Microsatellite Instability Classification by NGS

Microsatellite Instability

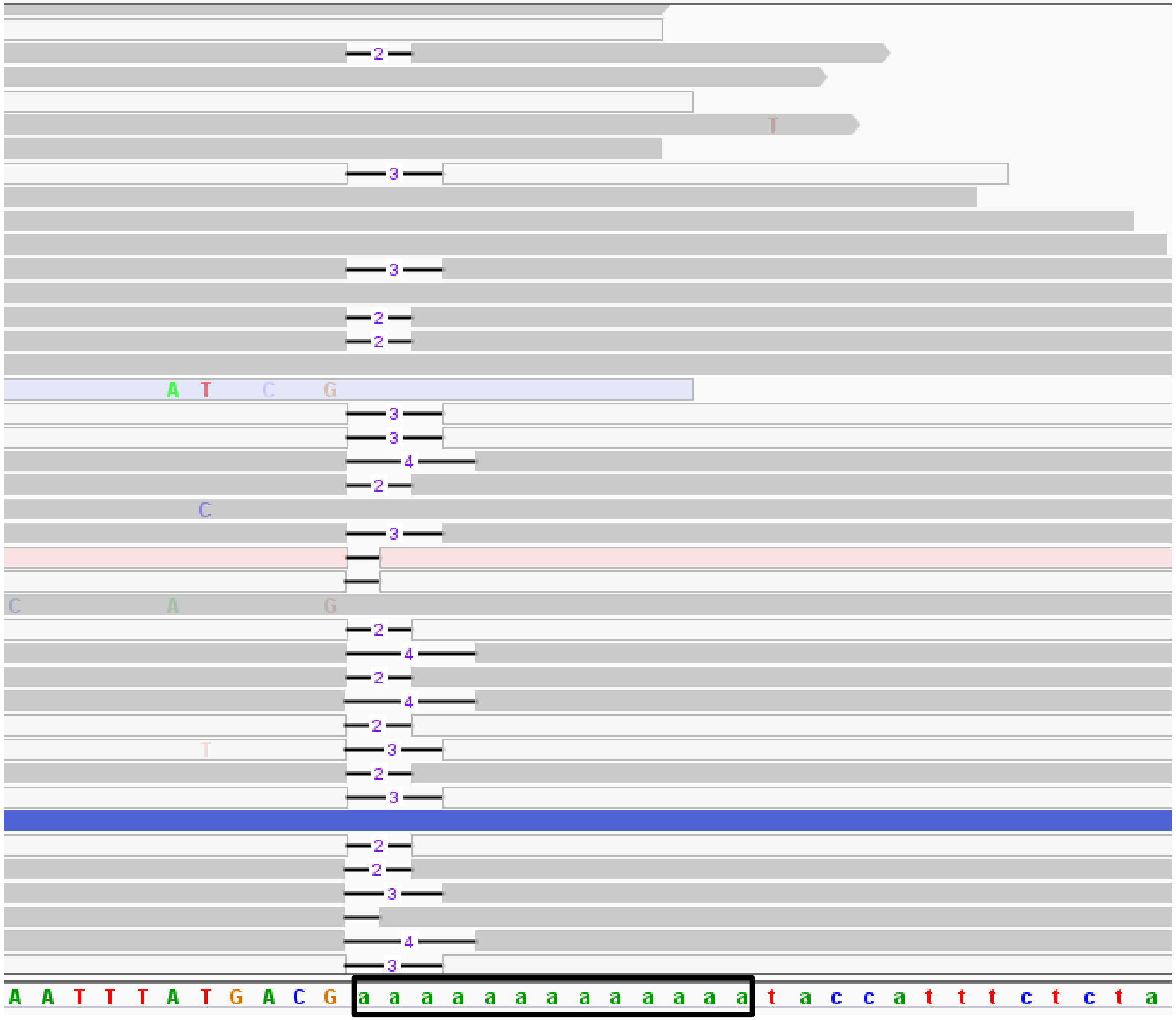
- Characterized by the “spontaneous loss or gain of nucleotides from repetitive DNA tracts”
  - Caused by defective mismatch repair
- Is a diagnostic phenotype for certain cancer types with clinical implications
  - Actionable marker for immune-checkpoint-blockade therapy

