Differential Methods in Modern Biological Data Analysis

Biostatistics Program

Ph. D. Dissertation Defense,

Micah Thornton Ph.D. Candidate

September 3rd 2021

Southern Methodist University
(Department of Statistical Science)

University of Texas Southwestern
(Department of Population and Data Sciences)
(Department of Bioinformatics)

Harmonic Analysis for Sequence Data

Part 1/3

Slides – [2-20]

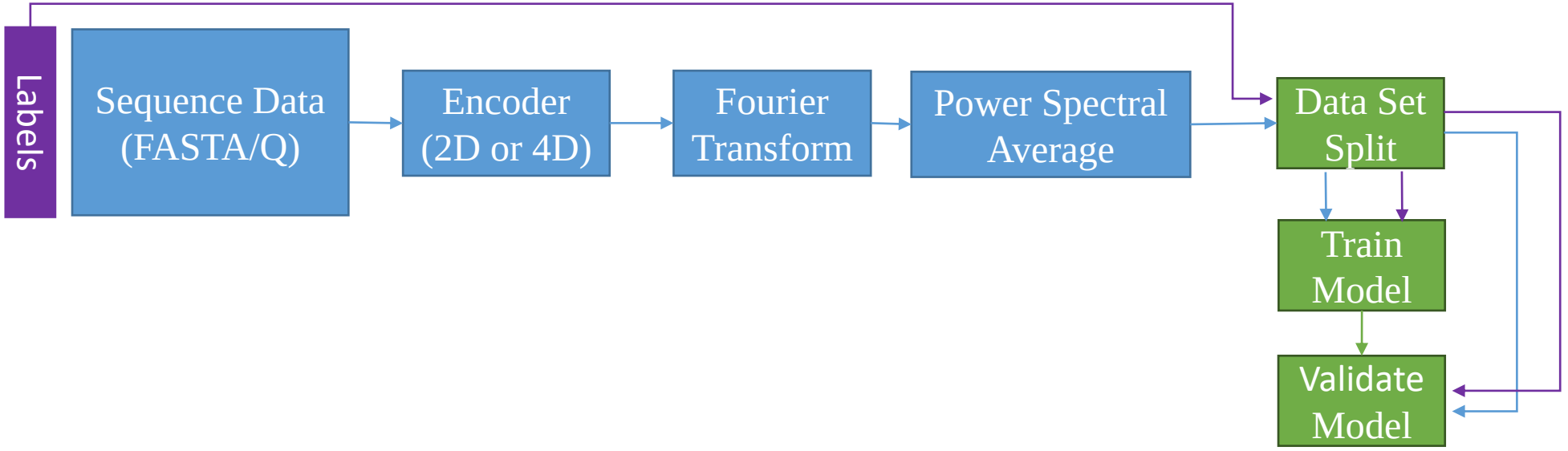
Questions – Harmonic Analysis

1. Does harmonic information, through Fourier coefficients and related spectra, allow
    1. meaningful **characteristic classification** with standard approaches,
    2. or delineate useful **clusters?** How does this relate to other approaches?

Hypotheses & Procedures (1)

1.1) Can **harmonic analysis** be applied to genetic sequences to **classify characteristics** with standard numerical approaches?

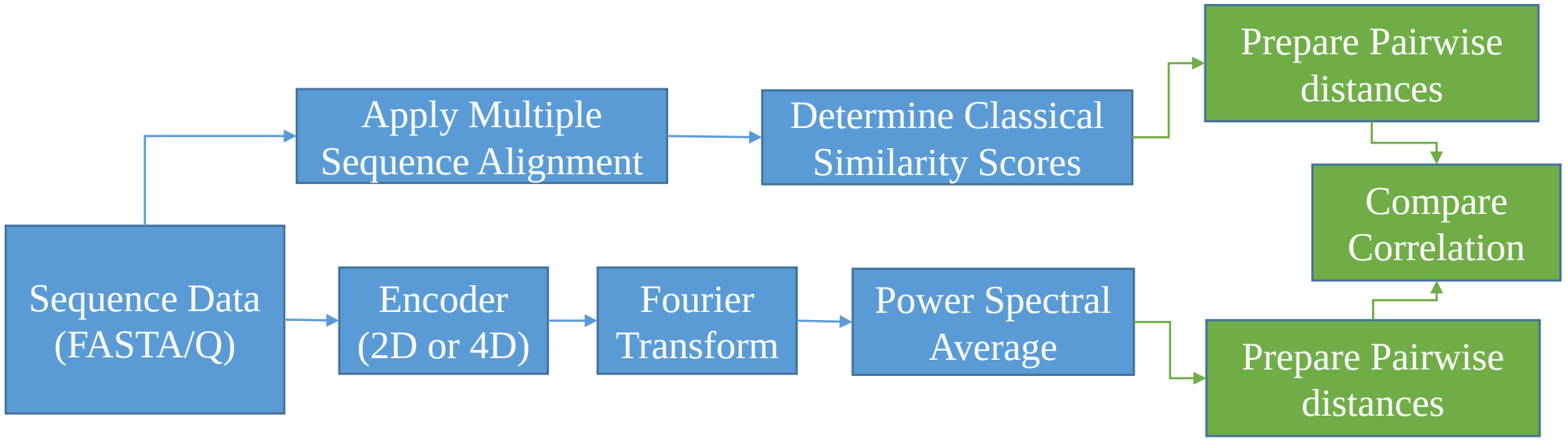
- Hypothesis: The **Fourier coefficients provide** summary **characteristics** of genetic sequencing data **suitable for classifying some attributes** of the original data.



Hypotheses & Procedures (1)

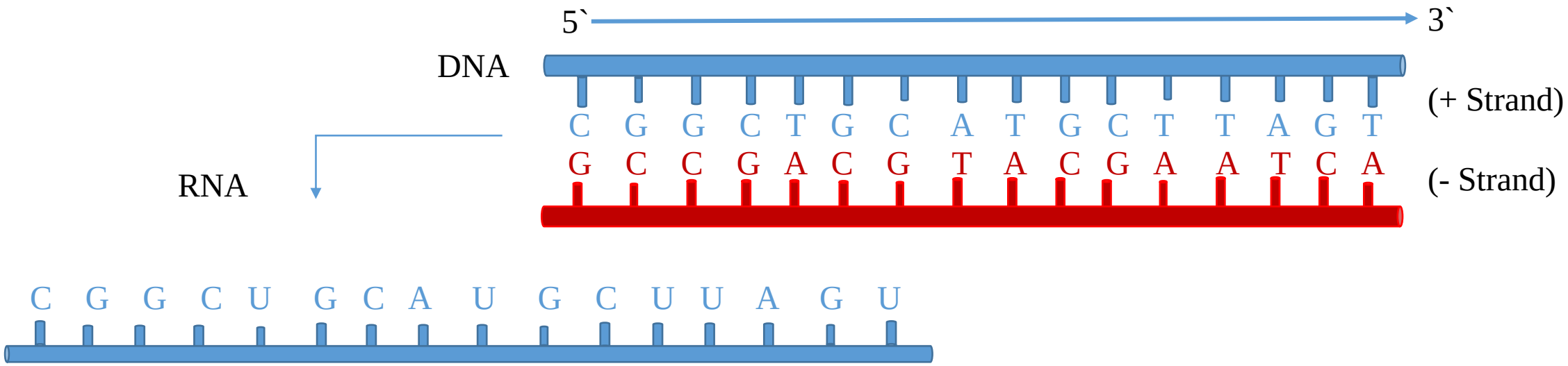
1.2) Can **harmonic analysis** be applied to genetic sequences to **indicate useful clusters**?

- Hypothesis: Clusters **can be determined** using standard approaches **with the power spectra**.



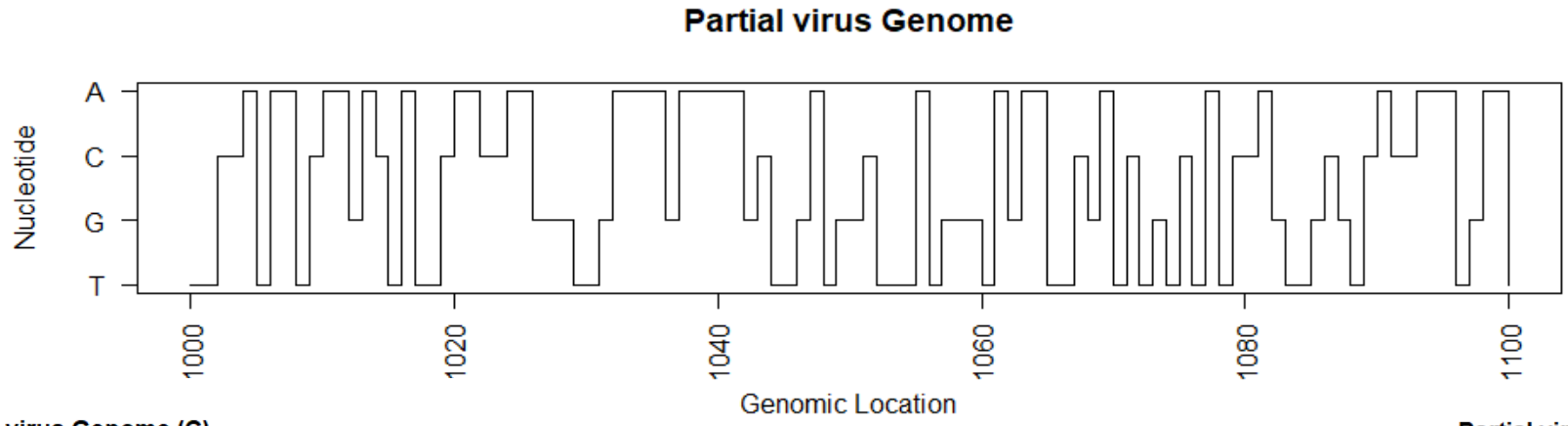
Background - Genetic Sequence Data

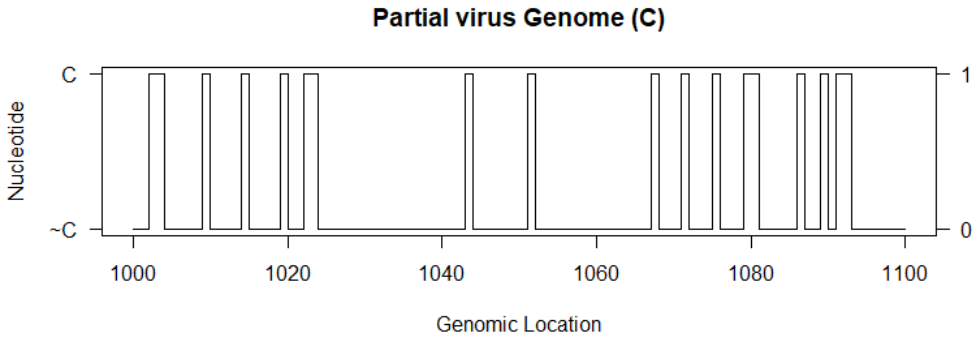
- Has directional information and an associated observation in each category.

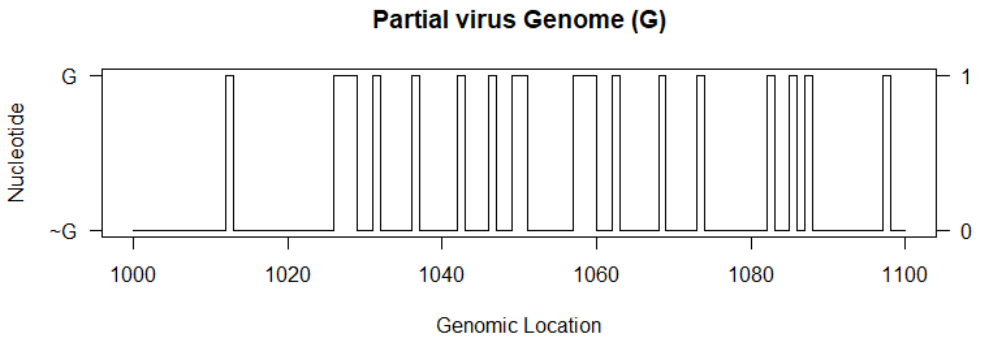


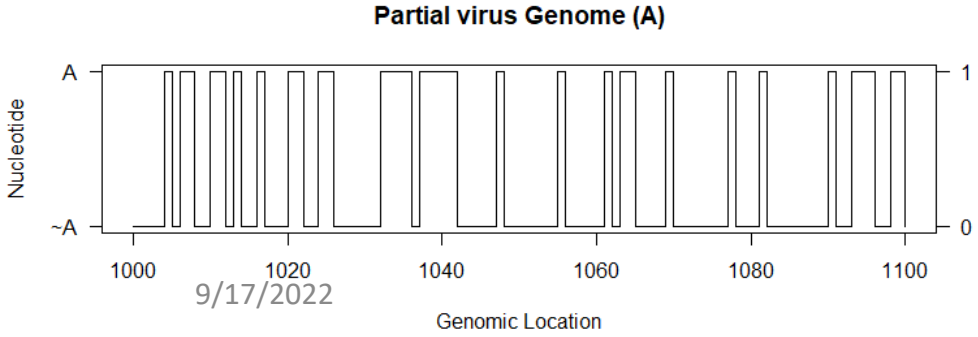
- Usually stored in text-based files that contain ordered letters indicating the nucleotide at a specific location.

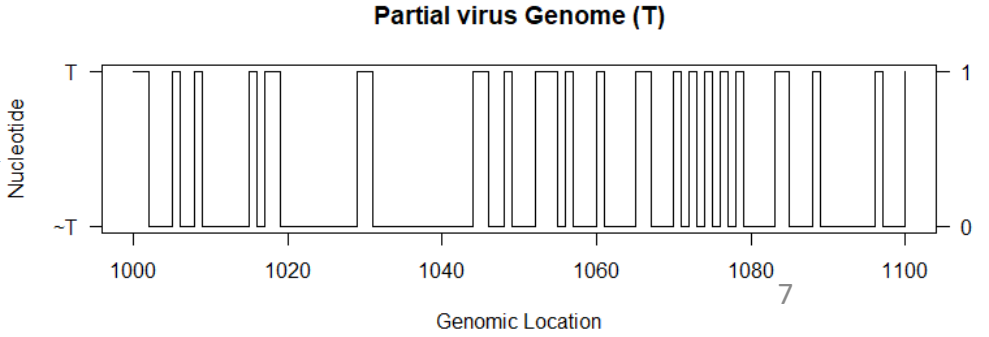
Taking Fourier Transforms of Genomic Signals



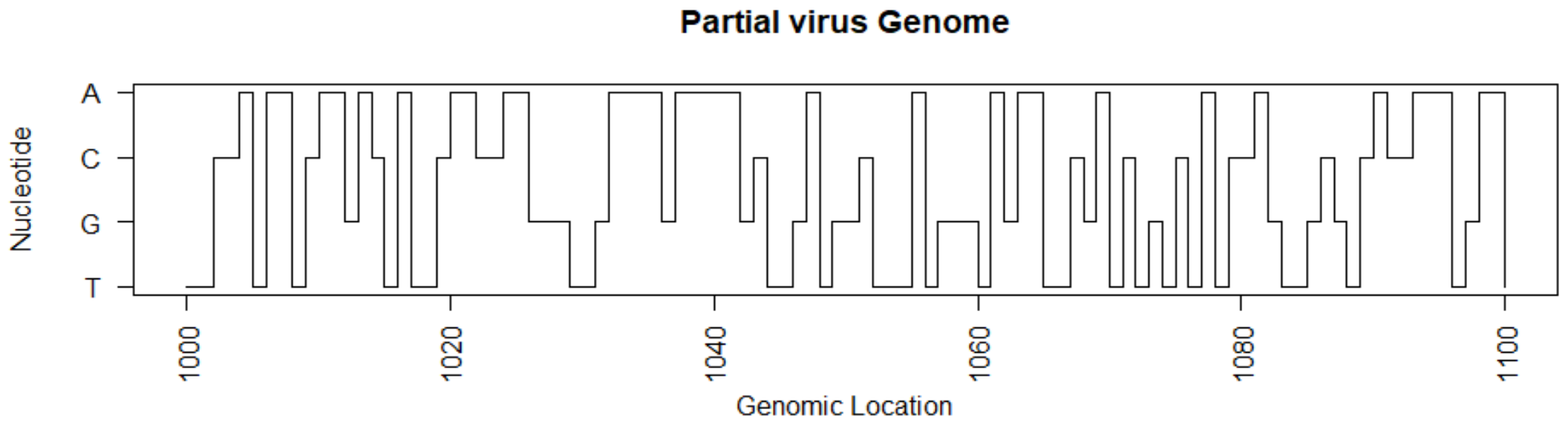


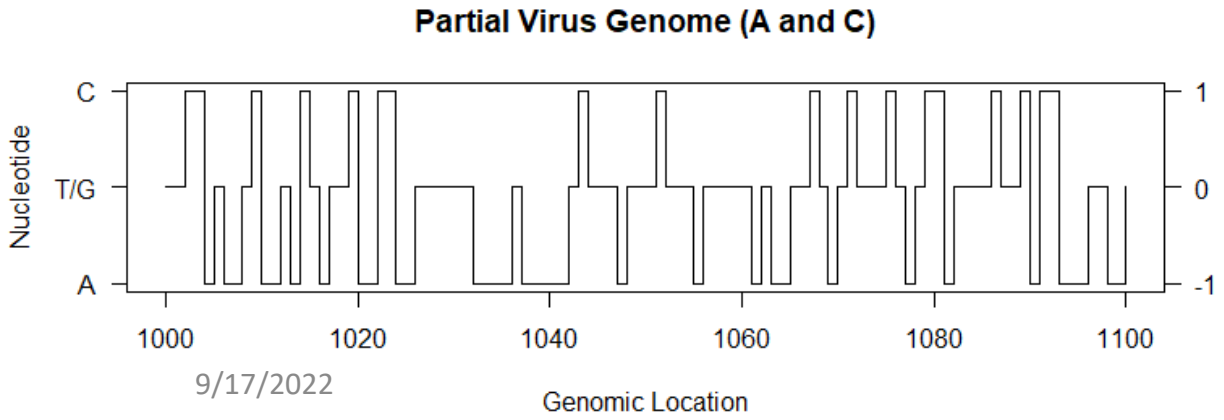


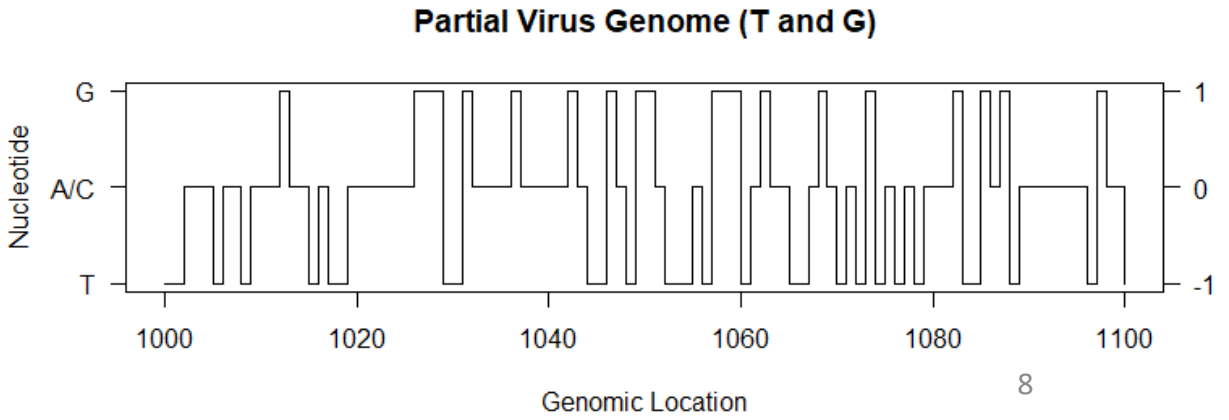




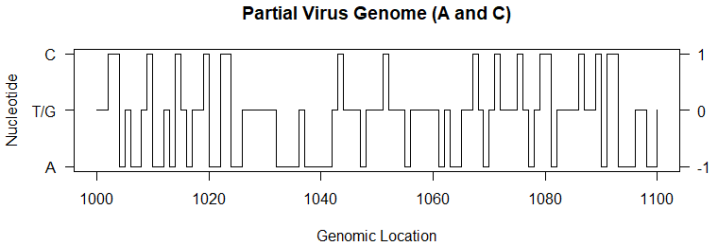
Taking Fourier Transforms of Genomic Signals

