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 Dat
  - EHRs + genomics
  - Systems biology/ physiology data (more to come)

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

Human genomic variation and clinical sequencing

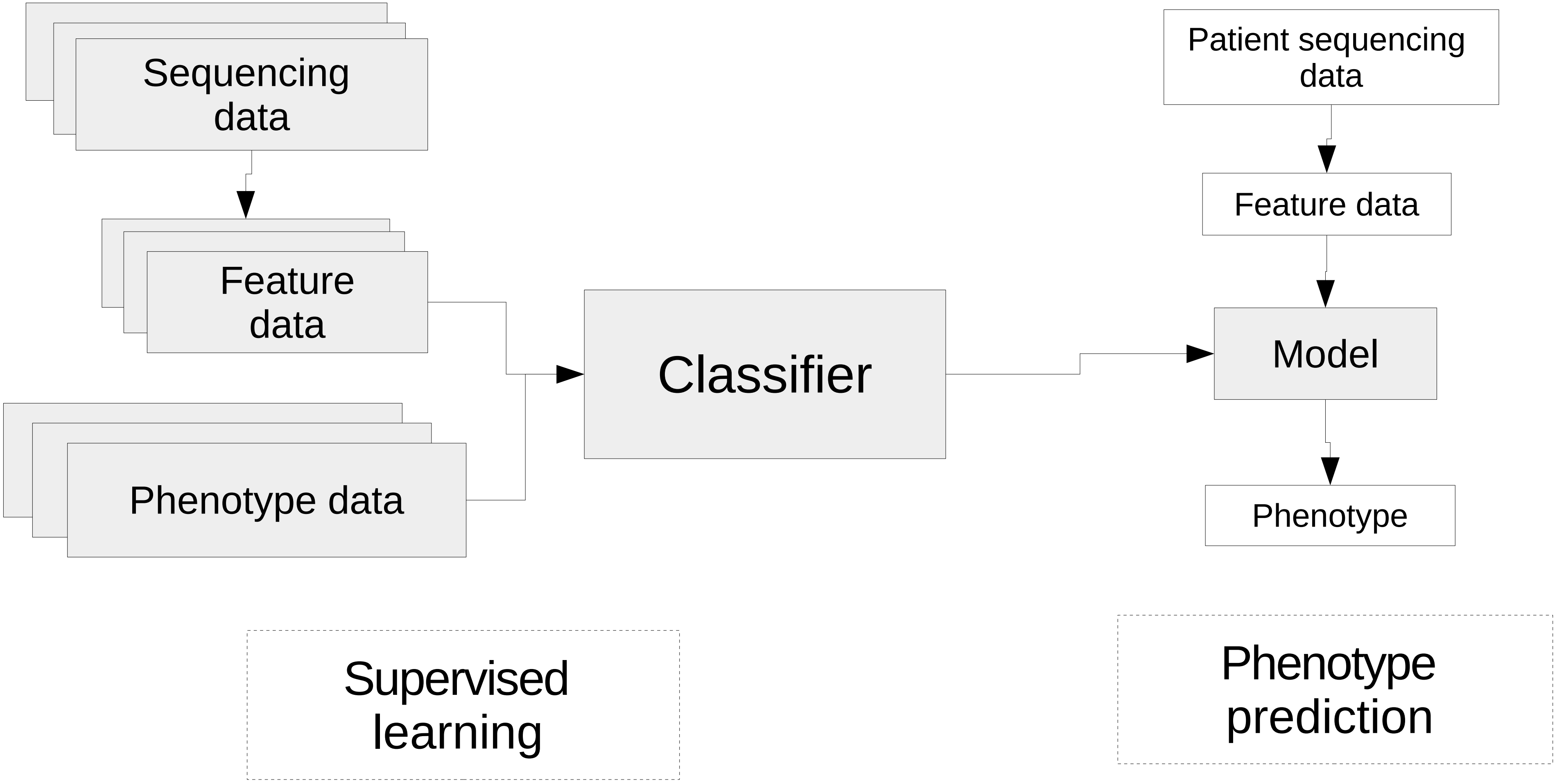
- 80 million variants identified in human genome (Jun 2015)
  - SNPs
  - structural (>50bp; CNV, translocations, etc.)