(Hause et al. 2016)

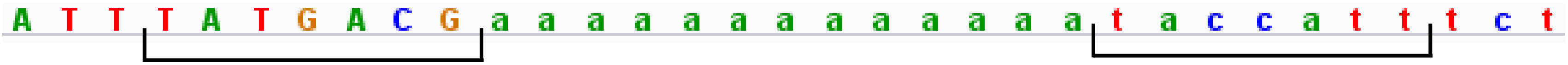
Sample

- 602 BAM/BAI files from patients in TCGA
  - 241 Colon/Rectal/Colorectal Adenocarcinoma
  - 361 Uterine Endometrial Carcinoma
- Annotated with MSI-PCR result
  - MSS
  - MSI-L, MSI-H
- 27 representative loci gathered from literature

Repeat Identification


2
T
3
3
2
2
A T C G
3
3
4
2
C
3
C A G
2
4
2
4
2
T 3
2
3
2
2
3
4
3
A A T T T A T G A C G a a a a a a a a a a a a a t a c c a t t t c t c t a

Repeat Identification


A T T T A T G A C G a a a a a a a a a a a a a t a c c a t t t c t

- Flanking regions
  - 7 bases long
    - Long enough to be unique, identifiable
    - Short enough to maximize usable reads
  - Immediately adjacent to the mononucleotide repeat

Repeat Identification

- Scan and find flanking regions
  - Ordered search
  - Brute-force fuzzy matching
    - Mismatch allowance = 2
- Validation: filter
  - Baseline = 90%
  - Modified for some loci

MSI Calling

- Binary classifier
  - Simplification of MSI-H and MSI-L
- Optimization
  - Accuracy
    - Sensitivity
    - Specificity
- Locus-based calls
  - More common in literature
- Overall MSI call based on all 27 loci

Machine Learning Approach

- TensorFlow
  - Python API developed by Google

- Minimize error in linear function, apply sigmoid curve to calculate p(MSI)

- Simplification
  - 5 loci with high number of BAMs with reads available, good depth at the loci

Datasets

- All files
  - 602 BAM files with MSS, MSI-L or MSI-H annotation
    - 241 COAD-READ, 361 UCEC
    - 212 MSI, 390 MSS
- Mode training set
  - Randomly generated set of 100 BAM files with MSS status to generate mode length
    - 43 COAD-READ, 57 UCEC
    - 0 MSI, 100 MSS

Datasets (continued)

- Training Set
  - 300 BAM files of mixed status to train the model
    - 118 COAD-READ, 182 UCEC
    - 130 MSI, 170 MSS
- Validation Set
  - 100 BAM files of mixed status to validate the model
    - 44 COAD-READ, 56 UCEC
    - 40 MSI, 60 MSS
- Test Set
  - 102 remaining BAM files of mixed status to test the model
    - 36 COAD-READ, 66 UCEC
    - 42 MSI, 60 MSS

Datasets (continued)

- Randomly generated sets
  - Representative of both cancer types

- Low coverage problem
  - Low coverage at certain loci excludes some files from some locus analyses
    - Actual size of datasets vary based on individual BAM coverage at any given locus

MSI Calling - Features

- Number of lengths
  - [11, 12, 11, 13, 14] = 4
- Distance from mode
  - Average distance from MSS sample mode of all reads
- Standard deviation
- Average Length

Locus-Based Calling

- Generate ML model for each locus individually
  - Weight for each feature, bias
  - Sigmoid activation function to generate a probability (0, 1)
- Choose loci with highest AUC
  - Exclude those that have no reads in many BAM files
  - Try different combinations of loci
- Determine MSI status based on number of loci with 'MSI' call

Locus-Based Calling (continued)

- Loci examined
  - BAT-26
  - MSI-07
  - MSI-09
  - H-06
  - MSI-06
  - MSI-04
  - HSPH1-T17

- Threshold: 0.500000
- Min no. loci: 3
- Total files: 93
- Correct predictions: 78
  - True pos: 24
  - True neg: 54
  - False pos: 2
  - False neg: 13
- Accuracy: **0.838710**
- Sensitivity: **0.648649**
- Specificity: **0.964286**

