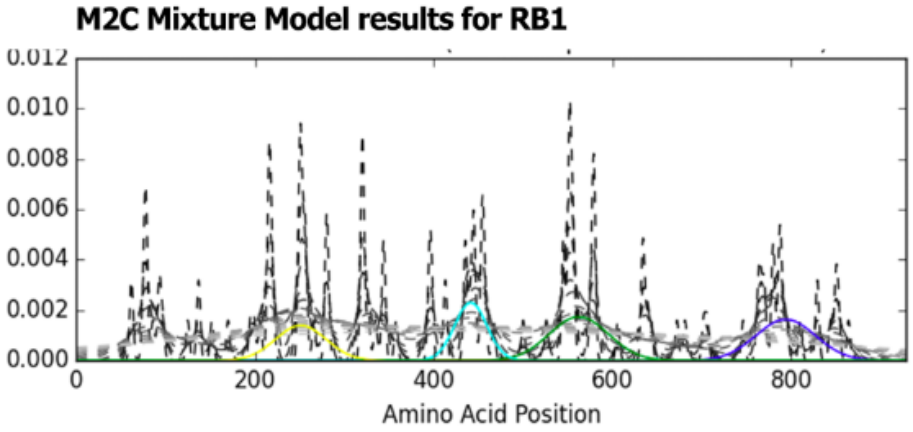
- oncodriveCLUST, a kernel-smoothing algorithm with a pre-defined single scale (Tamborero)
  - Found 37 genes, mostly with numerous small clusters (14 TSGs, 1 oncogene, 22 other)
- M^2C, a multiscale clustering algorithm (Poole) Identifies variable length mutation clusters using multiple continuous probability density functions
  - Found significant clusters in 21 out of the 37 genes