- 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

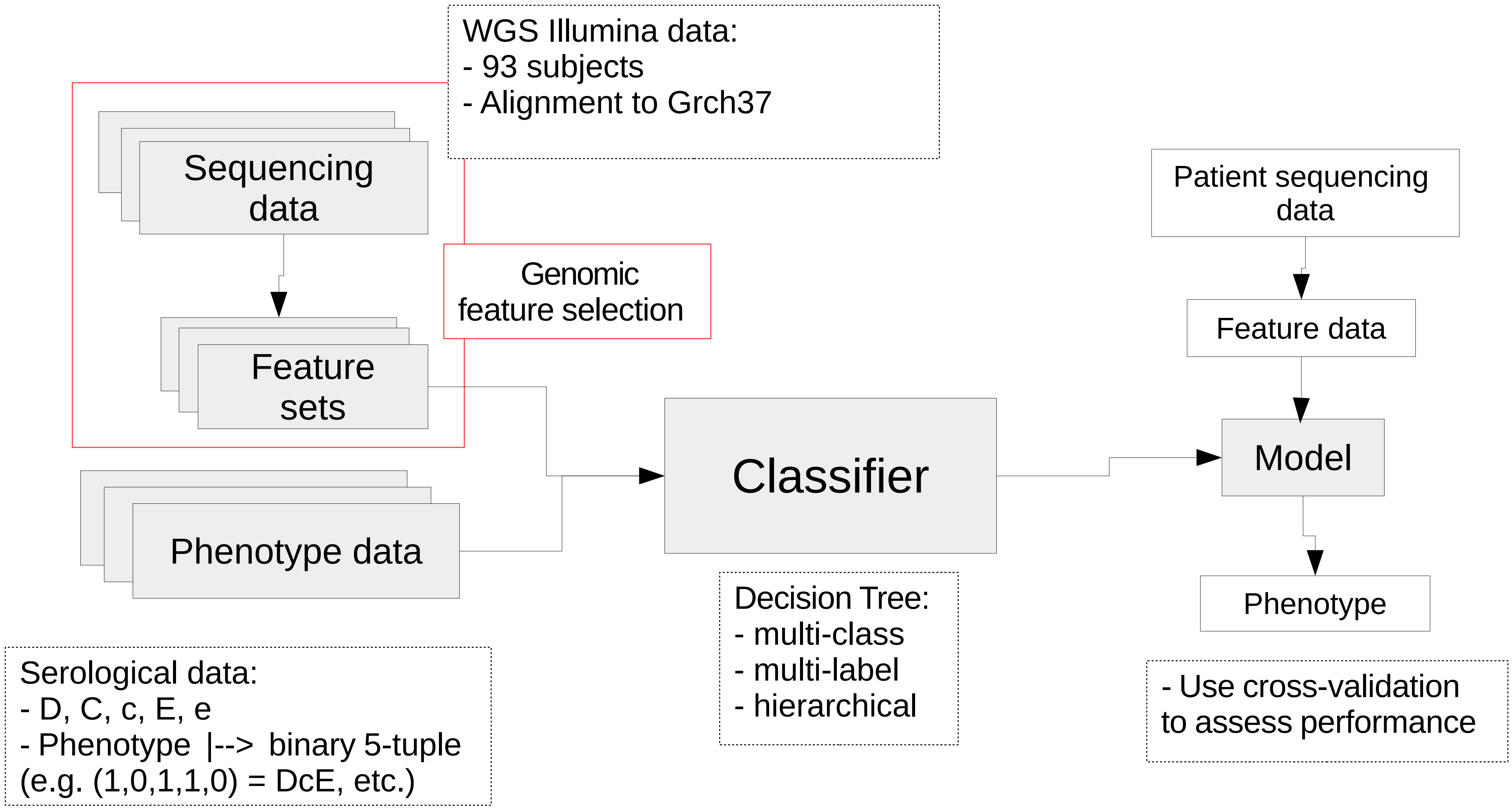
- Rh RBC antigen genomic region exemplifies “difficult”
  - Encodes for highly immunogenic antigens on RBC membranes

- 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 (e.g. heterozygosity; but also silent variation, etc)

- Clinical relevance:
  - Blood transfusion
  - Hemolytic disease of the newborn

Rh antigen prediction pipeline



WGS Illumina data:
- 93 subjects
- Alignment to Grch37

Sequencing data

Genomic feature selection

Feature sets

Phenotype data

Serological data:
- D, C, c, E, e
- Phenotype |--> binary 5-tuple
(e.g. (1,0,1,1,0) = DcE, etc.)