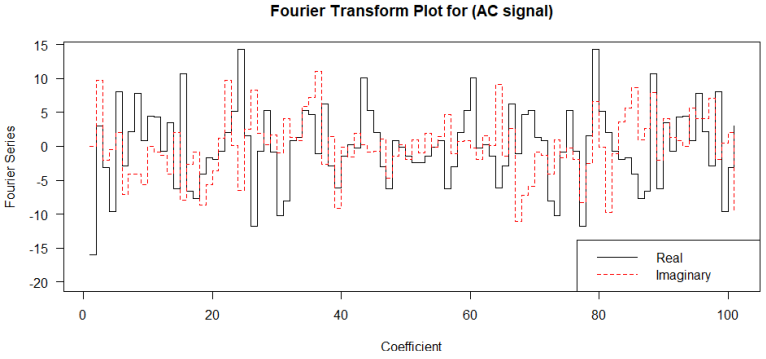


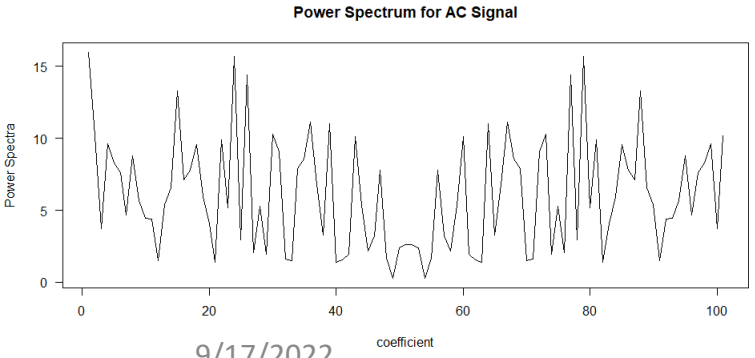




Taking Fourier Transforms of Genomic Signals



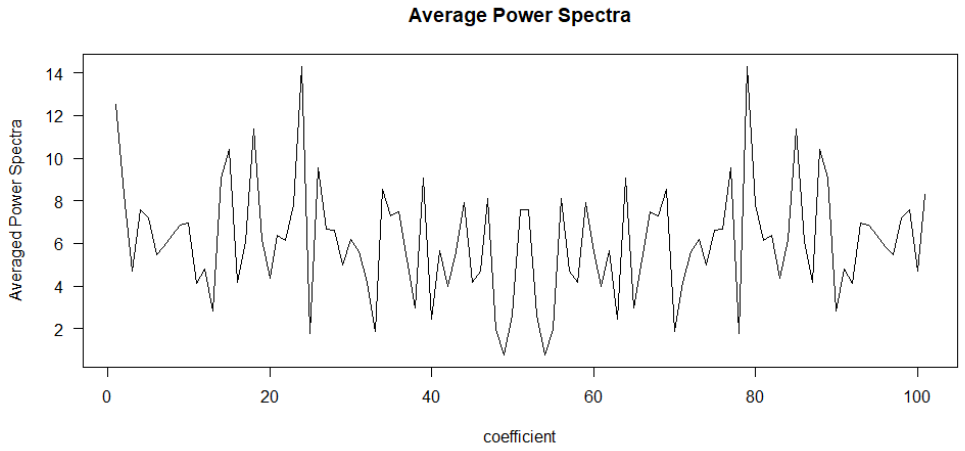


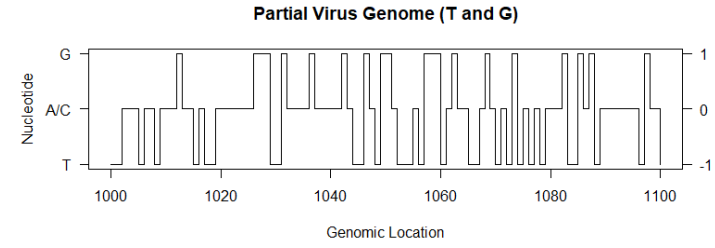


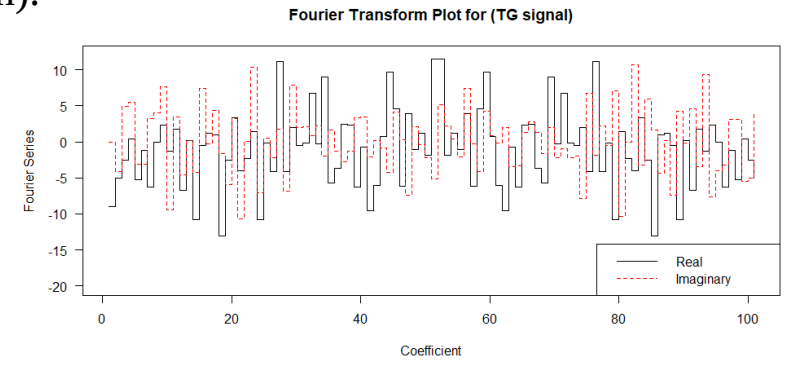
Discrete Fourier Transform

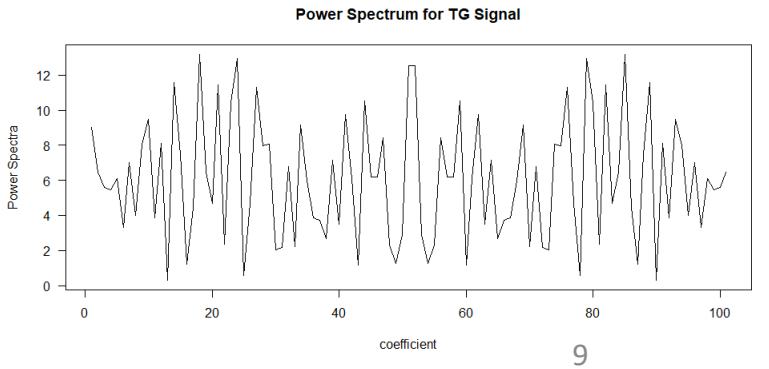
$$S(h) = \sum_{t=1}^{T} s(t) \cdot e^{-2\pi j(t-1)(h-1)(T^{-1})}$$

- π is the ratio of circumference to diameter (Euclidean).
- $e = \lim_{n\to\infty}(1 + n^{-1})^n$
- $j^2 = -1$
- T is the signal length,
  - $S(h)$ the h-element of the Fourier Series
  - $s(t)$ the t-element of the signal.

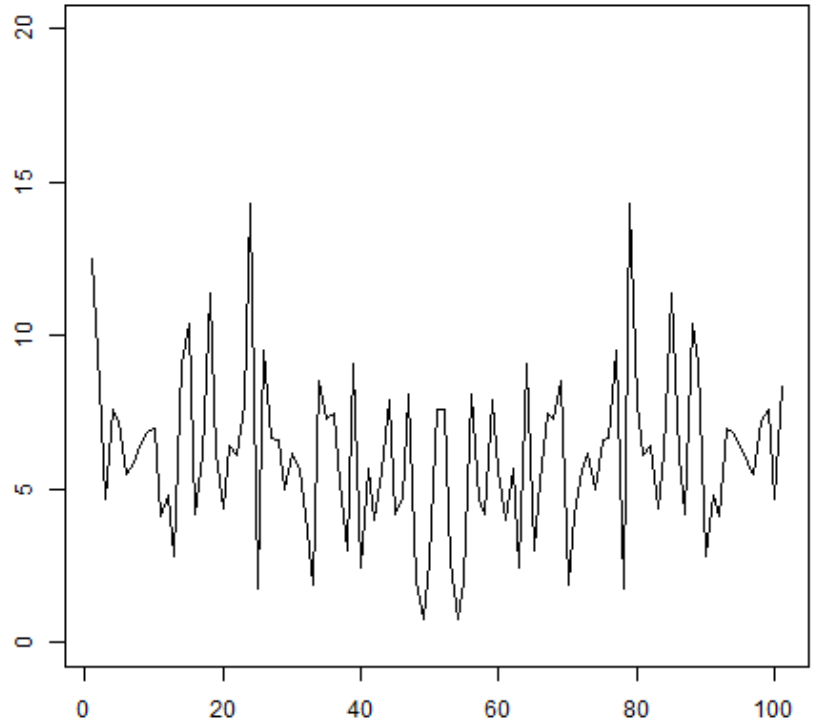








Scaling Genomic Power Spectra

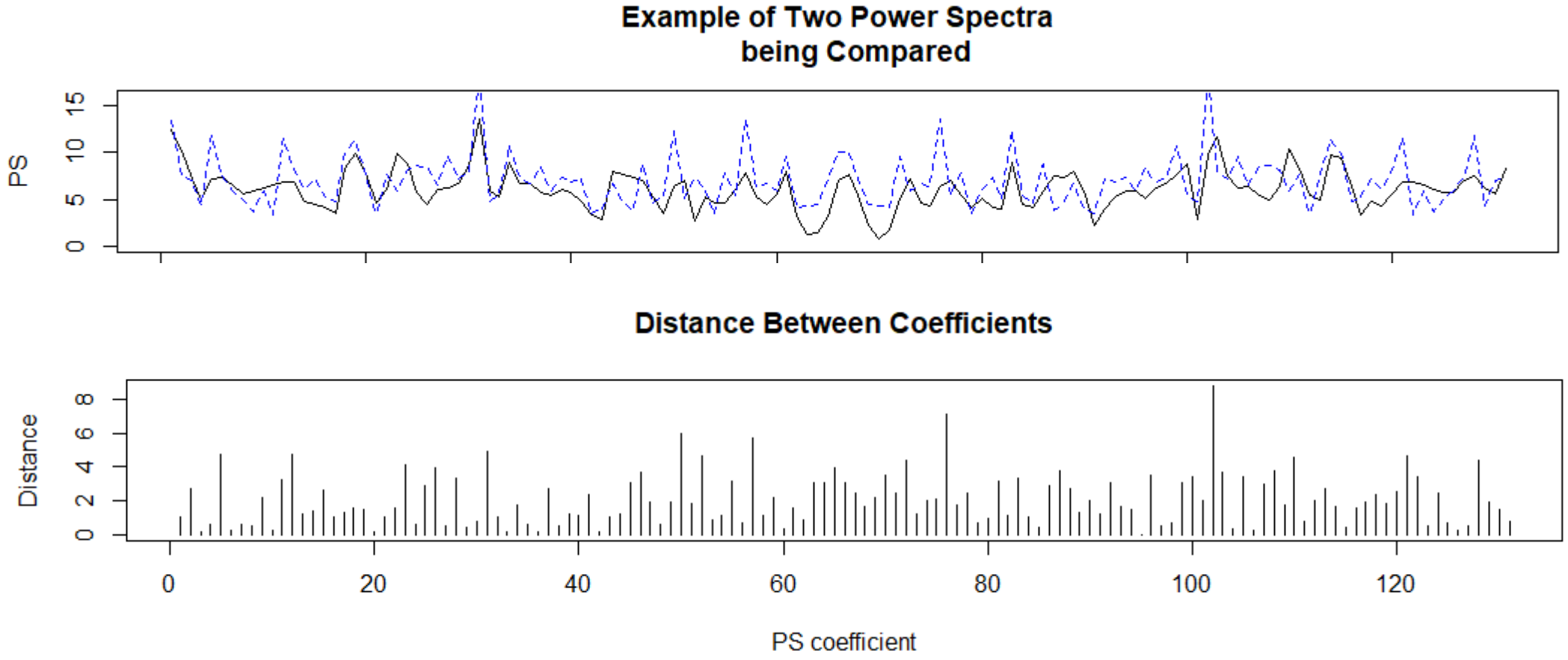


- When comparing two or more sequences, a natural distance would be the Euclidean distance between the two sequence power spectra.
  - Some Spectra will have to be Scaled to compute distances
  - Even Scaling Procedure from Yin and Yau 2015 is implemented in this study.

$$A_m(k) = \begin{cases} A_n(Q) & Q \in \mathbb{Z} \\ A_n(R) + (Q - R)(A_n(R+1) - A_n(R)) & Q \notin \mathbb{Z} \end{cases}$$

- Where, $Q = \frac{kn}{m}$, $R = \left\lfloor \frac{kn}{m} \right\rfloor$, and m and n represent the longer and shorter lengths respectively.

Computing distance between genomic PS



Data (SARS-CoV-2 Genomes)

- 1,397 virus genetic sequences for SARS-CoV-2 Genomes submitted from various collecting laboratories around the world.
- Sequences geographic origin information contained in the header for each observation.

- GISAID Initiative collected and maintained sequence data for download and analysis.

Data (SARS-CoV-2 Genomes)

Location	Observations (%)
Africa	35 (2.5)
East Asia	257 (18.4)
Europe	678 (48.5)
Middle East	153 (11)
North America	42 (3)
Oceania	38 (2.7)
South America	89 (6.4)
West Asia	105 (7.5)

Africa	East Asia	Europe	Middle East	North America	Oceania	South America	West Asia
Algeria	Beijing	Austria	Turkey	USA	Australia	Brazil	Bangladesh
Egypt	Chongqing	Belgium	Saudi Arabia	Puerto Rico	New Zealand	Chile	Cambodia
South Africa	Fujian	Czech Republic	Kazakhstan	Guam	Indonesia	Colombia	India
DRC	Guangdong	Denmark	Iran	Mexico	Malaysia	Costa Rica	Nepal
Gambia	Hangzhou	Finland	Israel			Uruguay	Sri Lanka
Senegal	Hong Kong	France	Kuwait				Vietnam
	Jiangsu	Georgia					Thailand
	Jiangxi	Germany					
	Jingzhou	Greece					
	Shandong	Spain					
	Shenzhen	Sweden					
	Sichuan	Hungary					
	Taiwan	Portugal					
	Tianmen	Poland					
	Wuhan	Russia					
	Yunnan	Romania					
	Zhejiang	Slovakia					
	Lishui	Italy					
	Japan						
	Guangzhou						

Multi-Class procedure results (10-fold CV – 138/fold)

Classification Scheme	Overall-Accuracy	Average Sensitivity	Average Specificity	CPU Time (s)
ECOC-SVM	0.4624	0.8072	0.2065	2547
Random Forest (25 Trees)	0.7802	0.6019	0.9503	238
Random Forest (50 Trees)	0.7881	0.6195	0.9523	473
Random Forest (100 Trees)	0.7953	0.6210	0.9534	938
Random Forest (500 Trees)	0.7967	0.6238	0.9335	4676
Random Forest (1000 Trees)	0.7996	0.6243	0.9544	9431
Multinomial Logistic Regression (50 Coefficients)	0.3958	0.1277	0.8099	10792
Multinomial Logistic Regression (100 Coefficients)	0.3257	0.1328	0.7583	44942
Neural Network (1 Hidden Layer, 100 Neurons)	0.6707	0.4640	0.9209	2245
Neural Network (1 Hidden Layer, 250 Neurons)	0.6779	0.4982	0.9249	5267
Neural Network (1 Hidden Layer, 500 Neurons)	0.6707	0.4827	0.9217	12255
Neural Network (1 Hidden Layer, 1000 Neurons)	0.6521	0.4722	0.9164	23012
Neural Network (2 Hidden Layers, 250, 150 Neurons)	0.6679	0.4642	0.9201	3905
Pseudo-Quadratic Discriminant Analysis (1000 coefficients)	0.1102	0.2801	0.7495	49
Pseudo-Quadratic Discriminant Analysis (3000 coefficients)	0.073	0.1637	0.7637	231
Pseudo-Quadratic Discriminant Analysis (5000 coefficients)	0.0709	0.1536	0.7640	966

10 Fold CV for SARS-CoV-2 Data

Supervised Learner	k-Mer Vectors (Interval)	DFT Power Spectra (Interval)
Naive Bayes	0.424 (0.411, 0.438)	**0.593 (0.580, 0.607)**
Regression Tree	0.179 (0.169, 0.189)	0.191 (0.181, 0.202)
K-Nearest Neighbors ($k = 10$)	0.722 (0.710, 0.734)	**0.776 (0.765, 0.787)**
Random Forest (500)	0.651 (0.639, 0.664)	**0.805 (0.795, 0.816)**
Neural Network (1 HL - 30 N)	0.505 (0.492, 0.519)	**0.580 (0.567, 0.593)**
SVM	0.688 (0.676, 0.700)	0.712 (0.699, 0.724)

- **Random Forest (500 trees) - best regional classifier**,
  - Better results for the DFT

- DFT Power Spectra provide better criteria, the intervals provide (accuracy estimation +/- standard error).