Examining M2C Results





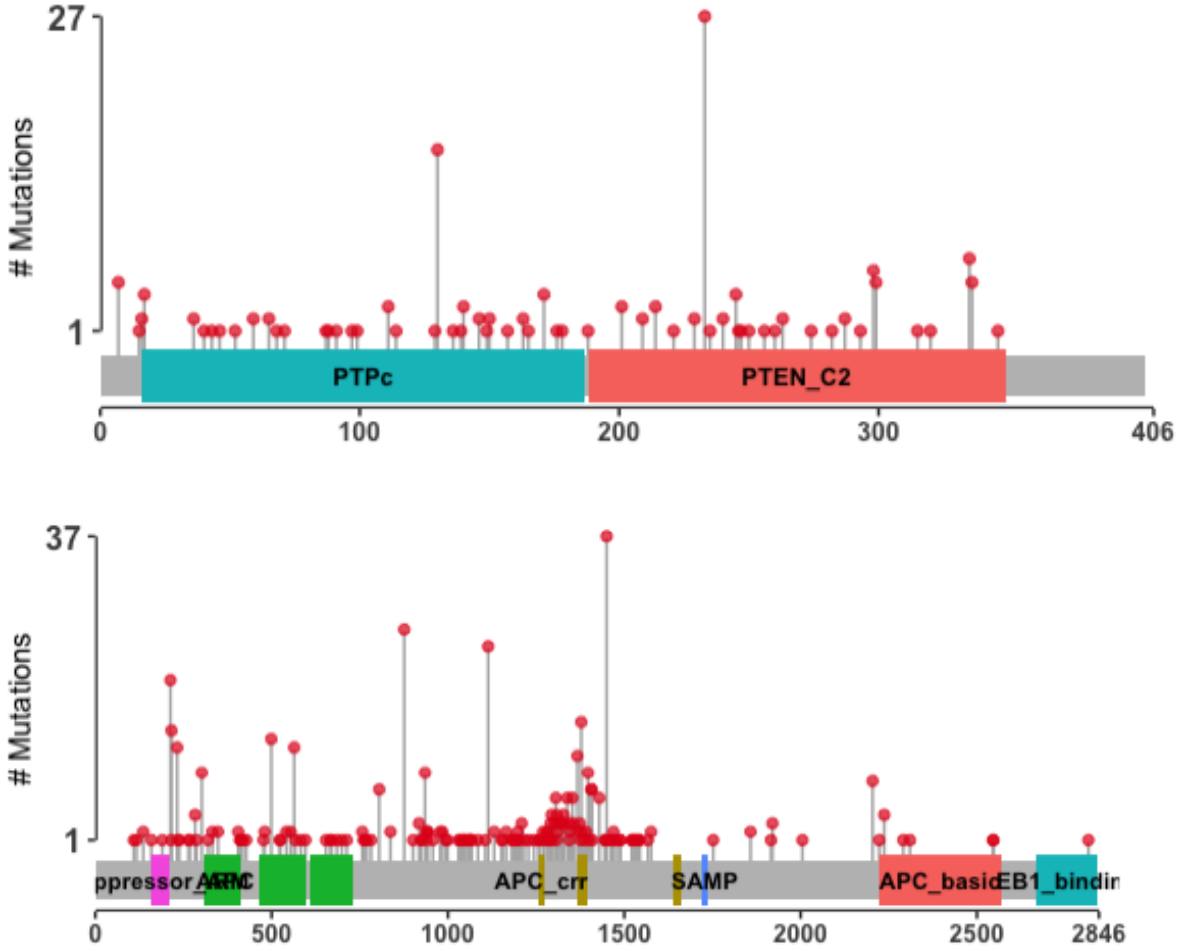




Discussion

- Unsatisfying ”clusters” using existing methods
  - Neither algorithm is meant for looking only at nonsense mutations.
  - oncodriveCLUST uses unsophisticated genome-wide background mutation rate. M2C slightly better, but still uses gene-wide rate calculated using synonymous mutations.
  - Nonsense mutations also don’t seem to cluster in hotspots like missense mutations, are more spread out over large regions
    - However, we found nonsense mutations were NOT necessarily always uniformly distributed, even if they weren’t tightly clustered.

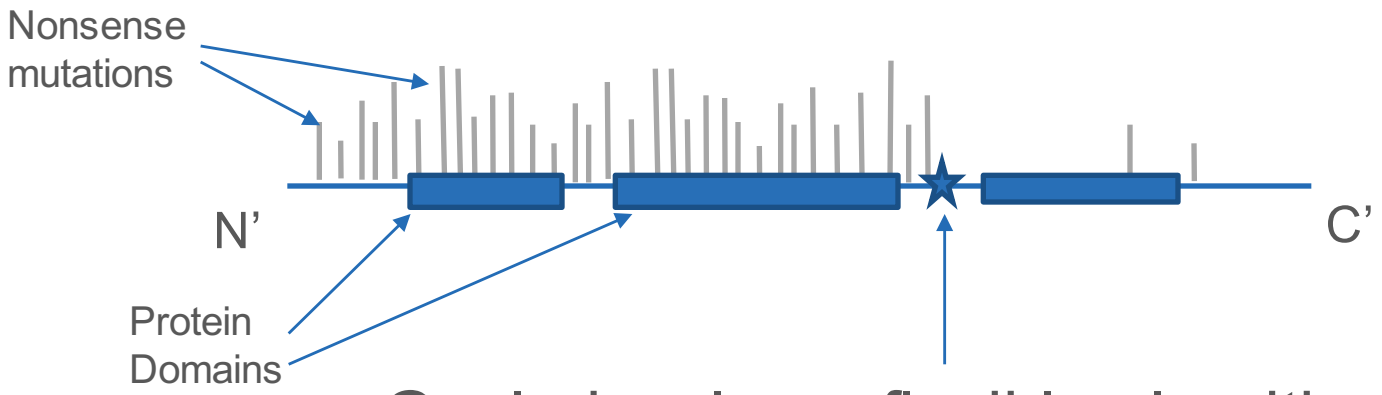
Is there a boundary beyond which nonsense mutations are depleted?



