# Clinical phenotype prediction from highly-polymorphic structurally-variant genotypes

Tim Farrell
Course Project, BE562
December 11, 2015
tmf@bu.edu

# Motivation

- Increased use of technology in clinic gives Data
  - EHRs + genomics
  - Systems biology/ physiology data (more to come)

- Trend accelerating:
  - Precision Medicine Initiative in 2015
  - *Nature* Big Data in Biomedicine feature in Nov 2015
  - IBM Watson, Google Life Sciences (now Verily), etc.

# Human genomic variation and clinical sequencing

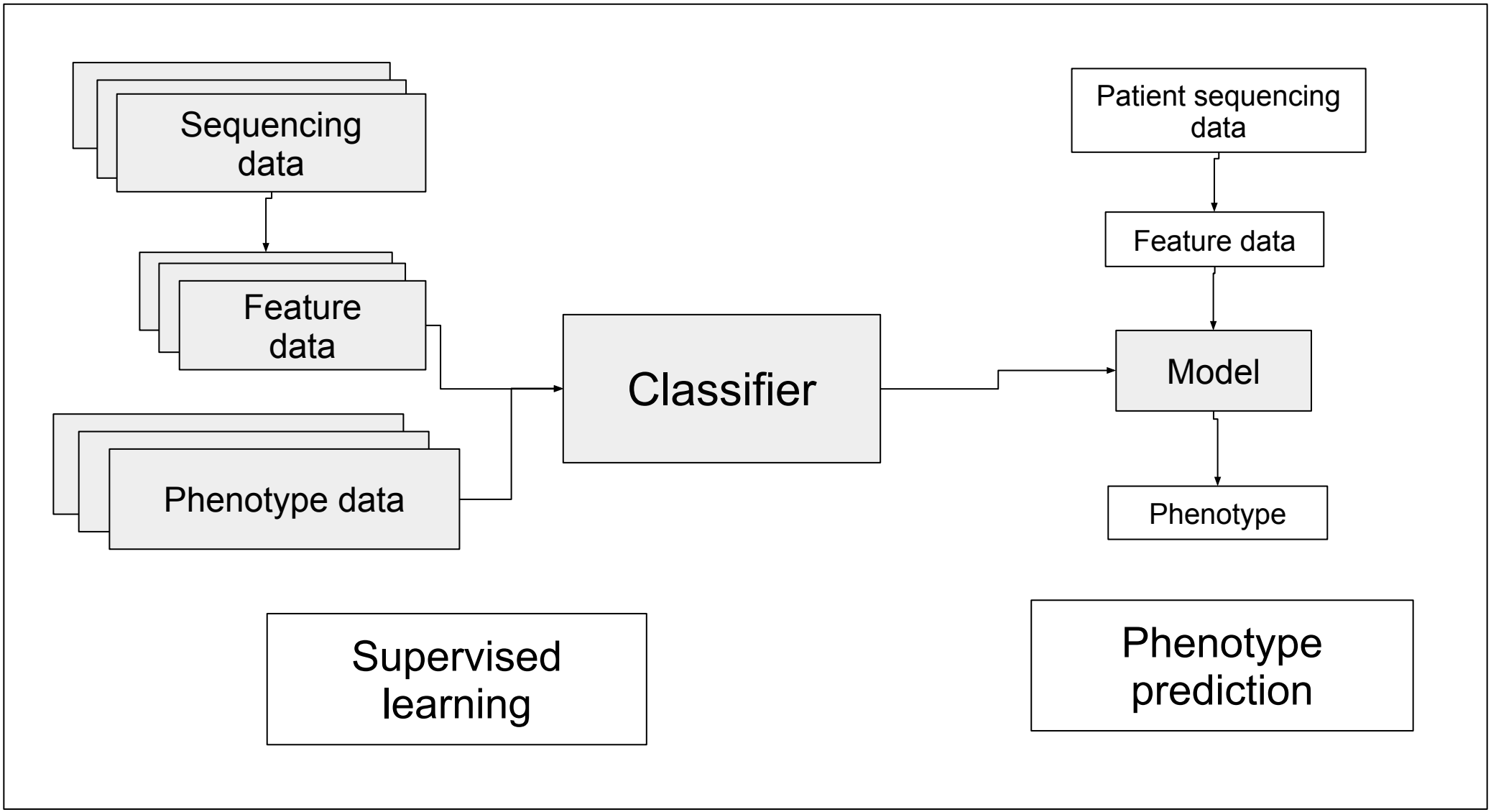
- 80 million variants identified in human genome (Jun 2015)
  - SNPs
  - structural (>50bp; CNV, translocations, etc.)