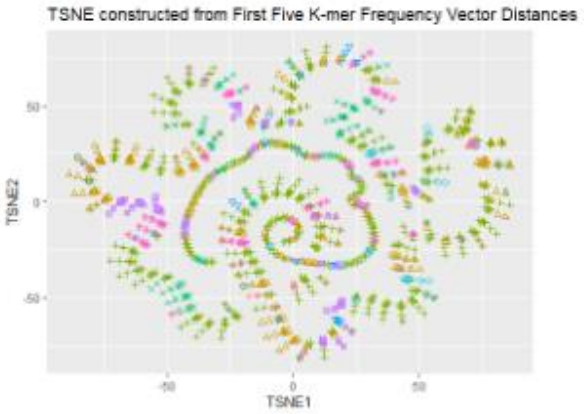
Conclusions (1.1)

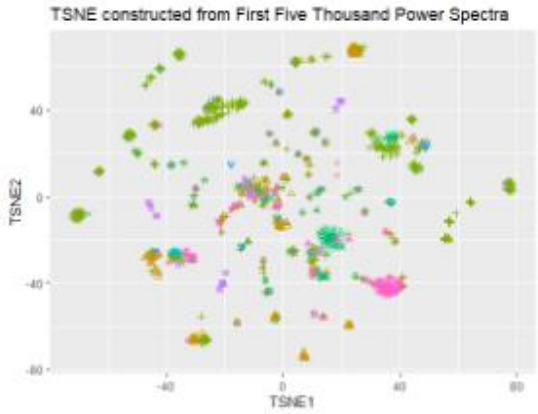
- Regional variations in virus data **can** be numerically summarized by the Power Spectrum.
  - The relationship can later be learned by standard supervised approaches with up to 80% accuracy for differentiating SARS-CoV-2 genome regions.
- Due to even-scaling procedure the PS computation procedure does not require alignment prior to computation.
  - K-mer counting also does not require alignment, therefore k-mer count vectors (for k = 1,2,…,5) provide an alternative set of numerical values on which to categorize sequences.
  - These k-mer count vectors do not provide as good of values (although there are admittedly less of them) for class differentiation.

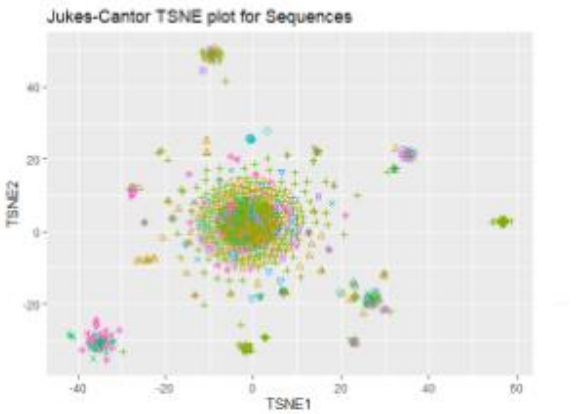
Clustering Sequences

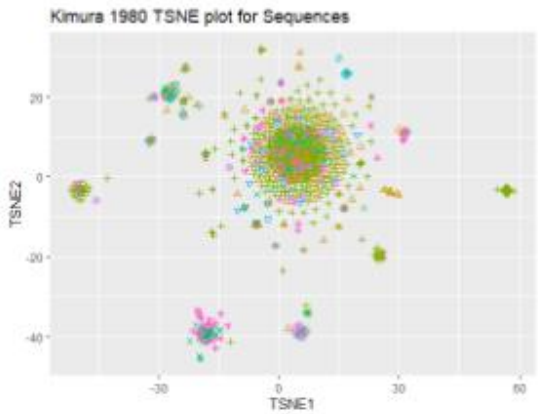
- Visualization of the sequences by their Power Spectra is possible through the usual techniques:
  - T-SNE
  - PCA
  - UMAP

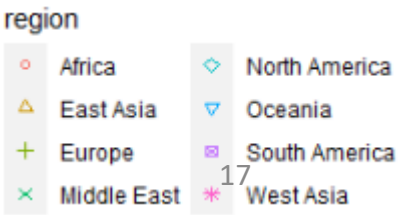
- Statistical Procedures are also applicable to the power spectra,
  - MANOVA – Are the multivariate mean power spectra the same across classes or different?







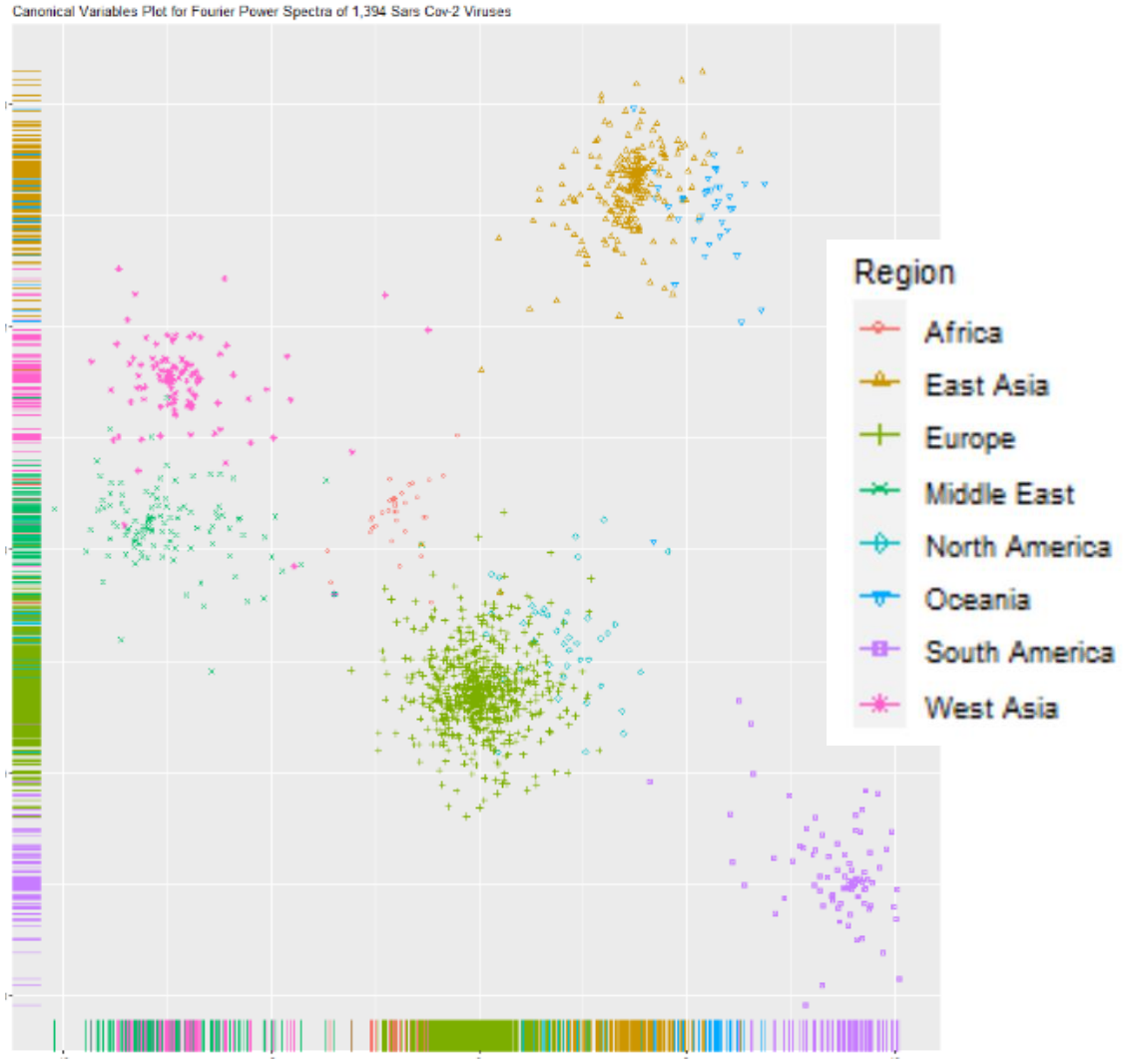




Method	Pilai Statistic	Approximate F	Numerator DF	Denominator DF	p-value
k-mer, $k = 5$	2.17	5.84	700	9072	$< 2 \times 10^{-16}$
FCPS	6.49	5.05	7000	2772	$< 2 \times 10^{-16}$

Clustering Sequences

- The ability to apply numerical procedures to the power spectra is very useful in clustering the data
- The distance calculation for power spectra is a simple and quick procedure.
- Some other distance estimation techniques require alignment of sequences to each other prior to computation, these are much more expensive by comparison.



Canonical Variables Plot (Created with Labels)

Conclusions (1.2)

- The distance metrics produced by the PS capture different information from some more classical techniques (such as Markov approaches like the JC).
- A lot of the information can be summarized by a filtered subset of the power spectra, a few coefficients/components
- Supervised filters created by machine learners may not always provide the best differentiators.

Predictive and Explanatory Modeling of Compositional Protein Data

Part 2/3

Slides – [21-31]

Questions – Compositional Data

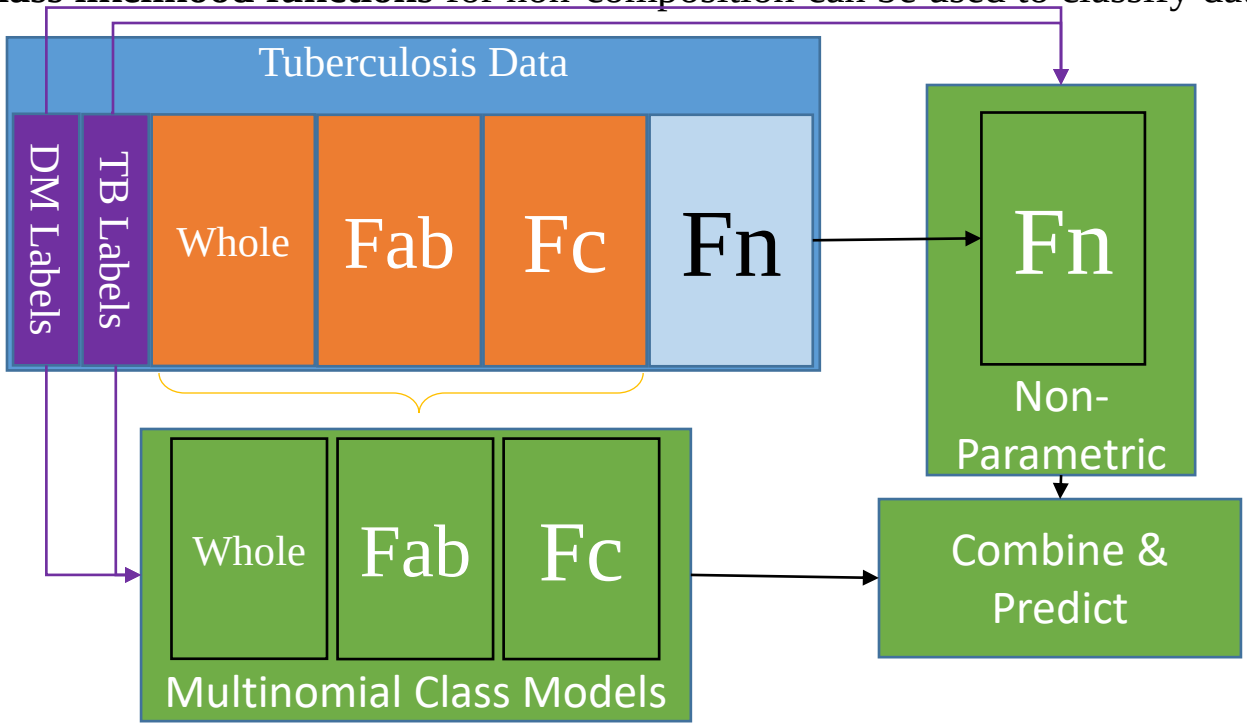
2. What are some statistical approaches for relating compositional glycan data to disease outcome?

1. Can a **semi-parametric model** utilizing multinomial likelihoods give reasonable classification estimates?
2. Could a transition-like **glycan rank proportion** model provide a valid classification procedure, that selects important pairs of glycans?

Hypotheses & Procedures (2)

2.1) Can a semi-parametric model utilizing multinomial likelihoods give reasonable classification estimates?

- Hypothesis: A Semi-Parametric approach which
  - combines **parametric likelihood functions** for capturing the composition contributions with
  - **empirical class likelihood functions** for non-composition can be used to classify data like this.



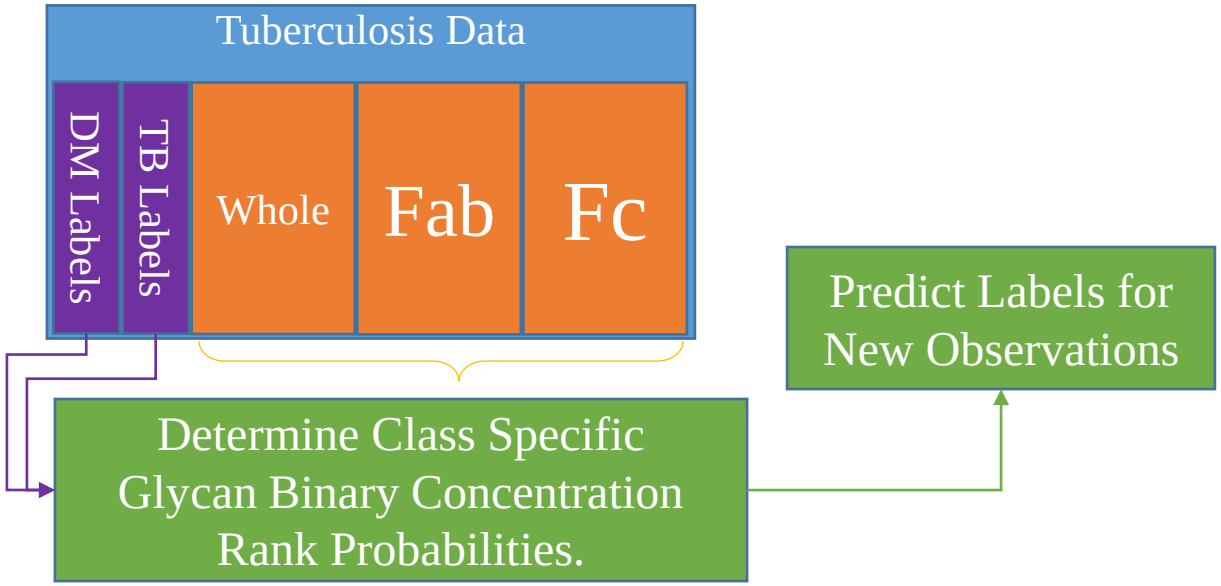
Hypotheses & Procedures (2)

2.2) Could a transition-like glycan rank proportion model provide a valid classification procedure, that selects important pairs of glycans?

- Class associated glycan rank probabilities can be used for prediction, and to determine the importance of pairings.