1. Utilized existing clustering algorithms

2. New algorithm development

Boundary-identification motivation



Goal: develop a flexible algorithm to identify genes with a boundary like this, while taking into account complexities of background mutation rate

Our algorithm: densityFindR

Calculate mutation densities upstream of 3 initial points

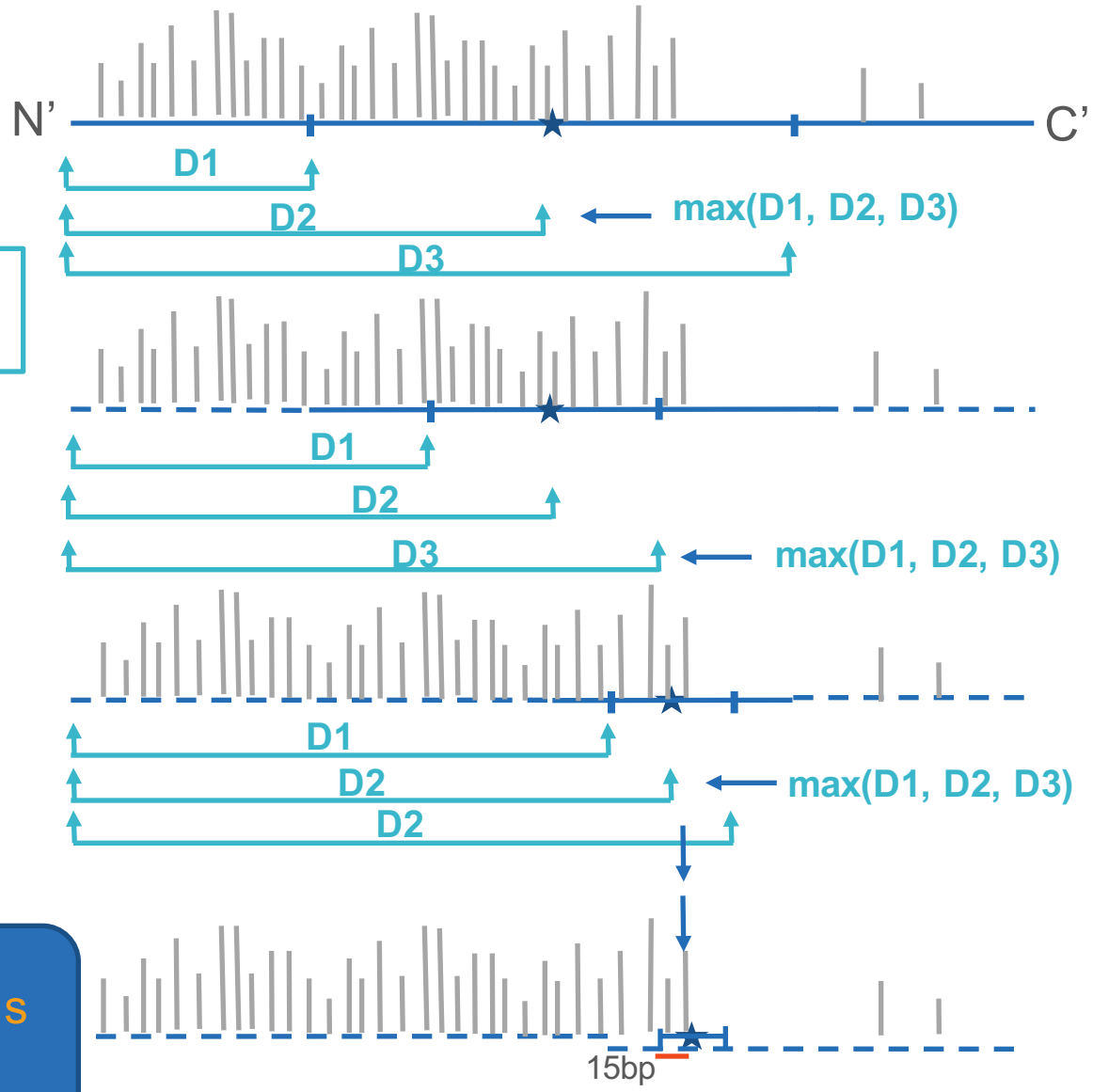
Max(D1, D2, D3) is chosen as the center of the next three points