Probability-Based Calling

- Generate ML model for each locus individually
  - As before

- Choose loci with highest AUC
  - As before

- Determine MSI status based on average p(MSI) across loci examined

Probability-Based Calling (example)

- BAM file: TCGA-00-0000
  - BAT-26
    - p(MSI) = 0.998
  - MSI-07
    - p(MSI) = 0.488
  - MSI-09
    - p(MSI) = 0.478
  - H-06
    - p(MSI) = 0.499
  - MSI-06
    - p(MSI) = 0.898
  - MSI-04
    - p(MSI) = 0.978
  - HSPH1-T17
    - p(MSI) = 0.408

- Locus-based call: MSS
  - 3 MSI, 4 MSS

- Probability-based call: MSI
  - Avg p(MSI): 0.678

Probability-Based Calling (continued)

- Loci examined
  - BAT-26
  - MSI-07
  - MSI-09
  - H-06
  - MSI-06
  - MSI-04
  - HSPH1-T17

- Threshold: 0.500000
- Min no. loci: 3
- Total files: 93
- Correct predictions: 78
  - True pos: 23
  - True neg: 55
  - False pos: 1
  - False neg: 14
- **Accuracy: 0.838710**
- **Sensitivity: 0.621622**
- **Specificity: 0.982143**

Probability-Based Calling (continued)

- Problem: some loci may be more indicative of MSI status than others
    - Need weights assigned to each locus individually

A Bigger ML Problem

- Solution: instead of individually looking at one locus with *m* features, consider all *n* loci with all *m* features at once
  - Machine learning problem in *n* x *m* dimensions
- Exclude loci with too few usable files

Computational Cost

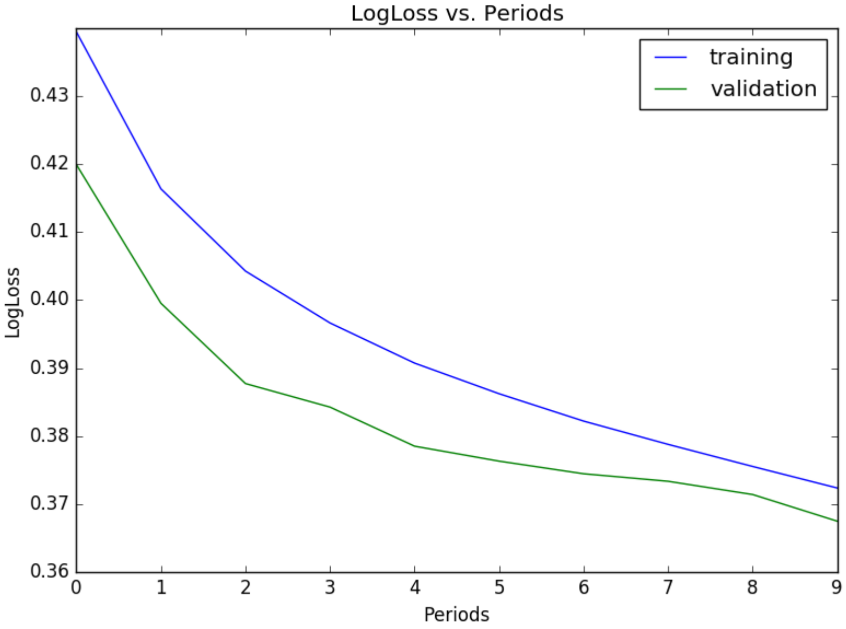
- 21 loci x 4 features each
  - 84-dimension model