- High discordance b/t sequencing tech and variant callers (VCs)

- Recent study on VC standardization reported 23% of human genome is “difficult” (i.e. not enough consensus among tools to make reasonable prediction)

- Gives low confidence for “predictive” clinical sequencing

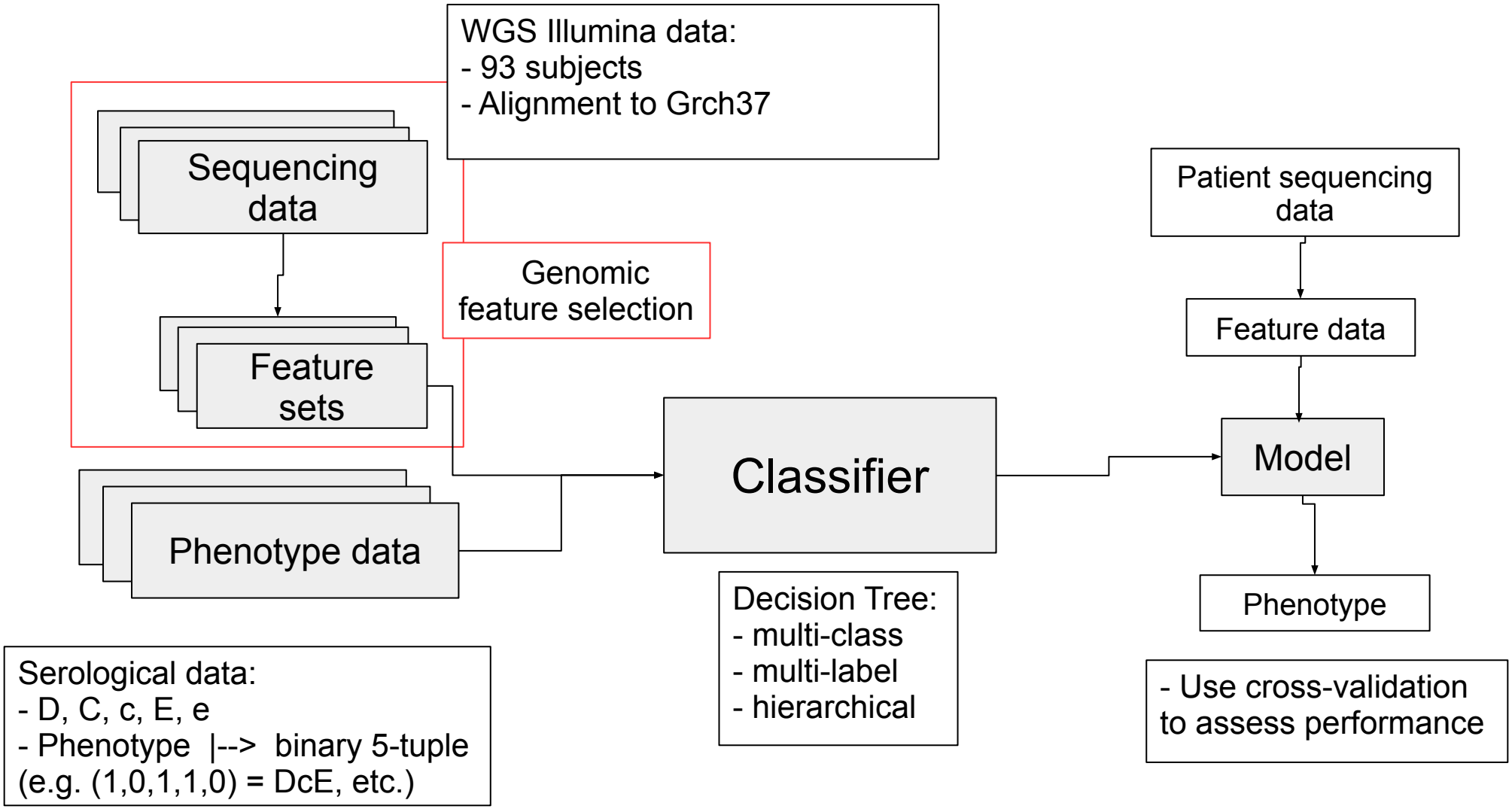
# Building better predictive models for automated clinical phenotyping