Repeat, narrow

STOP when 3 points within 15bp

Last center point is named the "boundary" for that gene

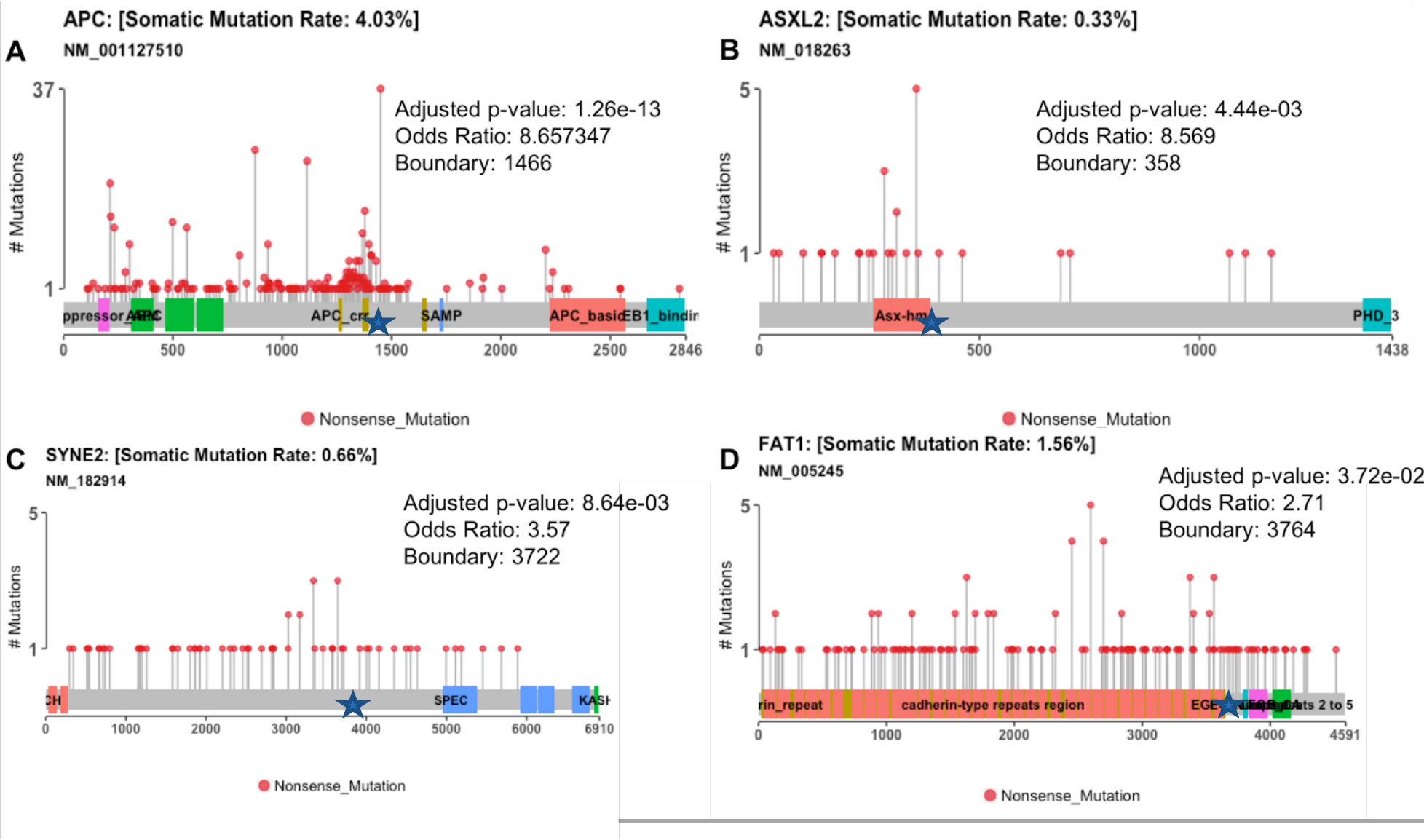
Finally, all genes are tested for enrichment of nonsense mutations upstream of the boundary, compared to downstream