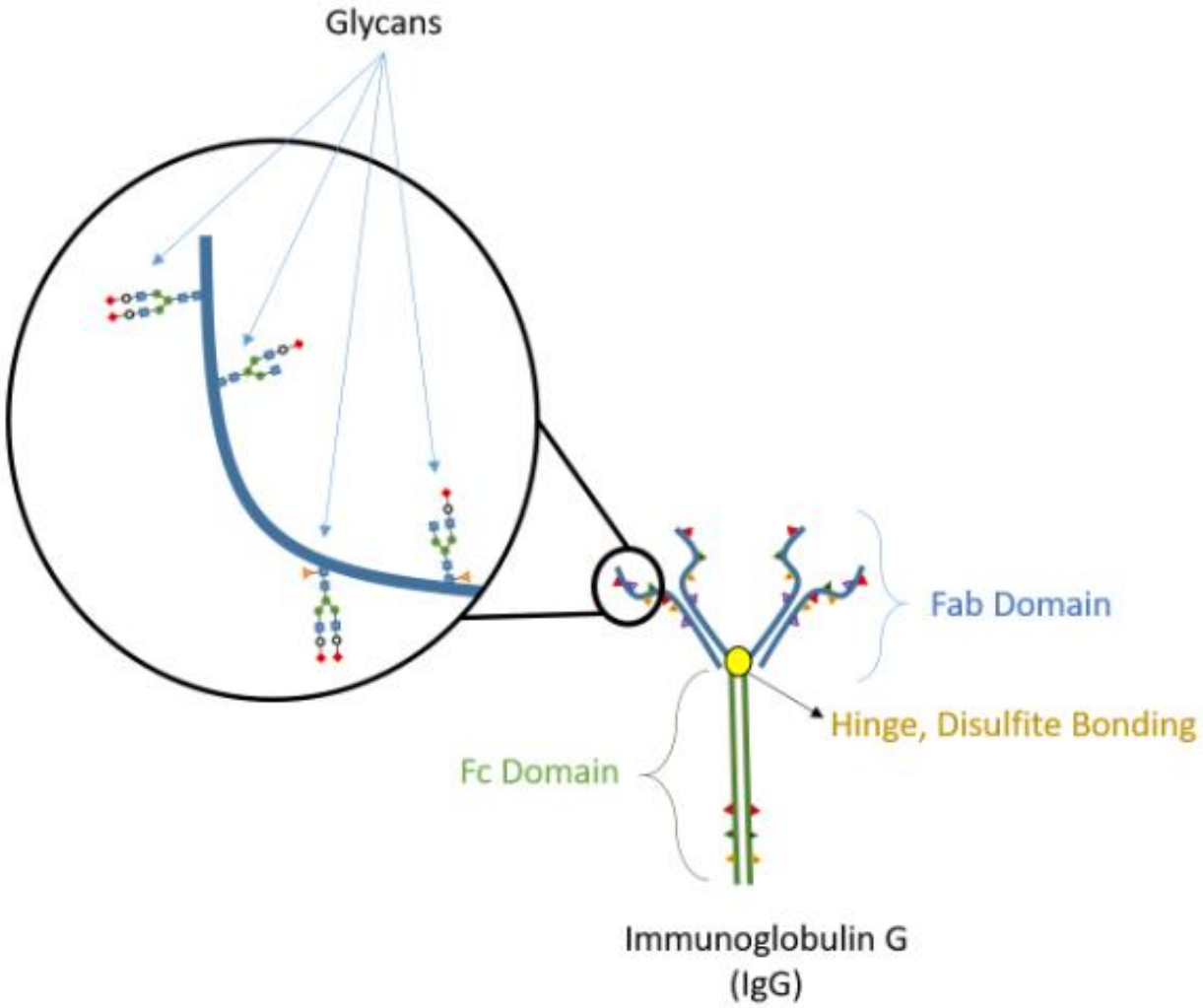
Whole – DM/ATB

	G2S2	G2S1 ….
Pat. 1	**10206**	3799
Pat. 2	**99428**	36322
…		



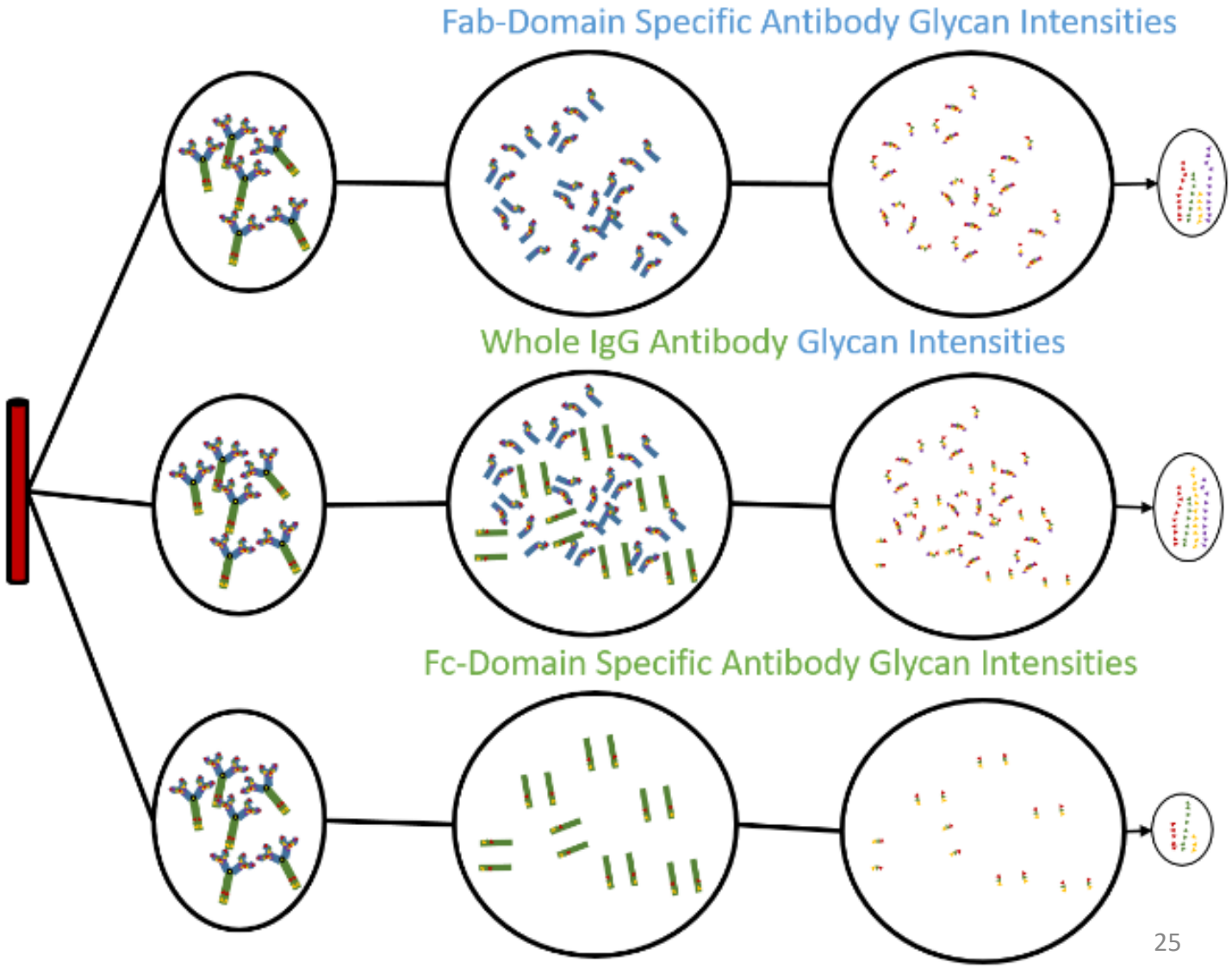
Glycan Data (Compositional Nature)

- Developed analyses could be used for any **compositional data**.
- Glycan data is available for some tuberculosis patients,
- Can compositional data models be used to differentiate disease outcomes?



Glycan Data (Compositional Nature)

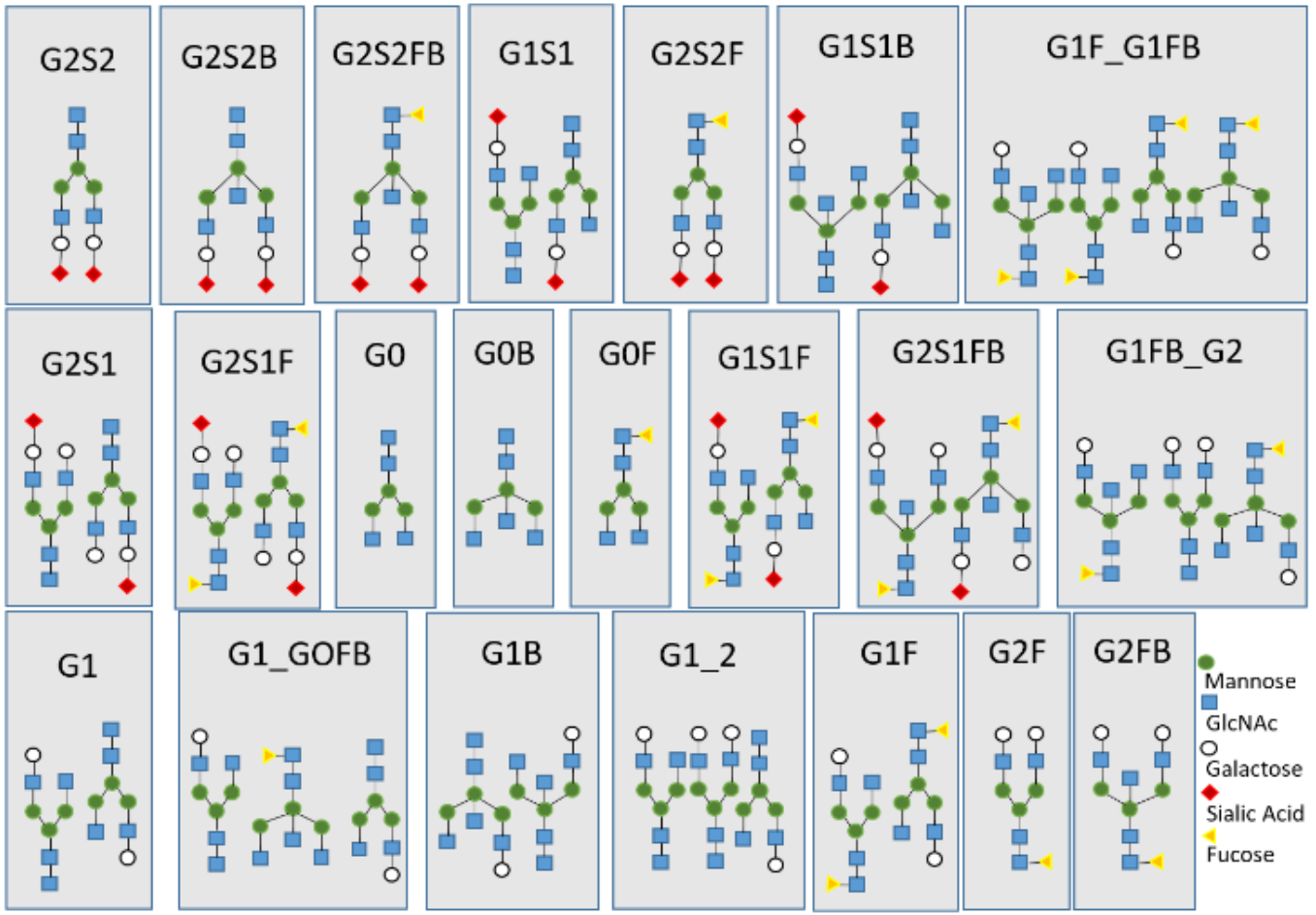
- Compositions for three different treatments.
- Three types of compositional data, three total compositions for modeling.



Glycan Data

- 21 categories per composition
- The models proposed should be general enough to encode arbitrary numbers of compositions and composition elements.

Figure 3.3. Types of Glycan Structures



Semi-Parametric Model

Within Composition Log Likelihood (Multinomial)

$$\ell(\boldsymbol{\pi_{(k)}}(C)|\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_N}) = \sum_{i=1}^{N} \left(\log \left(\sum_{j=1}^{J_k} x_{ij} \right) - \sum_{j=1}^{J_k} \log (x_{ij}!) + \sum_{j=1}^{J_k} x_{ij} \log \left(\pi_{j(k)}(C) \right) \right)$$

Outside Composition (Kernel Density Estimation within Classes)

$$\hat{f}_n(z) = \frac{1}{q} \sum_{i=1}^{q} \frac{1}{h} R \left(\frac{z - Z_i}{h} \right)$$

$$\log \left(P_{\text{nonpara}}(\tilde{c} = c) \right) \propto \sum_{l=1}^{L} \log \left(\hat{f}_{nl(c)}(\tilde{x_l}) \right)$$

Normalize to Combine Parametric/Nonparametric Score

$$K^*_{\text{para}}(c) = \frac{\log(P_{\text{para}}(\tilde{c} = c))}{\sum_{t=1}^{p} \log(P_{\text{para}}(\tilde{c} = c_t))}$$

$$K^*_{\text{nonpara}}(c) = \frac{\log(P_{\text{nonpara}}(\tilde{c} = c))}{\sum_{t=1}^{p} \log(P_{\text{nonpara}}(\tilde{c} = c_t))}$$

$$K^*(c) = K^*_{\text{para}}(c) + K^*_{\text{nonpara}}(c)$$

Select the Minimizing Class

$$\hat{c} = \{c | K^*(c) = \min \left(K^*(c_t) \ \ \forall \ \ t \in \{1, 2, \ldots, p\} \right)\}$$

Semi-Parametric Model Results

5-Class

	True Class	Predicted Class				
		ATB/DM	*ATB/ND*	*LTB/DM*	*LTB/ND*	*Neg/ND*
41 % Nonparametric Score Only	*ATB/DM*	7	9	0	2	0
	ATB/ND	7	12	2	3	1
	LTB/DM	0	3	0	2	2
	LTB/ND	1	4	1	9	3
	Neg/ND	1	0	3	5	6
41 % Parametric (Multinomial) Score Only	*ATB/DM*	15	2	1	0	0
	ATB/ND	10	4	1	5	5
	LTB/DM	3	0	0	3	1
	LTB/ND	0	5	2	10	1
	Neg/ND	1	5	2	2	5
43 % Combined Semiparametric Score	*ATB/DM*	14	3	0	0	1
	ATB/ND	11	8	1	3	1
	LTB/DM	2	2	0	3	2
	LTB/ND	0	3	3	9	2
	Neg/ND	1	2	2	4	5

Semi-Parametric Model Results

	True Class	Predicted Class		
		ATB	LTB	Neg
58% Nonparametric Score Only	ATB	32	10	1
	LTB	8	11	6
	Neg	0	10	5
57% Parametric (Multinomial) Score Only	ATB	31	7	5
	LTB	10	10	5
	Neg	5	4	6
65% Combined Semiparametric Score	ATB	34	8	0
	LTB	5	16	3
	Neg	3	9	2

3-Class (Tuberculosis)

2-Class (Diabetes)

	True Class	Predicted Class	
		DM	ND
62% Nonparametric Score Only	DM	8	17
	ND	15	43
67% Parametric (Multinomial) Score Only	DM	22	3
	ND	24	34
68% Combined Semiparametric Score	DM	20	5
	ND	21	34

Glycan Rank Proportion

- In general, composition element rank proportions.

Class B

Patient	glyA > B	gly A > C
1	1	0
2	1	1
3	1	1
4	1	0

Class A

Patient	glyA > B	gly A > C
1	1	1
2	1	1
3	0	1
4	1	1

Patient	glyA > B	gly A > C
1	0	0
2	0	1
3	0	1

Class C

Patient	glyA > B	gly A > C
1	0	0
2	1	1
3	1	1
4	0	0

Class D

Calculate Class Ranks Proportions and use in prediction

GRP Prediction (Full Data)

- 57.3,
- 58.2,
- 58.6,
- 65.3%

	True Class	Predicted Class				
		ATB/DM	*ATB/ND*	*LTB/DM*	*LTB/ND*	*Neg/ND*
Whole Glycan Only	*ATB/DM*	14	1	0	0	3
	ATB/ND	4	5	1	3	6
	LTB/DM	1	0	4	0	2
	LTB/ND	0	3	3	8	3
	Neg/ND	1	0	0	0	12
Fab Glycan Only	*ATB/DM*	12	3	0	0	1
	ATB/ND	8	7	0	3	6
	LTB/DM	0	0	5	0	2
	LTB/ND	1	0	4	7	4
	Neg/ND	2	1	1	2	8
Fc Glycan Only	*ATB/DM*	12	1	1	0	2
	ATB/ND	6	6	2	5	4
	LTB/DM	0	0	4	1	1
	LTB/ND	0	0	5	11	1
	Neg/ND	2	1	0	2	1
All Glycan compositions	*ATB/DM*	14	1	0	0	1
	ATB/ND	6	10	0	4	3
	LTB/DM	0	0	5	1	1
	LTB/ND	1	0	1	12	2
	Neg/ND	2	0	1	2	8

Epigenetic Modeling (Methylation Profile Smoothing)

Part 3/3

Slides – [33-41]

Questions – Methylation Smoothing

3. What are some capabilities of modeling **epigenetic data**?
    1. How frequently do point estimates produce inaccurate results? (Simulation Study)
    2. Do Read/Reference Length Play role in coverage variance? (Proof-Result)

Hypotheses & Procedures (3)

3.1) How frequently do point estimates produce inaccurate results?

- Hypothesis: The relationship between coverage and point estimate errors should be decreasing, with increased coverage point estimates will be incorrect in order less often.

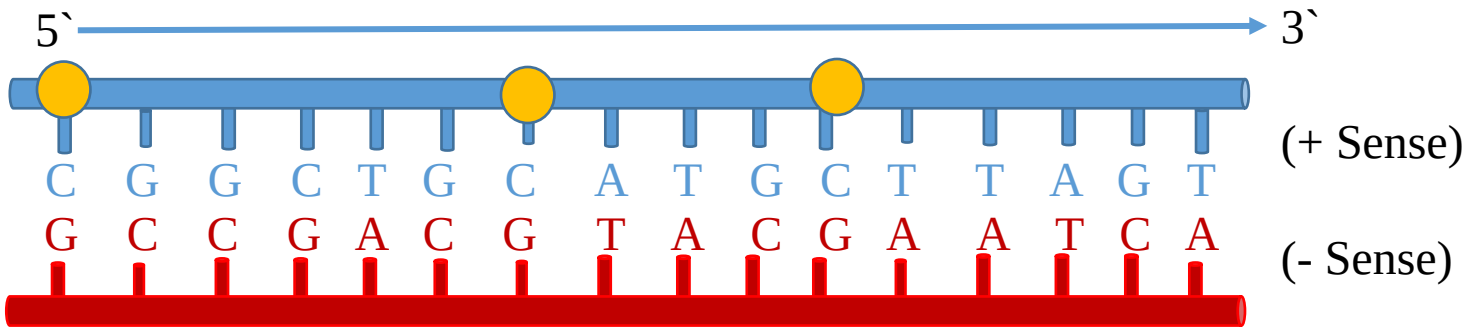
3.2) Do Read/Reference Length Play role in coverage variance?

- Hypothesis: There is a direct relationship between coverage variance and Read/Reference Length

Background - Genetic Sequence Data

- Usually stored in FASTA Files

Methylation is a epigenetic Cue which can give much more information!



Read Sequence Data

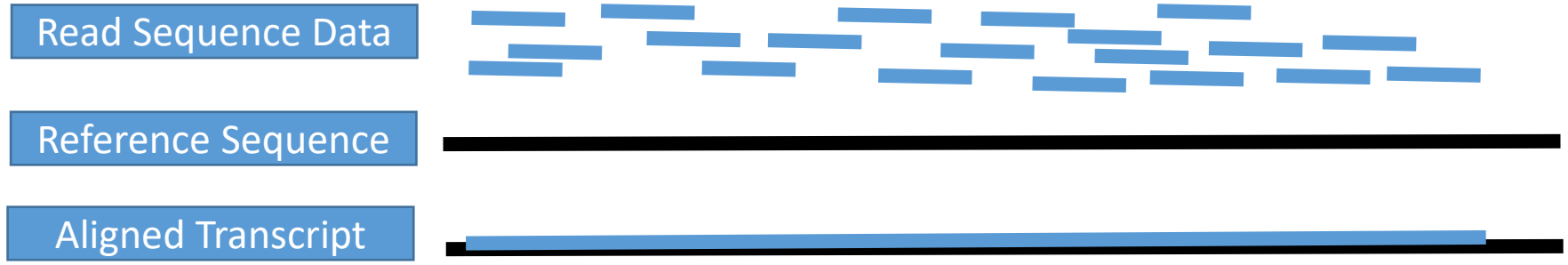
```
> Read 1 Information
ACGTACGAGCTGGTCCTAAGGTGTGCTCAGTATCCCTGGTATATGGT
> Read 2 Information
ACGTACGAGCTGGTCCACCGGTGTGCTCAGTATCCCTCCAGGATGGT
> Read 3 Information
ACGTACGAGCTGGTCCTAAGGTGTGCTCAGTATCGCTGGTATATGGT
```

Reference Sequence

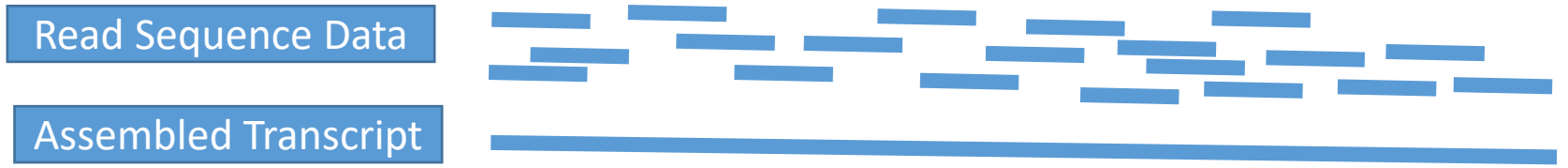
```
> Reference Information
ACGTAGTGTCTCTATATACTCTCTCTCCGGGAGAGTATGA
TCTCTGGTCATGATATTAACTGTGCTATATACGGTATAAG
TATGCTACGTACGAGCTGGTCCTAAGGTGTGCTCAGTATC
CCTGGTATATGGTTATATCGTGTGGTCCCAAACATCTCGC
GCGCGCGCGCGCGCGTCATATTAATATACGAGTCATGTCA
```

Background - Genetic Sequence Data

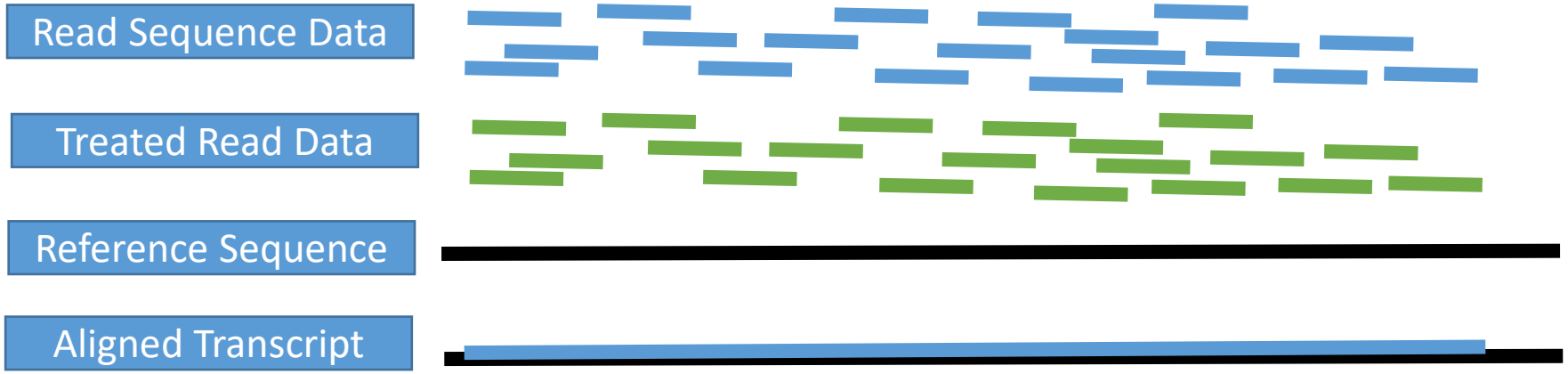
Alignment: Form transcript by mapping read sequences to reference.