Classifier

Decision Tree:
- multi-class
- multi-label
- hierarchical

Patient sequencing data

Feature data

Model

Phenotype

- Use cross-validation to assess performance

Feature selection: crude

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

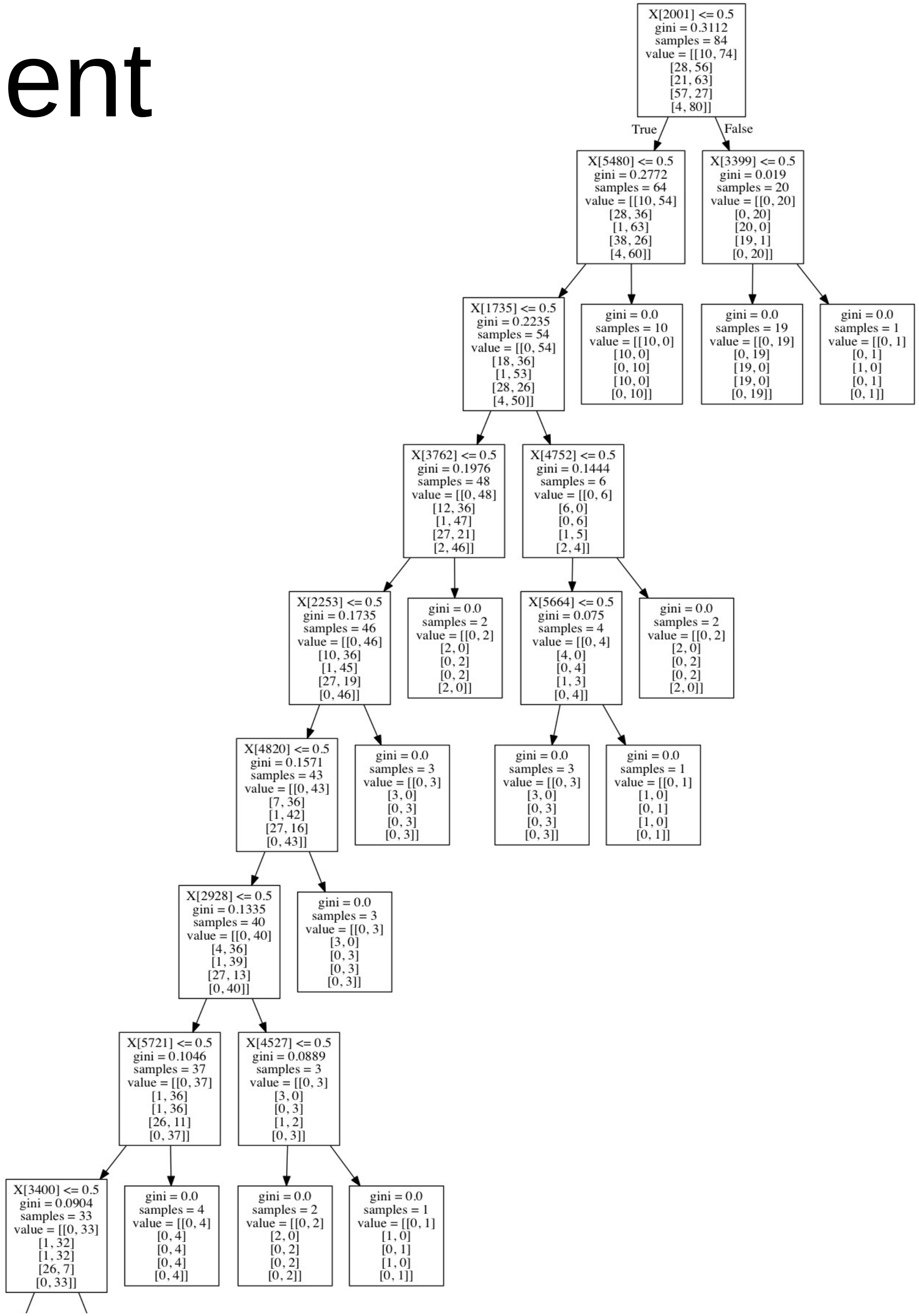
- Select
  - Whole exome
  - Variant positions associated with differential phenotypes:
    - dbRBC, ClinVar, dbSNP, dbVar, etc.
    - Call 'diff_genotype'
- Measure:
  - Categorical: call base with highest frequency
  - Position frequency/ max coverage
- Encode:
  - Encoding | Nonencoding
  - e.g. [(1,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