Applying new algorithm

- Supervised input using ~1000 highly mutated genes (>15 NMs), identified boundaries for all input genes
- Then used Fisher's exact test to identify genes with a significant difference in mutation density before and after boundary
  - Fisher's exact test comparing density of nonsense mutations before and after the boundary compared to background mutation
  - At this time, used synonymous mutation pre- and post-points to estimate background mutation rate
- Identified 64 genes that had significant odds ratios (FDR <0.1) indicating enrichment in the region upstream of the boundary
  - 15 listed as cancer genes in the COSMIC census

Manual Inspection of Genes



Future steps

- Improve boundary identification by exploring methods to randomly seed the start point
  - Repeat process several times to find consensus decision for density boundary
- Incorporate more sophisticated context-specific background mutation rate by including known mutational signatures into the null
- See if we can incorporate protein domain information from PFAM

Questions?

Thank you:

- Ronglai Shen, PhD, MSK
- Shuang Wang, PhD, Columbia
- Ya Wang, Columbia
- Arshi Arora, MSK
- Yufeng Shen, PhD, Columbia
- Amr Al-Zain, Columbia
- Katherine Croce, Columbia
- Rachel Madley, Columbia
- Olga Lyudovyk, Columbia

References: BRCA Subtypes

Ali HR, Rueda OM, Chin SF, Curtis C, Dunning MJ, Aparicio SA, et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biol. 2014;15(8):431. pmid:25164602

Aryee MJ, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics (Oxford, Engl). 2014;30(10):1363-9.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B. 1995;57(1):289-300.

Brenton JD, Tavare S, Caldas C, et al: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012, 486: 346-352.

Butcher LM, Beck S. Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data. Methods. 2015;72:21-8.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL,

Gao Y, Jones A, Fasching PA, et al. The integrative epigenomic-transcriptomic landscape of ER positive breast cancer *Clinical Epigenetics*. 2015;7:126. doi:10.1186/s13148-015-0159-0.

Holm K, Staaf J, Lauss M, et al. An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells *Breast Cancer Research : BCR*. 2016;18:27. doi:10.1186/s13058-016-0685-5.

iC10: a copy number and expression-based classifier for breast tumors. Version 1.1.3. Comprehensive R Archive Network; 2015 Sep 23 [cited 2016 Nov 8]. Available from: https://CRAN.R-project.org/package=iC10.

Jaffe AE, Murakami P, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41(1):200-9.

Kapp, A. V. & Tibshirani, R. [Are clusters found in one dataset present in another dataset?](https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxj029) Biostatistics 8, 9-31 (2007).

Michaut M, Chin S-F, Majewski I, et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific Reports*. 2016;6:18517. doi:10.1038/srep18517.

Morris TJ, et al. ChAMP: 450k chip analysis methylation pipeline. Bioinformatics. 2014;30:428-30.

Peters TJ, Buckley MJ, Statham AL, et al. *De novo* identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin*. 2015;8:6. doi:10.1186/1756-8935-8-6.

Teschendorff AE, Marabita F, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics (Oxford, Engl.). 2013a;29(2):189-96.

TCGA Network: Comprehensive molecular portraits of human breast tumors. Nature 2012, 490:61-70.

References: Clustering Nonsense

- Chang MT, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity andmutational specificity. Nature biotechnology. 2016;34(2):155-163.
- Forbes, S.A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43,D805–D811.
- Mayakonda A and Koeffler PH (2016). “Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies.” *BioRxiv*. doi: 10.1101/052662.
- Poole W, et al. Multiscale mutation clustering algorithm identifies pan-cancer mutational clusters associated with pathway-level changes in gene expression. PLoS Computational Biology. 2017;13(2):e1005347.
- Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238-44.

Thank you!

Identification of genes containing nonsense mutation clusters