Assembly: Form a transcript by matching most likely overlapping reads.



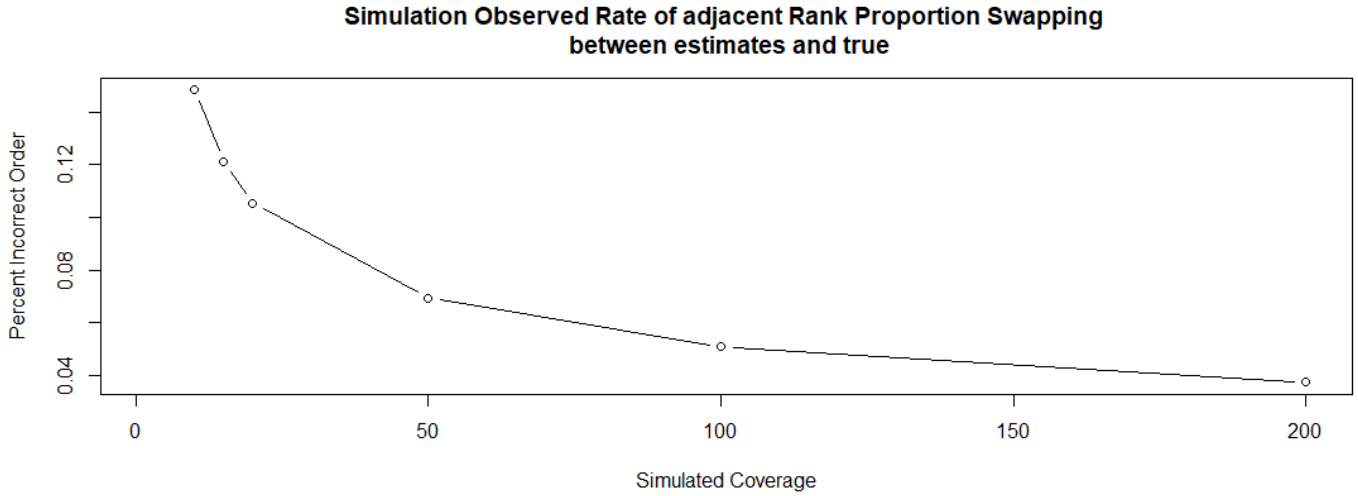
Background Epigenetic Signals

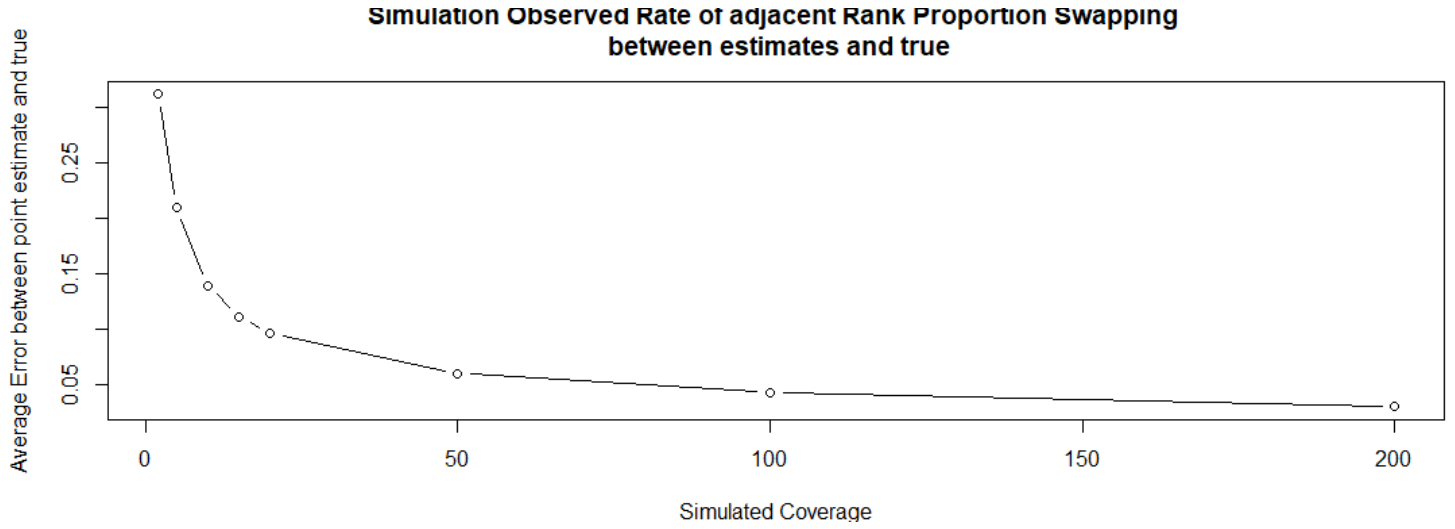


- Reads may be treated with Bisulfite first, so unmethylated cytosines are converted, then methylation is detected at mismatched locations post alignment.

Methylation Simulation Results

- Settings:
  - Number of Spots for Methylation: 1000
  - Length of genetic sequence of interest: 100,000
  - Size of Reads (100)
- Generation:
  - Methylation Ratio (True – Simulated) ~Beta(0.5)





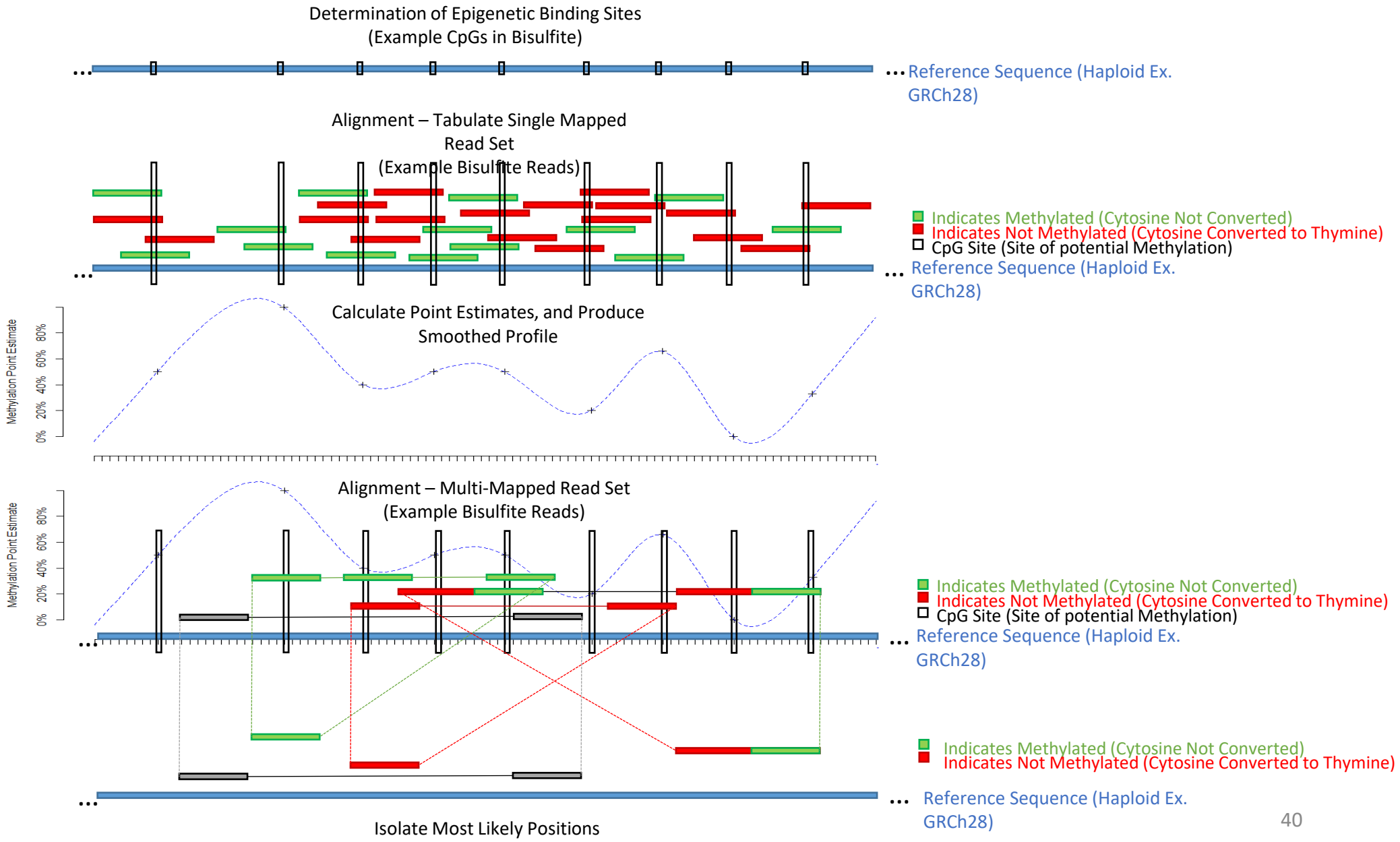
Coverage Variance Results

- Minimizer of coverage variance occurs when read length in reference to reference is:

Solving $\delta_{v_j}(s_r^*) = 0$ provides an optimum of

$$s_r^* = \frac{1}{10}\left(\sqrt{5s_f^2 - 10s_f + 9} + 5s_f + 3\right)$$

Methylation Demapping



Concluding Remarks

- Smoothing helps to determine more specific and potentially accurate methylation density estimates.
- The reference coverage is related to read length.
- Can be used to demap multi-mapped reads.

Acknowledgements

- My Advisor: Monnie McGee.
- My Supervisor/Mentor: Daehwan Kim.
- The other members of my PhD Committee: Dr. Heitjan and Dr. Sundararajan.
- My family/friends/supporters – Thank you!
- The Lenette Lu Lab

References (1)

- [1]Aitchison, J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological) 44, 2 (1982), 139–160.
- [2]Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Molecular Biology of the Cell. Garland Science, a member of the Taylor & Francis group, 29 West 35th Street, New York, NY 10001-2299, 2002.
- [3]Allaire, J., and Chollet, F. keras: R Interface to 'Keras', 2021. R package version 2.4.0.
- [4]Anastassiou, D. Genomic signal processing. IEEE Signal Processing Magazine 18, 4 (2001), 8–20.
- [5]Anderson, T. W. An introduction to multivariate statistical analysis. 1962.
- [6]Armăşelu, A. New spectral applications of the Fourier transforms in medicine, biological and biomedical fields. In Fourier Transforms High tech Application and Current Trends. In tech Open, 2017, pp. 235–252.
- [7]Braden, B. C., and Poljak, R. J. Structural features of the reactions between antibodies and protein antigens. FASEBJ9, 1 (Jan 1995), 9–16.
- [8]Brereton, R. G., and Lloyd, G. R. Partial least squares discriminant analysis: taking the magic away. Journal of Chemometrics 28, 4 (2014), 213–225.
- [9]Brieman, L. Random forests. Machine Learning 45 (2001), 5–32.
- [10]Brieman, L., Friedman, J., Stone, C. J., and Olshen, R. Classification and Regression Trees, first edition ed. Wadsworth, 1984.
- [11]Brillinger, D. R. Time Series Data Analysis and Theory. Society for Industrial and Applied Mathematics, 2001.
- [12]Centers for Disease Control, and Prevention. Science brief: Emerging sars-cov-2 variants. Website, 2021.
- [13]Cortes, C., and Vapnik, V. Support vector networks. Machine Learning 20, 3 (1995), 273–297.129
- [14]Corum, J., and Zimmer, C. Bad news wrapped in protein: Inside the coronavirus genome. The New York Times (Apr 2020).
- [15]Dunn, K. Process improvement using data, 2010. Available through download at https://learnche.org/pid/PID.pdf (2020), 430–0461
- [16]Elbe, S., and Buckland-Merrett, G. Data, disease and diplomacy: Gisaid's innovative contribution to global health. Global Challenges (2017), 33–46.
- [17]Fuentes, A. R., Ginori, J. V. L., and Ãbalo, R. G. Detection of coding regions in large DNA sequences using the short time Fourier transform with reduced computational load, 2006.
- [18]Geladi, P. Notes on the history and nature of partial least squares (PLS) modelling. Journal of Chemometrics 2, 4 (1988), 231–246.
- [19]Goodman, L. A. On simultaneous confidence intervals for multinomial proportions.Technometrics7, 2 (1965), 247–254.

References (2)

- [20]Grenfell, B. T., Bjorndstad, O. N., and Kappey, J.Travelling wavesand spatial hierarchies in measles epidemics.Nature414, 6865 (2001), 716.
- [21]Hansen, K. D., Langmead, B., and Irizarry, R. A.Bsmooth: from wholegenome bisulfite sequencing reads to differentially methylated regions.GenomeBiology13, R83 (2012).
- [22]Hastie, T., Tibshirani, R., and Friedman, J.Theelementsofstatisticallearning:datamining,inferenceandprediction. Springer New York, 2008.
- [23]Heideman, M., Johnson, D., and Burrus, C.Gauss and the history of thefast Fourier transform.IEEEASSPMagazine1, 4 (1984), 14–21.
- [24]Hinton, G. E., and Roweis, S.Stochastic neighbor embedding.Advancesinneuralinformationprocessingsystems15 (2002), 857–864.
- [25]Hirabayashi, J.GlycanProfiling. Springer Japan, Tokyo, 2008, pp. 56–59.
- [26]Hoang, T., Yin, C., Zheng, H., Yu, C., Lucy He, R., and Yau, S. S.-T.A new method to cluster dna sequences using fourier power spectrum.Journaloftheoreticalbiology372 (May 2015), 135–145. 25747773[pmid].
- [27]Hopfield, J. J.Neural networks and physical systems with emergent collectivecomputational abilities.ProcedingsoftheNationalAcademyofSciencesUSA79, 8 (1982), 2554–2558.
- [28]Hron, K., Templ, M., and Filzmoser, P.Imputation of missing values forcompositional data using classical and robust methods.ComputationalStatistics&DataAnalysis54, 12 (2010), 3095–3107.130
- [29]Johnson, K. A., and Goody, R. S.The original michaelis constant: Trans-lation of the 1913 michaelis–menten paper.Biochemistry50, 39 (Oct 2011),8264–8269.
- [30]Jukes, T. H., Cantor, C. R., et al.Evolution of protein molecules.Mammalianproteinmetabolism3 (1969), 21–132.
- [31]Katoh, K., Misawa, K., ichi Kuma, K., and Miyata, T.MaFFT: a novelmethod for rapid multiple sequence alignment based on fast Fourier transform.Nucleicacidsresearch30, 14 (Jul 2002), 3059–3066. 12136088[pmid].
- [32]Kawata, T.On the fourier series of a stationary stochastic process.Z.WahrscheinlichkeitstheorieverwGebiete6 (1966), 224–245.
- [33]Kim, D., Langmead, B., and Salzberg, S. L.Hisat: a fast spliced alignerwith low memory requirements.Naturemethods12, 4 (2015), 357–360.
- [34]Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L.Graph-based genome alignment and genotyping with hisat2 and hisat-genotype.NatureBiotechnology37, 8 (Aug 2019), 907–915.
- [35]Kimura, M.A simple method for estimating evolutionary rate of base substitu-tions through comparative studies of nucleotide sequences.JournalofMolecularEvolution16 (1980), 111–120.