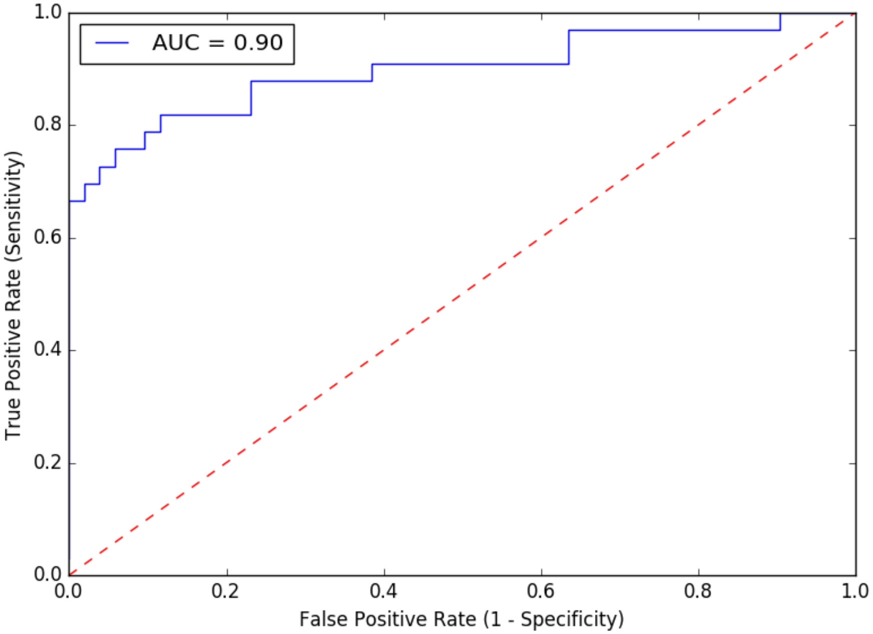
- L1 regularization
  - Drives weights of non-influential features to 0
  - Minimizes the number of features that contribute to the model

Machine Learning v2 - Sets

- Training Set
  - 124 MSI, 142 MSS
- Validation Set
  - 33 MSI, 52 MSS

Machine Learning v2 - Results





Machine Learning v2 – Test Set

- Loci examined
  - MSI-11
  - MSI-14
  - H-10
  - HSPH1-T17
  - BAT-26
  - BAT-25
  - MSI-04
  - MSI-06
  - MSI-07
  - MSI-01
  - MSI-03
  - MSI-09
  - H-09
  - H-08
  - H-01
  - H-03
  - H-02
  - H-04
  - H-07
  - H-06
  - H-05

- Threshold: 0.500000
- Total files: 85
- Correct predictions: 73
  - True pos: 27
  - True neg: 46
  - False pos: 6
  - False neg: 6
- Accuracy: 0.858824
- Sensitivity: 0.818182
- Specificity: 0.884615

Machine Learning v2 – Test Set

- Loci examined
  - MSI-11
  - MSI-14
  - H-10
  - HSPH1-T17
  - BAT-26
  - BAT-25
  - MSI-04
  - MSI-06
  - MSI-07
  - MSI-01
  - MSI-03
  - MSI-09
  - H-09
  - H-08
  - H-01
  - H-03
  - H-02
  - H-04
  - H-07
  - H-06
  - H-05

- Upper threshold: 0.550000
- Lower threshold: 0.450000
- Total files: 85
- Indeterminate files: 9
- Predictions: 76
- Correct predictions: 67
  - True pos: 25
  - True neg: 42
  - False pos: 3
  - False neg: 6
- **Accuracy: 0.881579**
- **Sensitivity: 0.806452**
- **Specificity: 0.933333**
- **Indeterminate: 10%**

Summary

- Maximum accuracy of 88% on test set
  - 10% labeled as ‘Indeterminate’

- Machine learning using all loci and all features is most:
  - Accurate
  - Sensitive/Specific

Further Investigations

- Add a label for ‘MSI-L’ and ‘MSI-H’
- Research clinical relevance of MSI diagnosis
  - Fine-tune parameters to favor either false positives or false negatives
- Add features
  - Earth mover distance
- Regularization
  - L2 regularization
    - Penalizes any weight that is too large, drives weights asymptotically to 0
  - L1 regularization
    - Penalizes total weight of all features, drives non-informative weights to 0

Further Investigations (continued)

- Get more BAM files
  - Training, validation and test sets are stale
  - Get BAM files with better coverage for more accurate results, better prediction

- Make a more complex model
  - Synthetic features
  - Feature crosses
  - Neural Nets
    - More complex models require additional samples to train, validate and test