# Rh RBC antigen genes

- Rh RBC antigen genomic region exemplifies "difficult"
  - Encodes for highly immunogenic antigens on RBC membranes
- RhCE and RhD
  - Highly similar genes known to undergo complex rearrangements
- 50 known antigens
  - Most significant: D, C, c, E, e
  - Many-to-one relationship haplotypes-to-phenotype

# Rh antigen prediction pipeline



# Feature selection: crude

Build PFM for each sample for each gene's exon, then...

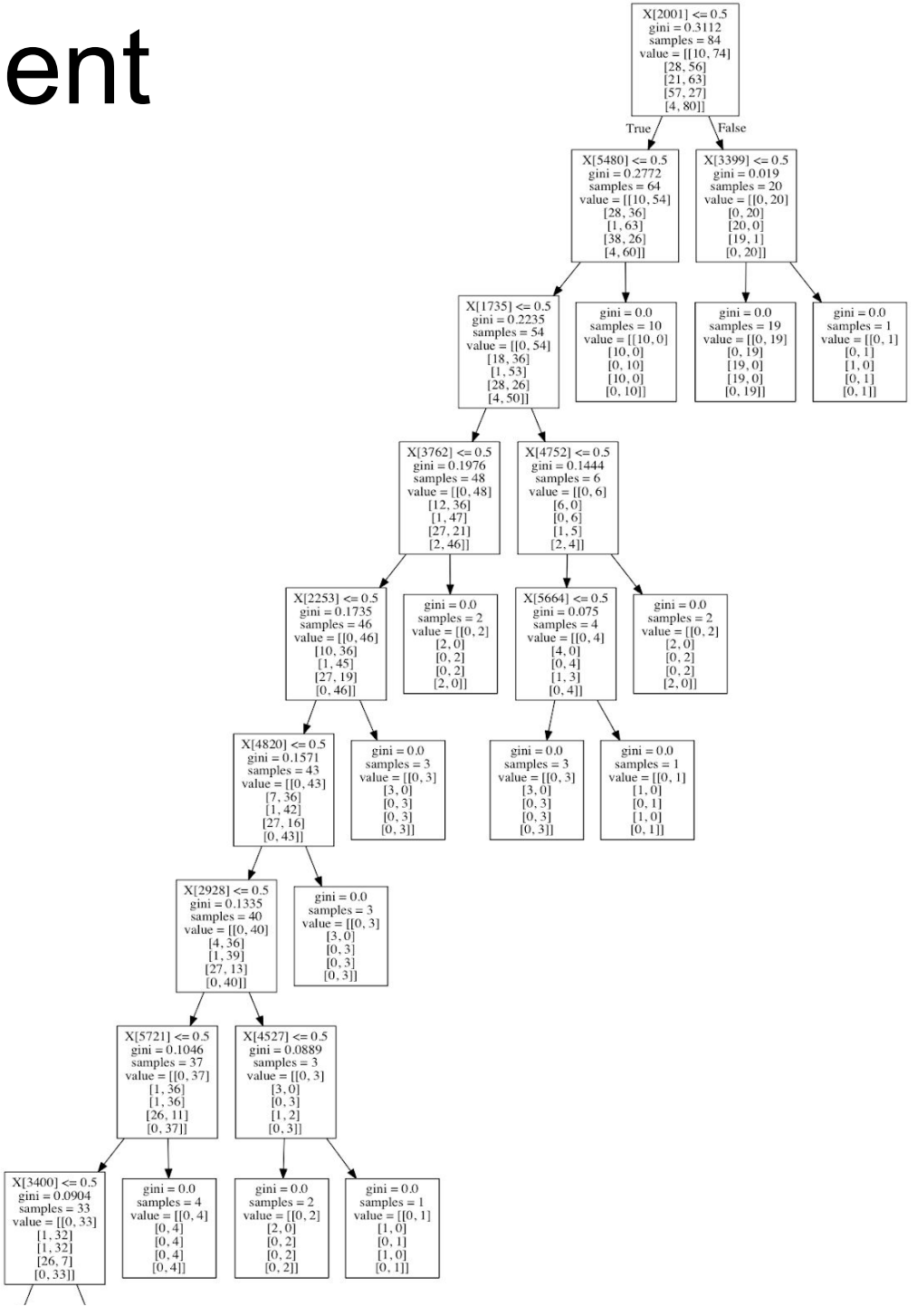
- Select
  - Whole exome
  - Variant positions associated with differential phenotypes:
    - dbRBC, ClinVar, dbSNP, dbVar, etc.
    - Call 'diff_genotype'
- Measure:
  - Categorical: call base with highest frequency
  - Position frequency/ max coverage
- Encode:
  - Encoding | Nonencoding
  - e.g. [(1, 4), (2, 3)] |--> [(1, 0, 0, 1), (0, 1, 1, 0)]