References (3)

- [36]Langmead, B., and Salzberg, S. L.Fast gapped-read alignment with bowtie2.Naturemethods9, 4 (2012), 357.
- [37]Li, H., and Durbin, R.Fast and accurate short read alignment with Bur-rows–Wheeler transform.Bioinformatics25, 14 (05 2009), 1754–1760.
- [38]Lipman, D., and Pearson, W.Rapid and sensitive protein similaritysearches.Science227, 4693 (1985), 1435–1441.
- [39]Lu, L. L., Das, J., Grace, P. S., Fortune, S. M., Restrepo, B. I., andAlter, G.Antibody Fc Glycosylation Discriminates Between Latent and ActiveTuberculosis.TheJournalofInfectiousDiseases222, 12 (02 2020), 2093–2102.
- [40]Lu, L. L., Das, J., Grace, P. S., Fortune, S. M., Restrepo, B. I., andAlter, G.Antibody Fc Glycosylation Discriminates Between Latent and ActiveTuberculosis.TheJournalofInfectiousDiseases222, 12 (02 2020), 2093–2102.
- [41]MATLAB.9.8.0.1417392(R2020a)Update4(R2020a). The MathWorks Inc.,Natick, Massachusetts, 2020.
- [42]Metzker, M. L.Sequencing technologies — the next generation.NatureReviewsGenetics11, 1 (Jan 2010), 31–46.131
- [43]Michaelis, L., and Menten, M. L.Die kinetik der invertinwirkung.BiochemischeZeitschrift49 (1913), 333 – 369.
- [44]Mittermayr, S., Bones, J., and Guttman, A.Unraveling the glyco-puzzle: Glycan structure identification by capillary electrophoresis.AnalyticalChemistry85, 9 (May 2013), 4228–4238.
- [45]Nunes, C. A., Freitas, M. P., Pinheiro, A. C. M., and Bastos, S. C.Chemoface: a novel free user-friendly interface for chemometrics.JournaloftheBrazilianChemicalSociety23 (2012), 2003–2010.
- [46]Oppenheim, A. V., Willisky, A. S., and Nawab, S. H.Signals&Systems2ndEdition. Pearson Education, Upper Saddle River, New Jersey 07458, 1997.
- [47]Paradis, E., and Schliep, K.ape 5.0: an environment for modern phyloge-netics and evolutionary analyses in R.Bioinformatics35 (2019), 526–528.
- [48]Pei, S., Dong, R., He, R. L., and Yau, S. S.-T.Large-scale genome com-parison based on cumulative fourier power and phase spectra: Central momentand covariance vector.ComputationalandStructuralBiotechnologyJournal17(2019), 982–994.
- [49]Polat, K., and Güneş, S.Classification of epileptiform eeg using a hybridsystem based on decision tree classifier and fast Fourier transform.AppliedMathematicsandComputation187, 2 (2007), 1017 – 1026.
- [50]Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., andNarasimhan, G.So you think you can pls-da?BMCBioinformatics21, 1(Dec 2020), 2.
- [51]Saitou, N., and Nei, M.The neighbor-joining method: a new method forreconstructing phylogenetic trees.MolecularBiologyandEvolution4, 4 (071987), 406–425.
- [52]Salcedo, A. C., and Baldasano Recio, J.Fourier analysis of meteorolog-ical data to obtain a typical annual time function.SolarEnergy32, 4 (1984),479 – 488.

References (4)

- [53]Sanchez, R., Yang, X., Maher, T., and Mackenzie, S. A.Discrim-ination of dna methylation signal from background variation for clinical diag-nostics.Internationaljournalofmolecularsciences20, 21 (Oct 2019), 5343.31717838[pmid].
- [54]Shu, J.-J., and Yong, K. Y.Fourier-based classification of protein sec-ondary structures.BiochemicalandBiophysicalResearchCommunications485,4 (2017), 731–735.132
- [55]Singleton, R.An algorithm for computing the mixed radix fast fourier trans-form.IEEETransactionsonAudioandElectroacoustics17, 2 (1969), 93–103.
- [56]Slatko, B. E., Gardner, A. F., and Ausubel, F. M.Overview of next-generation sequencing technologies.Currentprotocolsinmolecularbiology122,1 (Apr 2018), e59–e59. 29851291[pmid].
- [57]Song, Y., Cong, Y., Wang, B., and Zhang, N.Applications of fouriertransform infrared spectroscopy to pharmaceutical preparations.ExpertOpiniononDrugDelivery17, 4 (2020), 551–571. PMID: 32116058.
- [58]The MathWorks, I.ParallelComputingToolbox. Natick, Massachusetts,United State, 2019.
- [59]The MathWorks, I.BioinformaticsToolbox. Natick, Massachusetts, UnitedState, 2020.
- [60]The MathWorks, I.StatisticsandMachineLearningMathToolbox. Natick,Massachusetts, United State, 2020.
- [61]Thornton, M.The invariance of spectral-kolmogorov-type statistics for es-timating genomic similarity. In2019IEEE49thInternationalSymposiumonMultiple-ValuedLogic(ISMVL) (2019), IEEE, pp. 73–78.
- [62]Thornton, M.Genomic DFT in R, 2021.
- [63]Tibshirani, R.Regression shrinkage and selection via the lasso.JournaloftheRoyalStatisticalSociety:SeriesB(Methodological)58, 1 (1996), 267–288.
- [64]TIBSHIRANI, R.The lasso method for variable selection in the cox model.StatisticsinMedicine16, 4 (1997), 385–395.
- [65]Tonegawa, S.Somatic generation of antibody diversity.Nature302, 5909 (Apr1983), 575–581.
- [66]van der Maaten, L., and Hinton, G. Visualizing data using t-sne.Journalofmachinelearningresearch9, Nov (2008), 2579–2605.
- [67]Voss, R. F.Evolution of long-range fractal correlations and 1/f noise in dnabase sequences.Physicalreviewletters68, 25 (1992), 3805.
- [68]Waage, P., and Guldberg, C.Studier over affiniteten.ForhandlingeriVidenskabs-selskabetiChristiania1 (1864), 35–45.
- [69]Wang, H., Meng, J., and Tenenhaus, M. Regression modelling analysis on compositional data. InHandbookofPartialLeastSquares. Springer, 2010,pp. 381–406.133

References (5)

- [70]Wasserman, L.AllofNonparametricStatistics. Springer Texts in Statistics.Springer New York Inc., 2006.
- [71]Wold, H.Estimation of principal components and related models by iterativeleast squares.Multivariateanalysis (1966), 391–420.
- [72]Yin, C.Phylogenetic analysis of DNA sequences or genomes by Fourier trans-form, August 2020.
- [73]Yin, C., Chen, Y., and Yau, S.A measure of DNA sequence similarityby Fourier transform with applications on hierarchical clustering.Journaloftheoreticalbiology359 (06 2014), 18–28.
- [74]Yin, C., and Yau, S. S.-T.An improved model for whole genome phylogeneticanalysis by Fourier transform.JournalofTheoreticalBiology382 (2015), 99 –110.
- [75]Zhang, Y., Park, C., Bennett, C., Thornton, M., and Kim, D.Hisat-3n: a rapid and accurate three-nucleotide sequence aligner.GenomeResearch(2020).
- [76]Zhou, Y., Zhou, L.-Q., Yu, Z.-G., and Anh, V.Distinguish coding andnoncoding sequences in a complete genome using fourier transform. InThirdInternationalConferenceonNaturalComputation(ICNC2007) (2007), vol. 2,IEEE, pp. 295–2

Thankyou for Your Attention

Questions? Suggestions? Critiques?

Backup Slides

Coverage Variance Related To Read Length

$$X_i \overset{\text{iid}}{\sim} \text{DUNIF}[0, s_f - (s_r - 1)]$$

$$\iff \text{P}(X_i = x) = \begin{cases} \frac{1}{s_f - (s_r - 1)} & \{x \in \mathbb{Z} | 1 \leq x \leq s_f - (s_r - 1)\} \\ 0 & \text{otherwise} \end{cases}$$

$$\forall i = 1, 2, \ldots, n_r$$

$$C_{ij} = \begin{cases} 1 & x_i \leq l_j \leq x_i + (s_r - 1) \\ 0 & \text{otherwise} \end{cases}$$

Coverage Variance Related To Read Length

$$
\begin{aligned}
\mathrm{P}(C_{ij}=1) &= \mathrm{P}(X_i \le l_j \cap l_j \le X_i + (s_r - 1)) \\
&= \mathrm{P}(l_j - (s_r - 1) \le X_i \le l_j) \\
&= \sum_{k=l_j-(s_r-1)}^{l_j} \mathrm{P}(X_i = k) = \frac{s_r - 1}{s_f - (s_r - 1)} \quad \forall (i,j) \\
&\implies C_{ij} \overset{\text{iid}}{\sim} \mathrm{Bern}\left(p = \frac{s_r - 1}{s_f - (s_r - 1)}\right) \quad \text{In } i \text{ only}
\end{aligned}
$$

Coverage Variance Related To Read Length

$$H_j = \sum_{i=1}^{n_r} C_{ij} \implies H_j \sim \text{Bin}\left(n = n_r, p = \frac{s_r - 1}{s_f - (s_r - 1)}\right)$$

It can be shown that,

$$\text{Var}(H_j) = np(1-p) = n_r \cdot \frac{s_r - 1}{s_f - (s_r - 1)} \cdot \frac{s_f - 2s_r + 2}{s_f - (s_r - 1)}$$

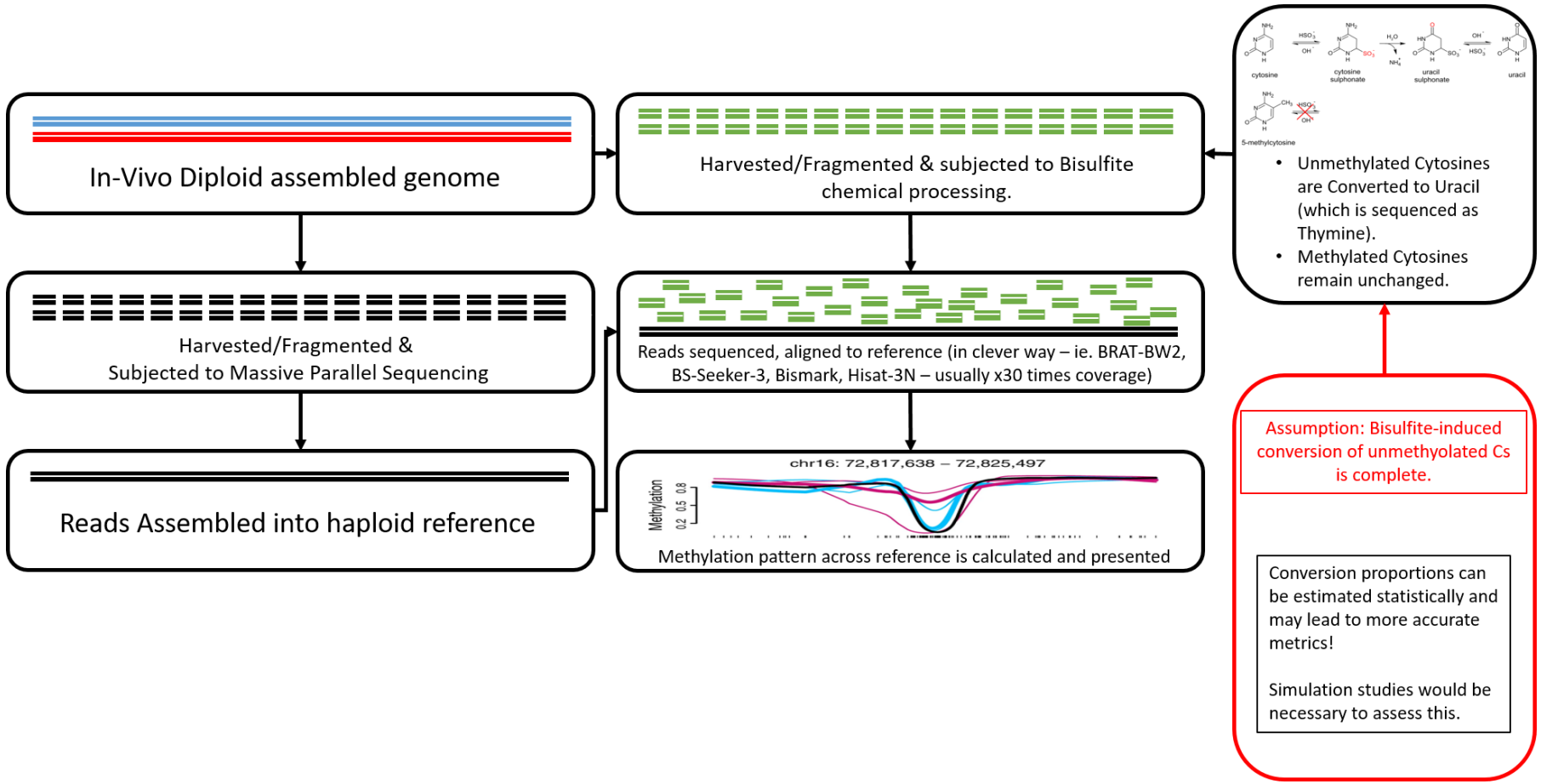
$$\implies \frac{\partial \text{Var}(H_j)}{\partial s_r} = -n_r \cdot s_f \cdot \frac{3s_r - s_f - 3}{(-s_r + s_f + 1)^3}$$

Or more correctly, considering the first order differencing (as s_r is actually a discrete variable as noted above), and letting $v_j(s_r, s_f, n_r) = \text{Var}(H_j)$:

$$\delta_{v_j}(s_r) = n_r \cdot \left(\frac{s_f s_r - 2s_r^2}{s_f^2 - 2s_f s_r + s_r^2} - \frac{s_f s_r - 2s_r^2 + 2s_r - s_f + 2s_r - 2}{s_f^2 - 2s_f s_r + s_f + s_r^2 - 2s_r + 1} \right)$$

Solving $\delta_{v_j}(s_r^*) = 0$ provides an optimum of

$$s_r^* = \frac{1}{10}\left(\sqrt{5s_f^2 - 10s_f + 9} + 5s_f + 3\right)$$


In-Vivo Diploid assembled genome
Harvested/Fragmented & subjected to Bisulfite chemical processing.
cytosine
cytosine sulphonate
uracil sulphonate
uracil
5-methylcytosine
• Unmethylated Cytosines are Converted to Uracil (which is sequenced as Thymine).
• Methylated Cytosines remain unchanged.
Harvested/Fragmented & Subjected to Massive Parallel Sequencing
Reads sequenced, aligned to reference (in clever way – ie. BRAT-BW2, BS-Seeker-3, Bismark, Hisat-3N – usually x30 times coverage)
Assumption: Bisulfite-induced conversion of unmethyolated Cs is complete.
Reads Assembled into haploid reference
chr16: 72,817,638 – 72,825,497
Methylation
0.2 0.5 0.8
Methylation pattern across reference is calculated and presented
Conversion proportions can be estimated statistically and may lead to more accurate metrics!
Simulation studies would be necessary to assess this.

Background – Encoding Genomic Signals

- Representing a 4-category value observed by position
  - numerically,
  - preserving repetitive elements for subsequent analysis.

  Let i denote the locus of the ith position in a string representing a genomic signal, and G_i the actual base observed at that position, two possible encodings may be represented

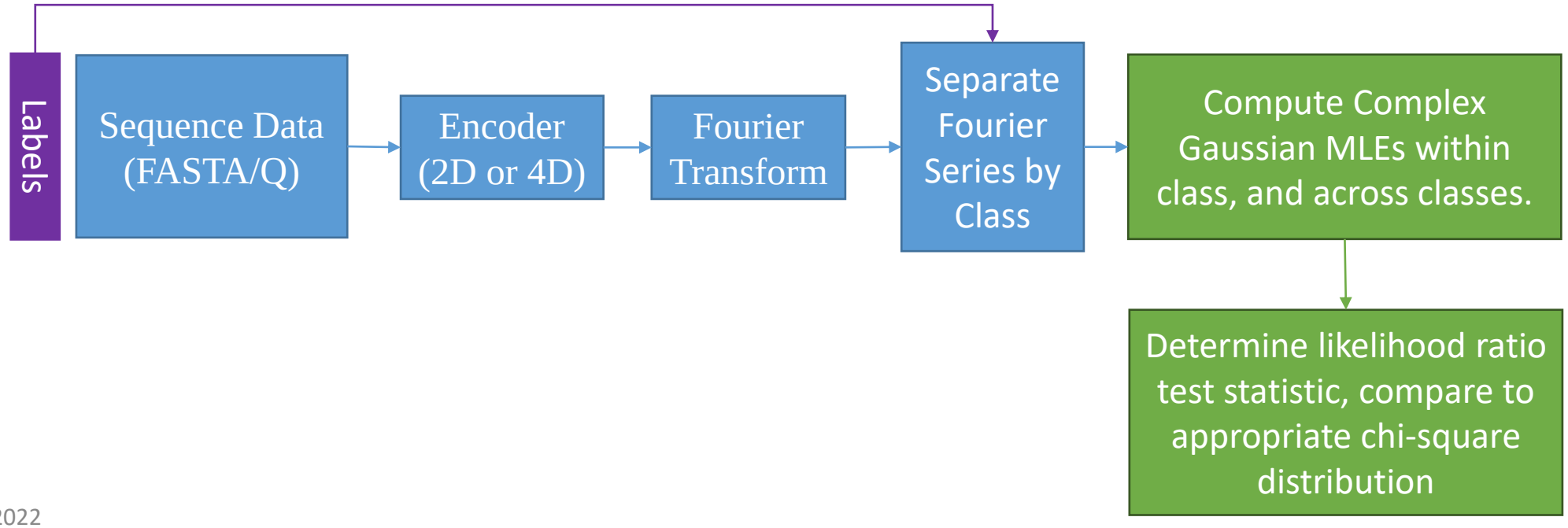
  $G_i \equiv' X'$ is 1 (True) if the ith nucleotide is X and 0 otherwise

$$b : \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \mapsto \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \right\} \quad S_i = \begin{pmatrix} G_i \equiv 'A' \\ G_i \equiv' C' \\ G_i \equiv' G' \\ G_i \equiv' T' \end{pmatrix} \quad \text{or} \quad S_i = \begin{pmatrix} G_i \equiv' C' \\ G_i \equiv' G' \end{pmatrix} - \begin{pmatrix} G_i \equiv' A' \\ G_i \equiv' T' \end{pmatrix}$$

Hypotheses & Procedures (1)

1.3) To provide some **statistically** valid approach to **comparing autocorrelation among ensembles**?

- Hypothesis: A suitable derivation for determining the likelihood ratio test distribution is provided and an algorithm for computing p-values is provided.