Feature typeset assessment

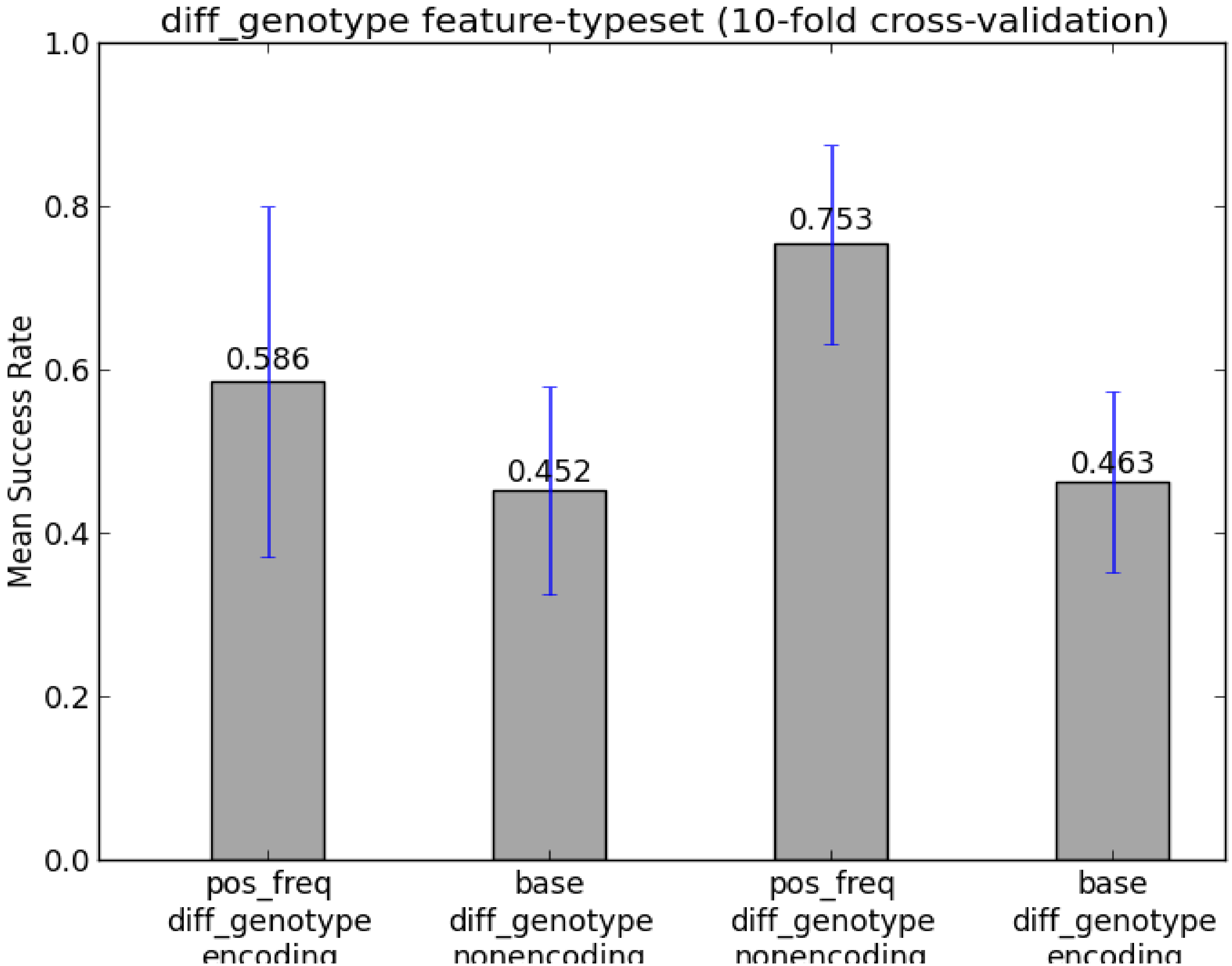
For each feature typeset:

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

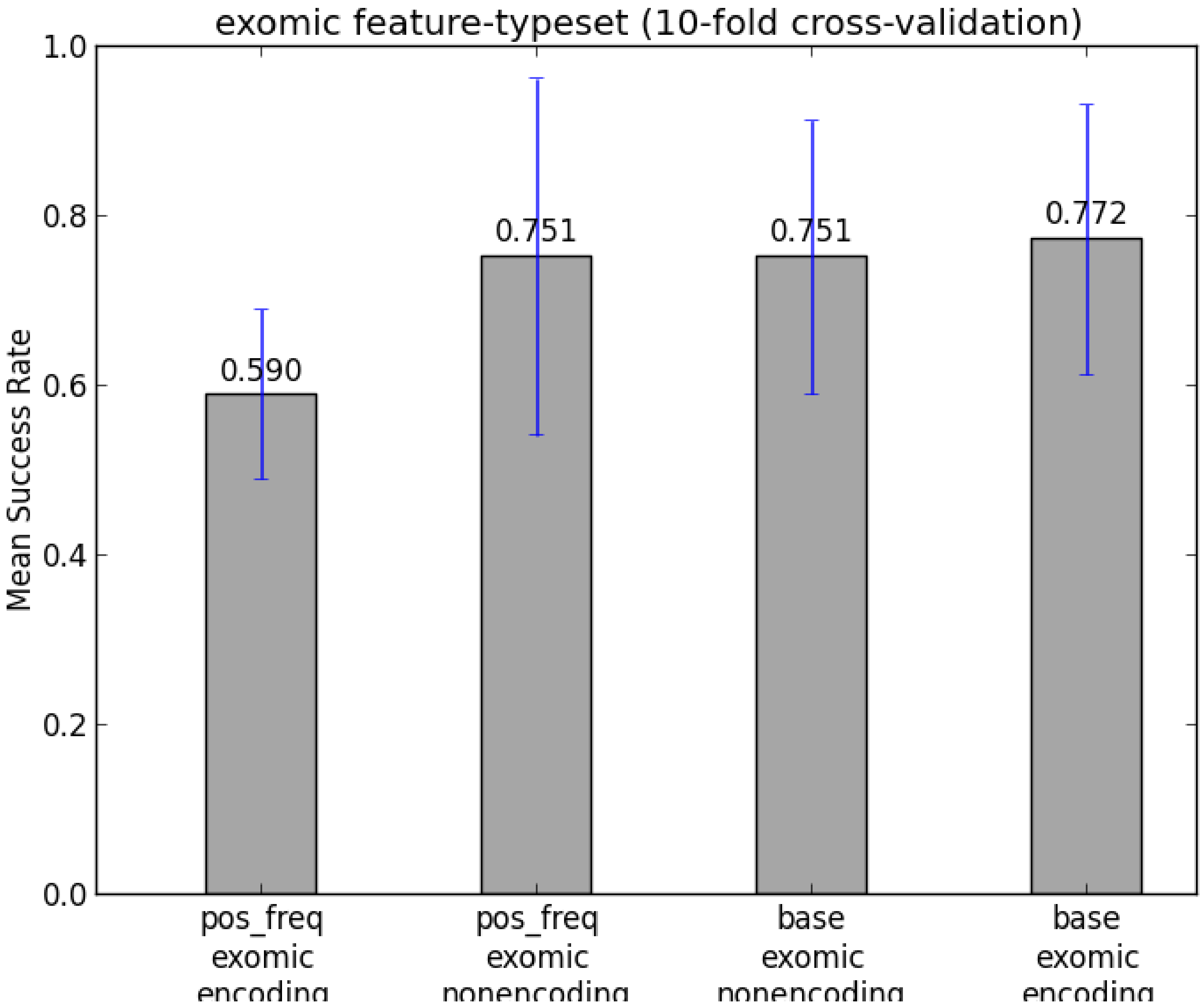
(b) measure success rate


X[2001] <= 0.5 gini = 0.3112 samples = 84 value = [[10, 74] [28, 56] [21, 63] [57, 27] [4, 80]]
True
False
X[5480] <= 0.5 gini = 0.2772 samples = 64 value = [[10, 54] [28, 36] [1, 63] [38, 26] [4, 60]]
X[3399] <= 0.5 gini = 0.019 samples = 20 value = [[0, 20] [0, 20] [20, 0] [19, 1] [0, 20]]
X[1735] <= 0.5 gini = 0.2235 samples = 54 value = [[0, 54] [18, 36] [1, 53] [28, 26] [4, 50]]
gini = 0.0 samples = 10 value = [[10, 0] [10, 0] [0, 10] [10, 0] [0, 10]]
gini = 0.0 samples = 19 value = [[0, 19] [0, 19] [19, 0] [19, 0] [0, 19]]
gini = 0.0 samples = 1 value = [[0, 1] [0, 1] [1, 0] [0, 1] [0, 1]]
X[3762] <= 0.5 gini = 0.1976 samples = 48 value = [[0, 48] [12, 36] [1, 47] [27, 21] [2, 46]]
X[4752] <= 0.5 gini = 0.1444 samples = 6 value = [[0, 6] [6, 0] [0, 6] [1, 5] [2, 4]]
X[2253] <= 0.5 gini = 0.1735 samples = 46 value = [[0, 46] [10, 36] [1, 45] [27, 19] [0, 46]]
gini = 0.0 samples = 2 value = [[0, 2] [2, 0] [0, 2] [0, 2] [2, 0]]
X[5664] <= 0.5 gini = 0.075 samples = 4 value = [[0, 4] [4, 0] [0, 4] [1, 3] [0, 4]]
gini = 0.0 samples = 2 value = [[0, 2] [2, 0] [0, 2] [0, 2] [2, 0]]
X[4820] <= 0.5 gini = 0.1571 samples = 43 value = [[0, 43] [7, 36] [1, 42] [27, 16] [0, 43]]