# Feature typeset assessment

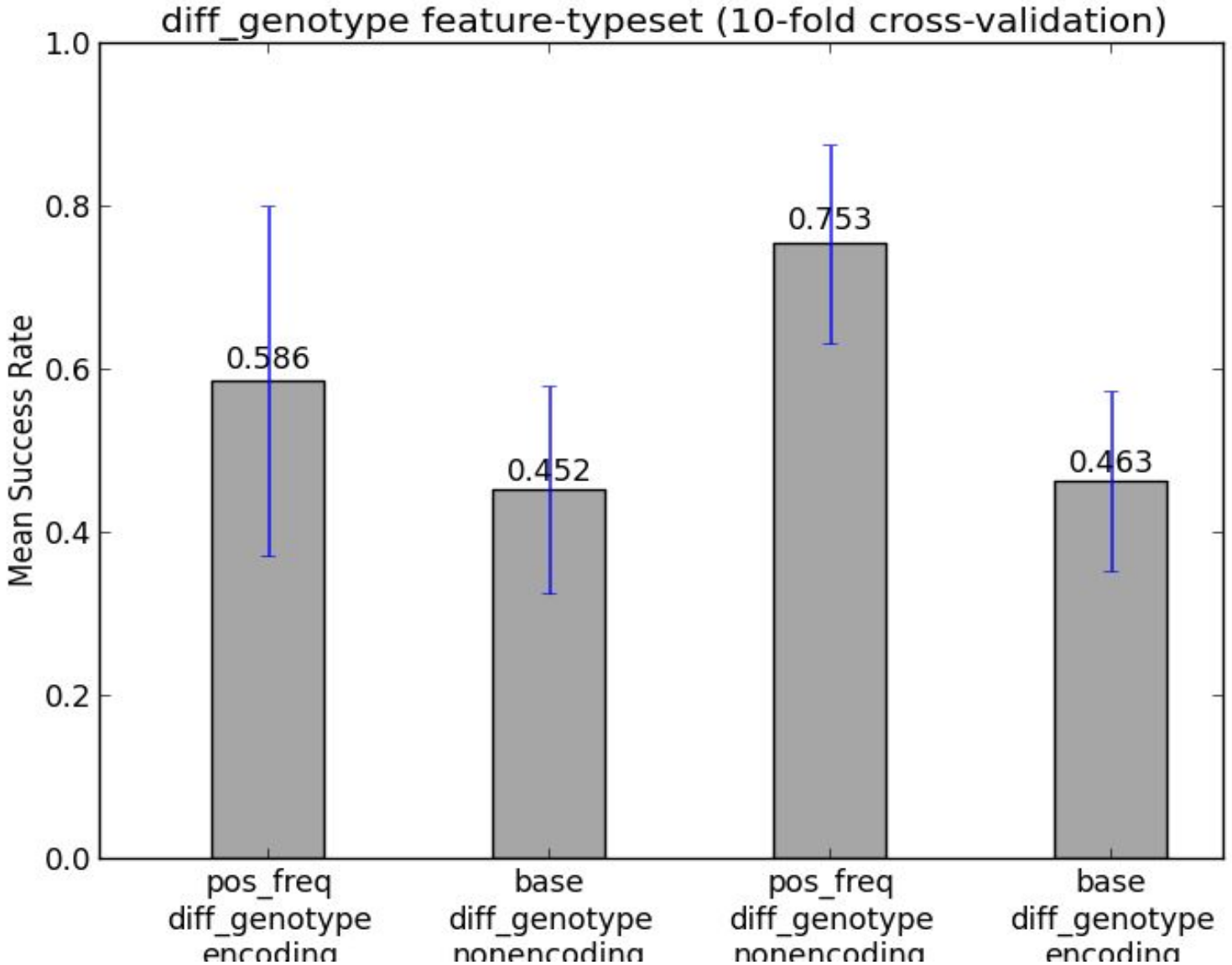
For each feature typeset:

(a) perform 10-fold cross-validation with DecisionTree classifier

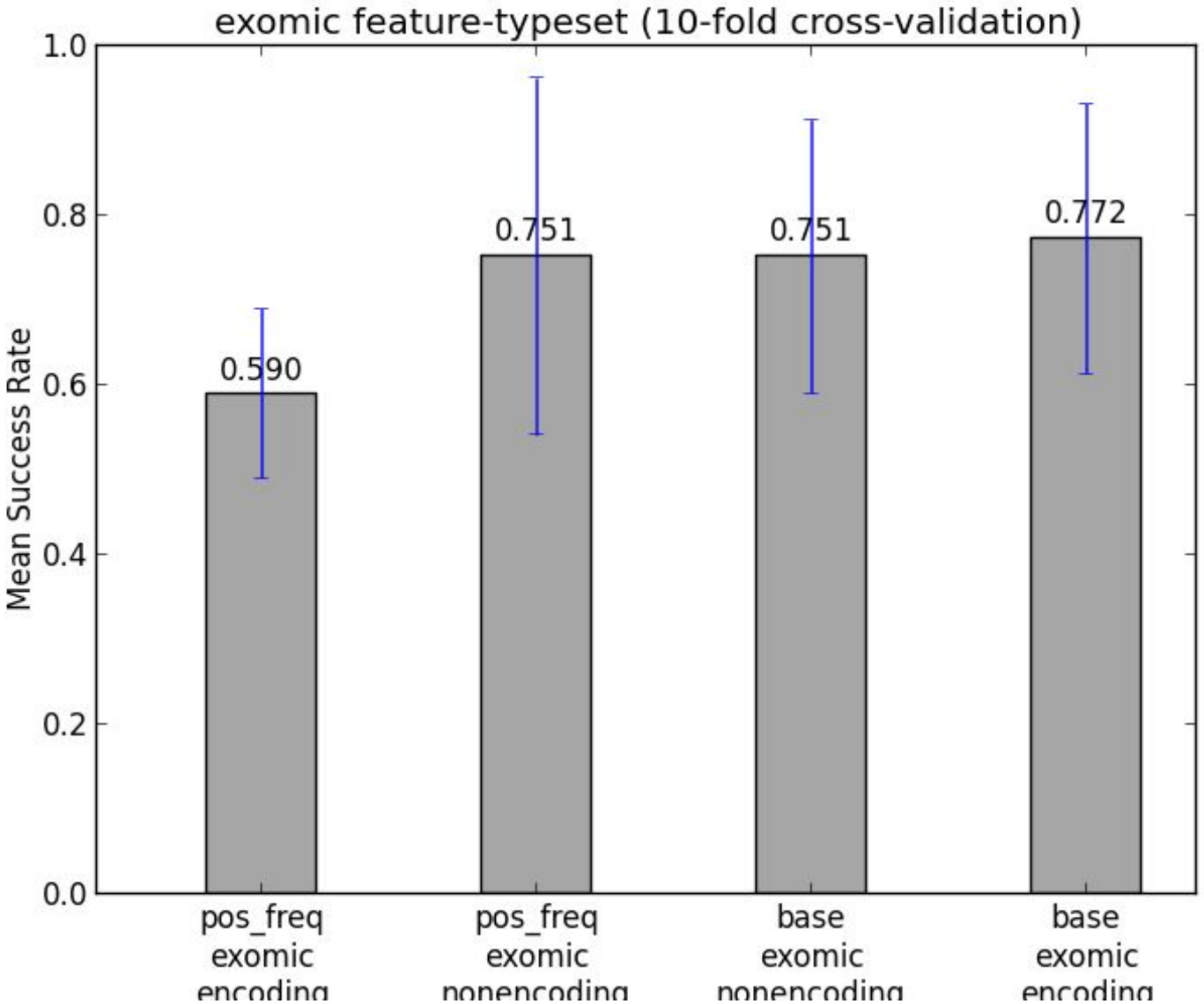
(b) measure success rate



# diff_genotype feature sets



# exomic feature sets



# Feature selection: fully-featured

- Use well-established bioinformatics tools to better characterize and differentiate genomic architectures
  - MEME/ DREME:
    - call motifs within exons to eliminate commonalities across genotypes
    - look for motifs in introns that may add specificity
  - Weeder: count motifs
  - HaplotypeCaller: calls SNPs and SV
- Still working on fitting the metrics generated with these together

# Future directions

- More data sources:
  - Long-read capable sequencing tech
  - Overlapping primer sets with barcodes

# References/ Thanks

[1] Jameson JL and Longo DL. 2015. Precision medicine – personalized, promising and problematic. *N Engl J Med.* 372(23): 2229-2234.

[2] Baker M. 2012. Structural variation: the genome's hidden architecture. *Nat Methods.* 9(2): 133-139.

[3] Silvestri GA, Vachani A, Whitney D, Elashoff M, Smith KP, Ferguson JS, Parsons E, Mitra N, Brody J, Lenburg ME, and Spira A. 2015. A bronchial genomic classifer for the diagnostic evaluation of lung cancer. *N Engl J Med* 373;3.

[4] Qiu P, Cai X, Ding W, Zhang Q, Norris ED and Greene JR. 2009. HCV genotyping using statistical classifcation approach. *J of Biomed Sci*, 16:62. doi:10.1186/1423-0127-16-62.

[5] Abel HJ , Duncavage EJ. 2014. Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. *Cancer Genetics* 206 (2014) 432e440.

[6] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W and Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 30(2): 246- 251.

[7] Seringhaus M and Gerstein M. 2008. Genomics confounds gene classification. *American Scientist*, 96(6) p.466-473.

Bill Lane, BWH Pathology

Peter Tonellato, DBMI HMS