Testing Sequence Data By Fourier Coefficients

- **For a random variable *Z* such as a specific frequency transform coefficient of a signal (recall these are r-dimensional, usually 2/4)**

$$\boldsymbol{Z} \sim N_r^{\mathbb{C}}(\boldsymbol{\mu_X}, \boldsymbol{\Sigma_{XX}}) \iff \begin{pmatrix} \boldsymbol{Z_{\Re}} \\ \boldsymbol{Z_{\Im}} \end{pmatrix} \sim N_{2r}\left(\begin{pmatrix} \boldsymbol{\mu_{X_{\Re}}} \\ \boldsymbol{\mu_{X_{\Im}}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_{XX_{\Re}}} & \boldsymbol{-\Sigma_{XX_{\Im}}} \\ \boldsymbol{\Sigma_{XX_{\Im}}} & \boldsymbol{\Sigma_{XX_{\Re}}} \end{pmatrix} \right)$$

- **Where the Multivariate Normal Distribution is given by,**

$$\boldsymbol{Q} \sim N_{2r}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff dP(\boldsymbol{Q} \leq \boldsymbol{q}) = \left((2\pi)^{2r} |\boldsymbol{\Sigma}| \right)^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{q}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{q}-\boldsymbol{\mu})} d\boldsymbol{q}$$

Testing Sequence Data By Fourier Coefficients

- **It can be shown that the MLE for these parameters are given**

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}_{\Re}} & -\widehat{\boldsymbol{\Sigma}_{\Im}} \\ \widehat{\boldsymbol{\Sigma}_{\Im}} & \widehat{\boldsymbol{\Sigma}_{\Re}} \end{pmatrix} \qquad \hat{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\boldsymbol{\mu}_{\Re}} \\ \widehat{\boldsymbol{\mu}_{\Im}} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \widehat{\mu_{\Re 1}} & \widehat{\mu_{\Re 2}} & \dots & \widehat{\mu_{\Re r}} \end{pmatrix}^T \\ \begin{pmatrix} \widehat{\mu_{\Im 1}} & \widehat{\mu_{\Im 2}} & \dots & \widehat{\mu_{\Im r}} \end{pmatrix}^T \end{pmatrix} = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{z_{n\Re}} \\ \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{z_{n\Im}} \end{pmatrix}$$

$$\widehat{\boldsymbol{\Sigma}_{\Re}} = \begin{pmatrix} \widehat{\sigma^2_{\Re 1}} & \widehat{\rho_{\Re 1,2}} & \dots & \widehat{\rho_{\Re 1,r}} \\ \widehat{\rho_{\Re 2,1}} & \widehat{\sigma^2_{\Re 2}} & \dots & \widehat{\rho_{\Re 2,r}} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\rho_{\Re r,1}} & \widehat{\rho_{\Re r,2}} & \dots & \widehat{\sigma^2_{\Re r}} \end{pmatrix} \qquad \begin{aligned} \widehat{\sigma^2_{\Re l}} &= \frac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Re l} - \widehat{\mu_{\Re l}}\right)^2 \\ \widehat{\rho_{\Re l,m}} &= \frac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Re l} - \widehat{\mu_{\Re l}}\right)\left(z_{n\Re m} - \widehat{\mu_{\Re m}}\right) \end{aligned}$$

$$\widehat{\boldsymbol{\Sigma}_{\Im}} = \begin{pmatrix} \widehat{\sigma^2_{\Im 1}} & \widehat{\rho_{\Im 1,2}} & \dots & \widehat{\rho_{\Im 1,r}} \\ \widehat{\rho_{\Im 2,1}} & \widehat{\sigma^2_{\Im 2}} & \dots & \widehat{\rho_{\Im 2,r}} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\rho_{\Im r,1}} & \widehat{\rho_{\Im r,2}} & \dots & \widehat{\sigma^2_{\Im r}} \end{pmatrix} \qquad \begin{aligned} \widehat{\sigma^2_{\Im l}} &= \frac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Im l} - \widehat{\mu_{\Im l}}\right)^2 \\ \widehat{\rho_{\Im l,m}} &= \frac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Im l} - \widehat{\mu_{\Im l}}\right)\left(z_{n\Im m} - \widehat{\mu_{\Im m}}\right) \end{aligned}$$

Testing Sequence Data By Fourier Coefficients

- **Which Allows Calculation of the Test Statistic**

Log-likelihood among all classes

$$\log L\left(\boldsymbol{\mu}=\mathbf{0},\boldsymbol{\Sigma}=\widehat{\boldsymbol{\Sigma}}\ ;\ \ (\boldsymbol{z_1},\boldsymbol{z_2},\ldots,\boldsymbol{z_N})^T\right)=\ell(\widehat{\boldsymbol{\Sigma}})$$

$$=\frac{N}{4}\cdot\log\left((2\pi)^{2r}|\widehat{\boldsymbol{\Sigma}}|\right)\sum_{n=1}^{N}\begin{pmatrix}\boldsymbol{z_{n\Re}}\\ \boldsymbol{z_{n\Im}}\end{pmatrix}^T\widehat{\boldsymbol{\Sigma}}^{-1}\begin{pmatrix}\boldsymbol{z_{n\Re}}\\ \boldsymbol{z_{n\Im}}\end{pmatrix}$$

Log-likelihood within all classes

$$\ell(\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}},\ldots,\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{K}})=$$

$$\sum_{k=1}^{K}\left(\frac{N_k}{4}\cdot\log\left((2\pi)^{2r}|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}|\right)\sum_{n_k=1}^{N_k}\left(\begin{pmatrix}\boldsymbol{z_{n_k\Re(k)}}\\ \boldsymbol{z_{n_k\Im(k)}}\end{pmatrix}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}{}^{-1}\begin{pmatrix}\boldsymbol{z_{n_k\Re(k)}}\\ \boldsymbol{z_{n_k\Im(k)}}\end{pmatrix}\right)\right)$$

Likelihood Ratio Test Statistic

$$\Lambda=-2\cdot\left(\ell(\widehat{\boldsymbol{\Sigma}})-\ell(\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}},\ldots,\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{K}})\right)\implies\Lambda\sim\chi^2_{K-1}$$

- **Distributed according to the chi-square distribution with degrees of freedom given by the number of classes – 1.**

Conclusions (1.3)

- The Fourier coefficients provide a numerical summary that can be used in this testing framework to give a statistical measure of the likelihood of observing data as or more extreme than observed if the data all came from the same class as opposed to different classes.

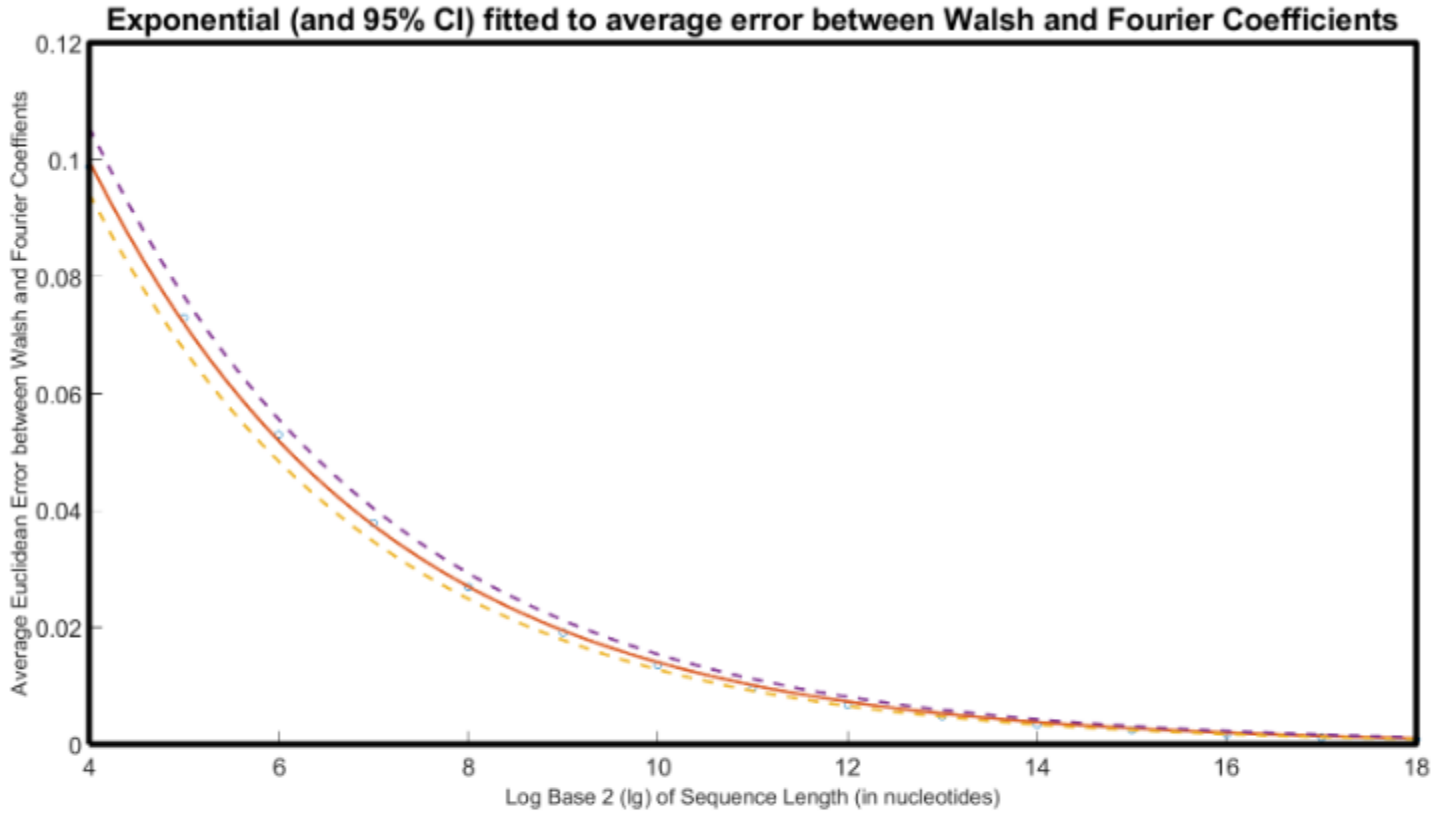
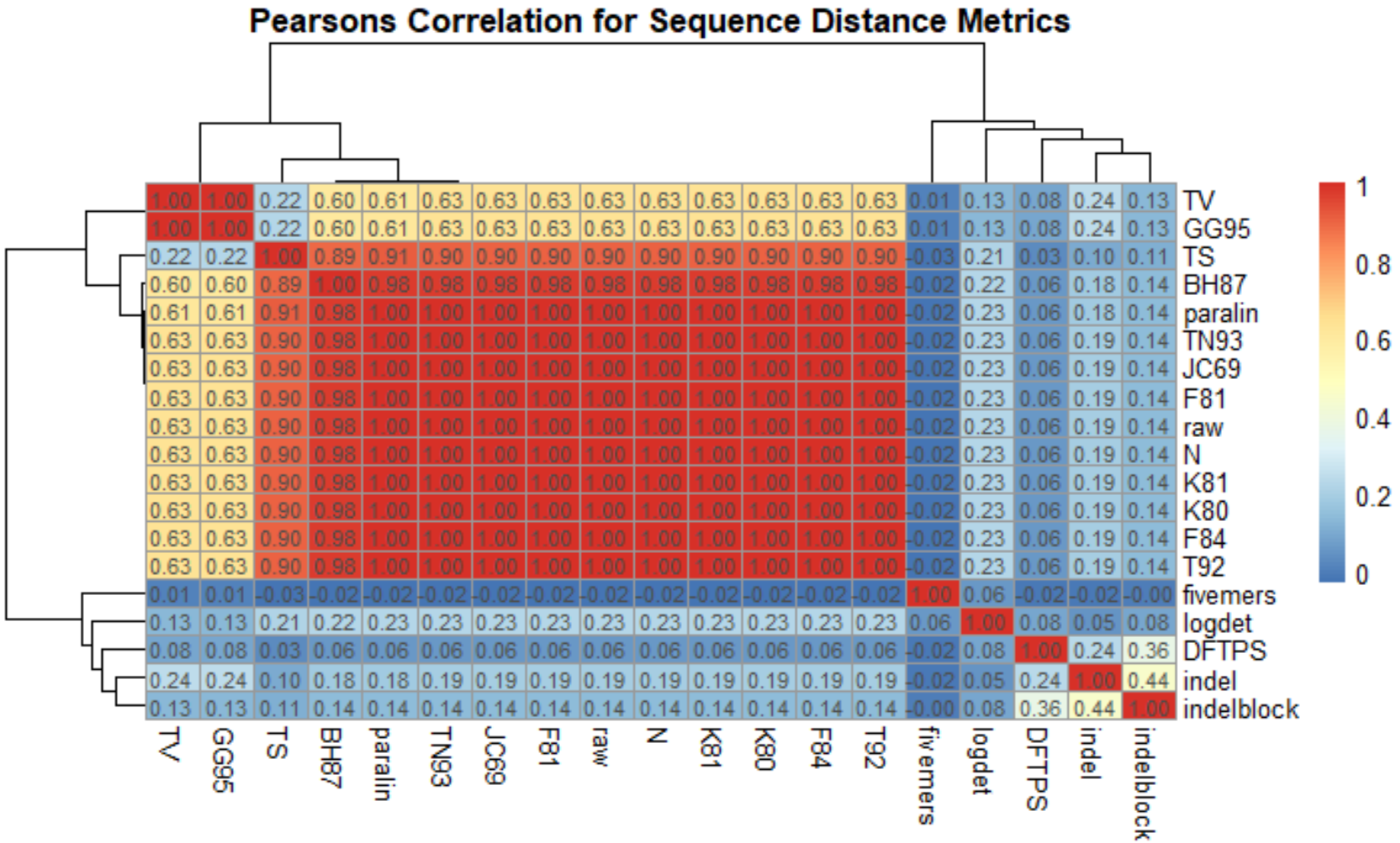


Figure 2.1. Exponentially Decreasing Error Between Walsh and Fourier Coefficients As Signal Length increases

Correlation of Pairwise Distances

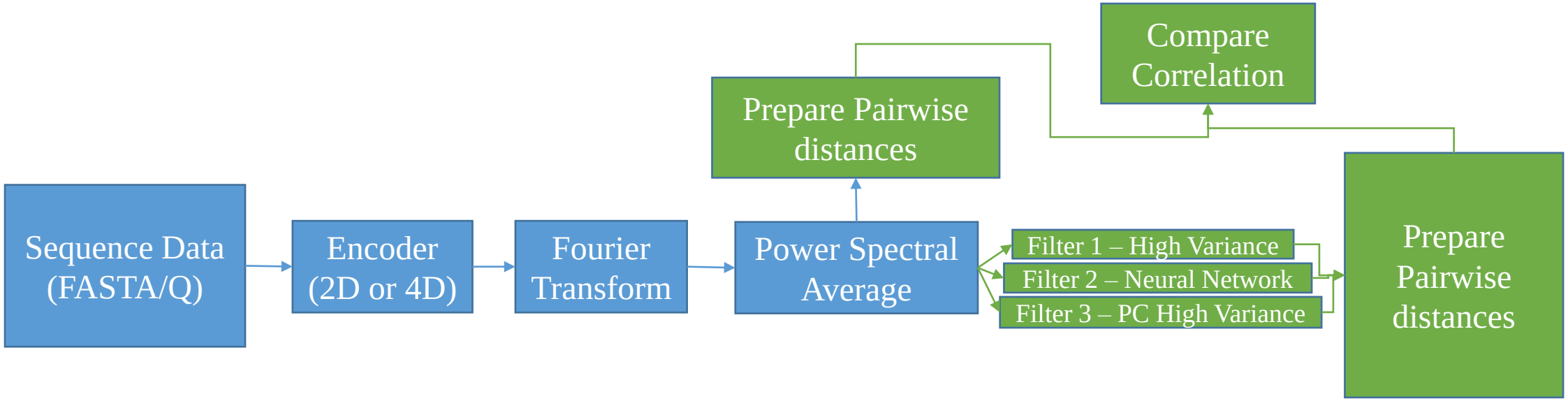


- There are two primary groupings of the clustering procedures (distance calculations) provided for the sequences.
- The first block includes the majority of the distance measures (from TV to T92) many of these distances are highly correlated with each other, these are all alignment requiring procedures.
- The Fourier Transform distance procedure (DFTPS) is contained by the second main block, which shows similarity between distances produced by the indel and indelblock distances, and almost no correlation with five-mer frequency distances.

Hypotheses & Procedures (1)

1.2) Can harmonic analysis be applied to genetic sequences to **achieve similar clustering** capabilities as other more intensive approaches?

- Sub-Hypothesis: Some **Filtered Subset** of the Power Spectra may provide a large part of the overall information contained in the full power spectra.



Filtering of Power Spectrum

- Subsets of coefficients can be used to achieve similar distances as the full set, three such techniques are also tested
  - Filter 1 – Select top Subset of PS coefficients in terms of variance across class
    - MVF – Minimum Variance Filter
  - Filter 2 – Select top Variance PS Coefficients and compute PCs
  - Filter 3 – Train NN to identify characteristic of interest and extract filters.