gini = 0.0 samples = 3 value = [[0, 3] [3, 0] [0, 3] [0, 3] [0, 3]]
gini = 0.0 samples = 3 value = [[0, 3] [3, 0] [0, 3] [0, 3] [0, 3]]
gini = 0.0 samples = 1 value = [[0, 1] [1, 0] [0, 1] [1, 0] [0, 1]]
X[2928] <= 0.5 gini = 0.1335 samples = 40 value = [[0, 40] [4, 36] [1, 39] [27, 13] [0, 40]]
gini = 0.0 samples = 3 value = [[0, 3] [3, 0] [0, 3] [0, 3] [0, 3]]
X[5721] <= 0.5 gini = 0.1046 samples = 37 value = [[0, 37] [1, 36] [1, 36] [26, 11] [0, 37]]
X[4527] <= 0.5 gini = 0.0889 samples = 3 value = [[0, 3] [3, 0] [0, 3] [1, 2] [0, 3]]
X[3400] <= 0.5 gini = 0.0904 samples = 33 value = [[0, 33] [1, 32] [1, 32] [26, 7] [0, 33]]
gini = 0.0 samples = 4 value = [[0, 4] [0, 4] [0, 4] [0, 4] [0, 4]]
gini = 0.0 samples = 2 value = [[0, 2] [2, 0] [0, 2] [0, 2] [0, 2]]
gini = 0.0 samples = 1 value = [[0, 1] [1, 0] [0, 1] [1, 0] [0, 1]]

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 together the metrics/ statistics generated from these for feature set

Future directions

- More/ better 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 M 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 re 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 sequenc 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