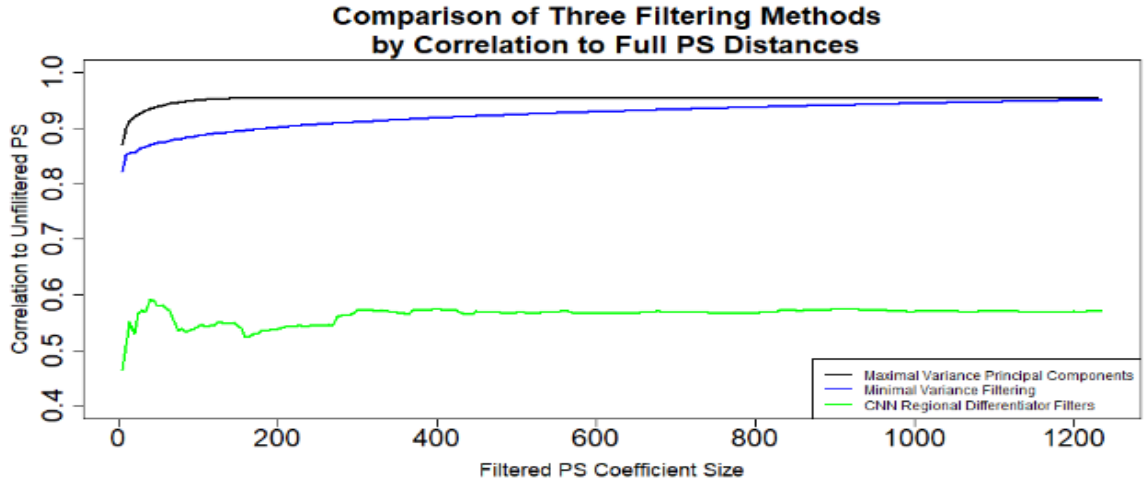


Figure 2.12. Filtering Methods Comparisons, by Correlation to Full PS Distances

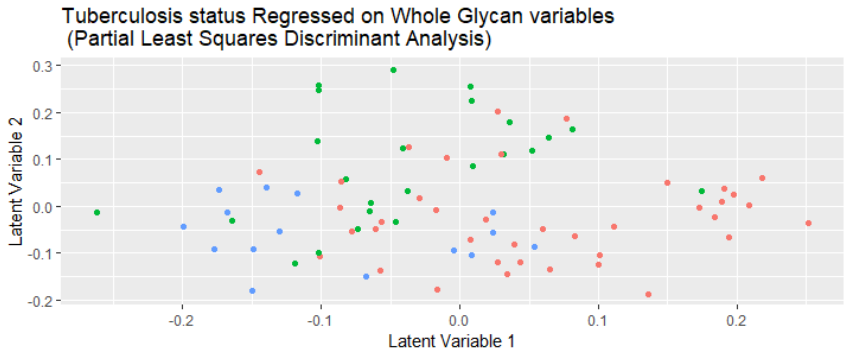
Table 2.5. Random Forest of 500 trees CV Accuracies for Region Classification from Filtered PS (%)

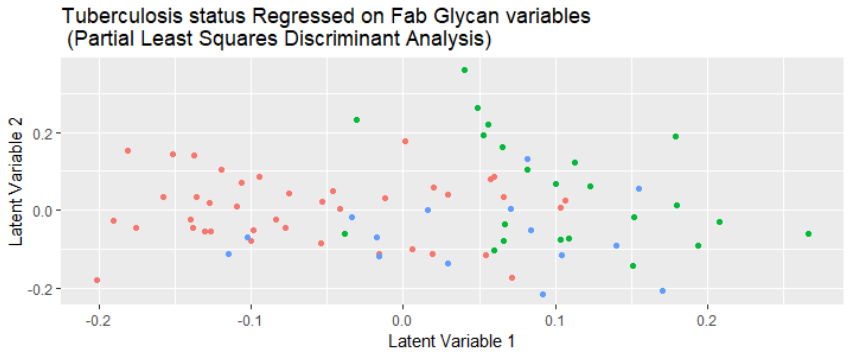
Filtered PS Size	MVF	AFL	MVPCF
50	24.833	**45.239**	21.759
100	24.917	**47.740**	38.874
250	31.003	45.528	**48.245**
500	28.706	48.244	**48.532**
1000	45.451	44.516	**48.532**

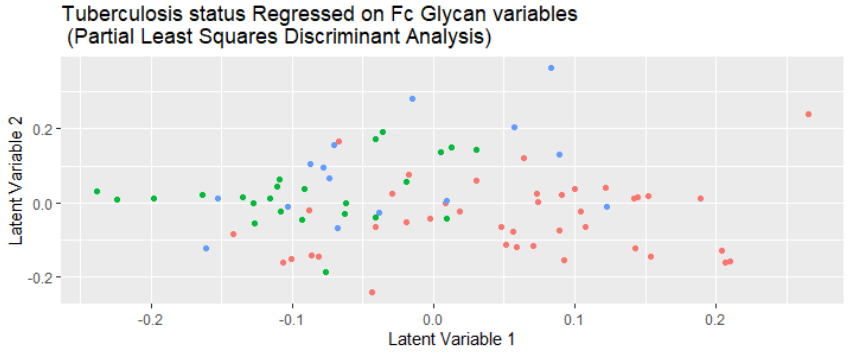
Partial Least Squares (1)

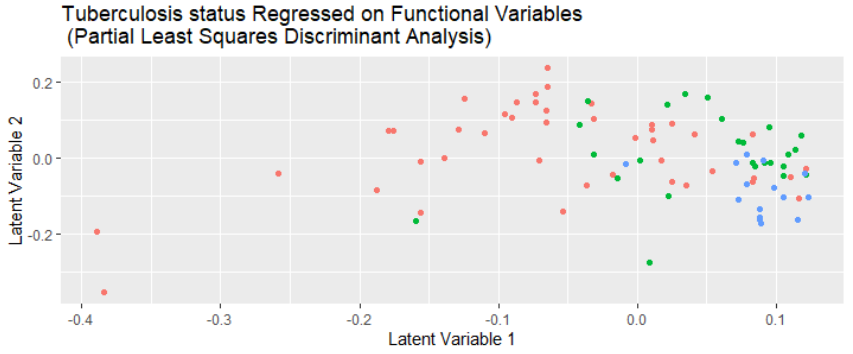
- PLS Models determine the principal components of the data (glycan assay data) and the outcomes, which when categorical are encoded in a class membership matrix.
- The scores for the data and outcomes are related using linear regression.
- maximize the variance within each block and the correlation between the data and the outcome.

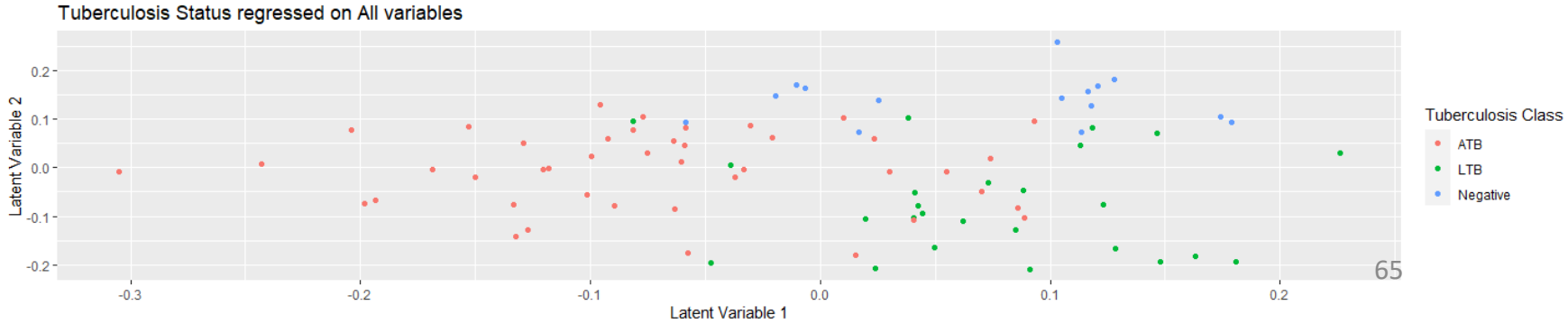
Partial Least Squares (2)







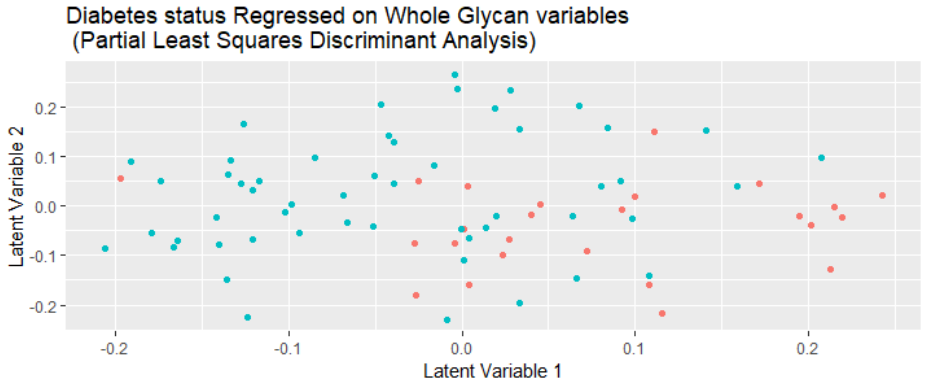


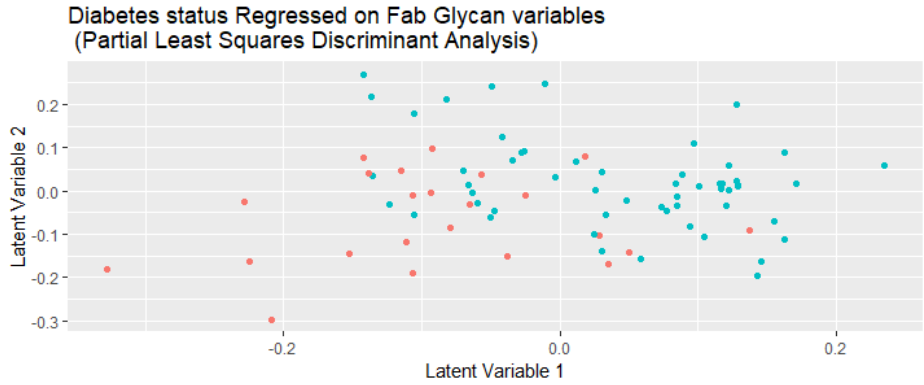


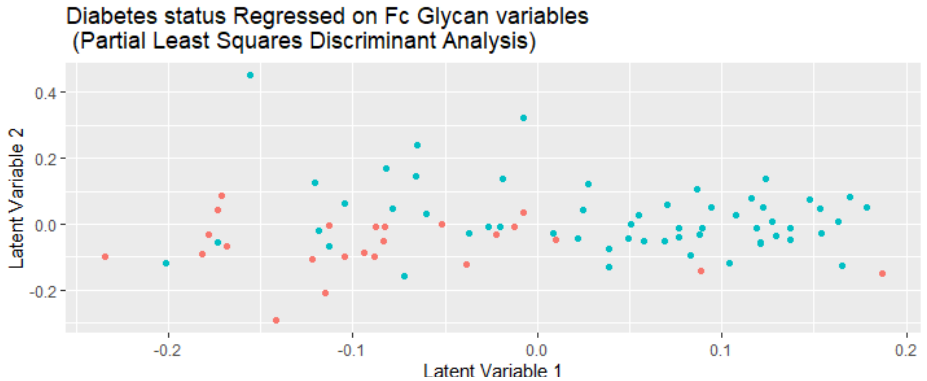
Partial Least Squares (3)

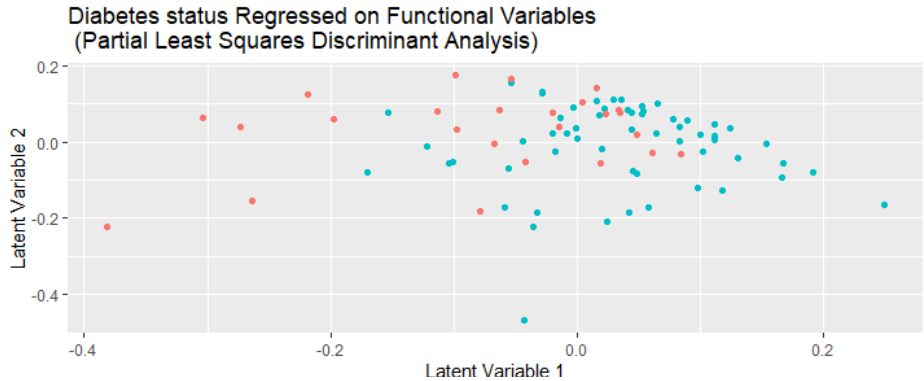
- PLS is also a predictive procedure, and once trained can be used to differentiate the class of an observation.
- It does so for a user specified number of components.

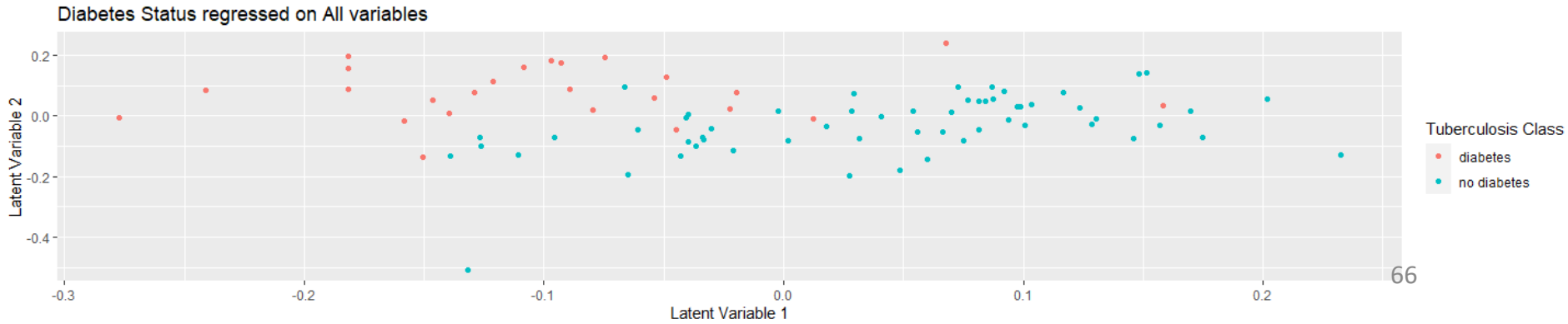
- Sometimes known as “Supervised Principal Components”.











Principal Components

Semi-Parametric Model Full Validation

	True Class	Predicted Class	
		DM	ND
84%, Nonparametric Score Only	DM	15	10
	ND	3	55
72% Parametric (Multinomial) Score Only	DM	23	2
	ND	21	37
79% Combined Semiparametric Score	DM	21	13
	ND	4	42

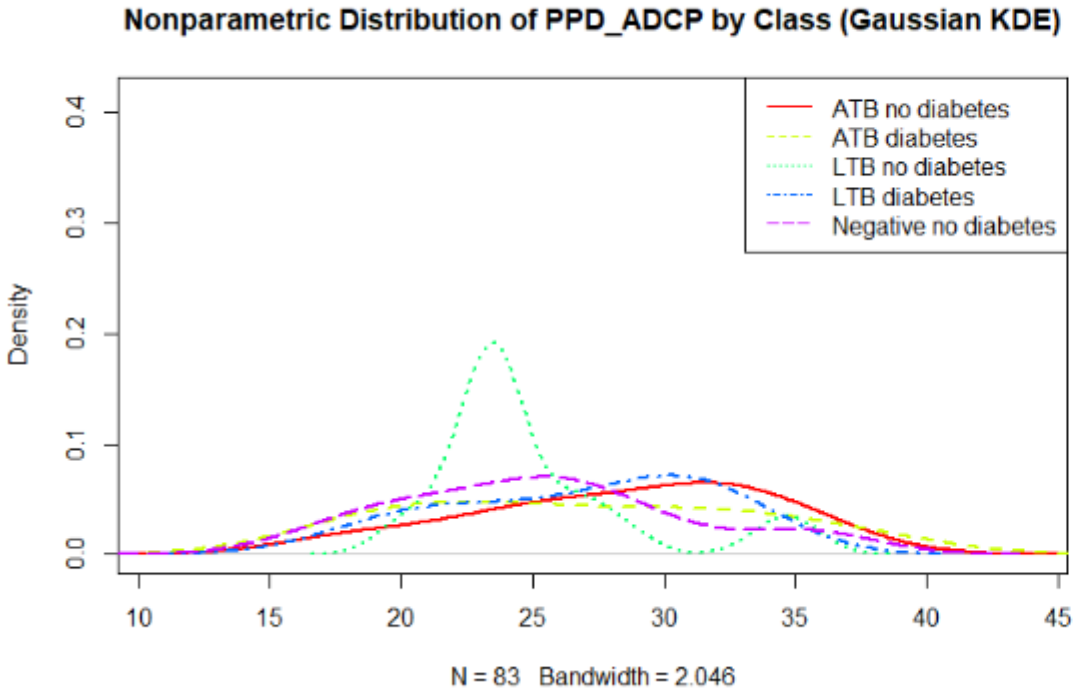
	True Class	Predicted Class		
		ATB	LTB	Neg
80% Nonparametric Score Only	ATB	33	9	1
	LTB	1	20	4
	Neg	0	1	14
Parametric (Multinomial) Score Only 65%	ATB	31	7	5
	LTB	7	16	2
	Neg	4	4	7
Combined Semiparametric Score 84%	ATB	35	2	1
	LTB	7	20	1
	Neg	0	2	12

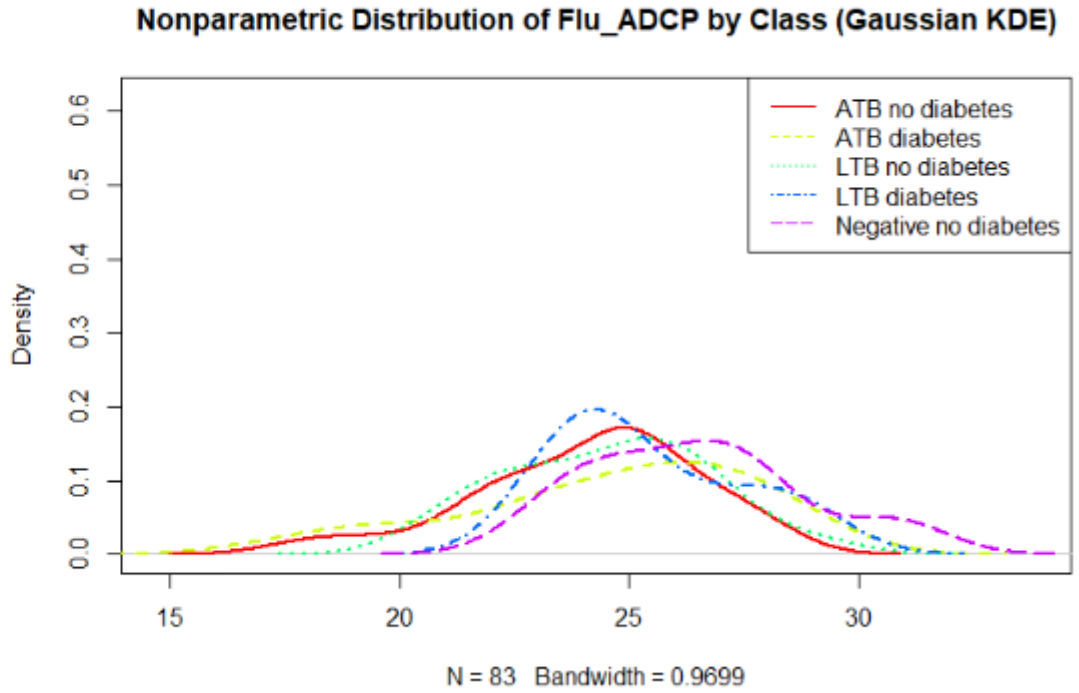
	True Class	Predicted Class				
		ATB/DM	*ATB/ND*	*LTB/DM*	*LTB/ND*	*Neg/ND*
Nonparametric Score Only 72.5 %,	*ATB/DM*	15	1	0	2	0
	ATB/ND	2	17	2	3	1
	LTB/DM	0	0	6	0	1
	LTB/ND	1	0	1	13	3
	Neg/ND	0	0	0	1	14
Parametric (Multinomial) Score Only 51.2 %	*ATB/DM*	16	1	1	0	0
	ATB/ND	10	7	1	5	2
	LTB/DM	3	0	1	3	0
	LTB/ND	0	5	1	11	1
	Neg/ND	1	2	2	2	8
Combined Semiparametric Score 78.3 %	*ATB/DM*	16	1	0	0	1
	ATB/ND	8	12	1	3	0
	LTB/DM	0	0	4	3	0
	LTB/ND	0	0	1	14	2
	Neg/ND	0	1	0	1	12

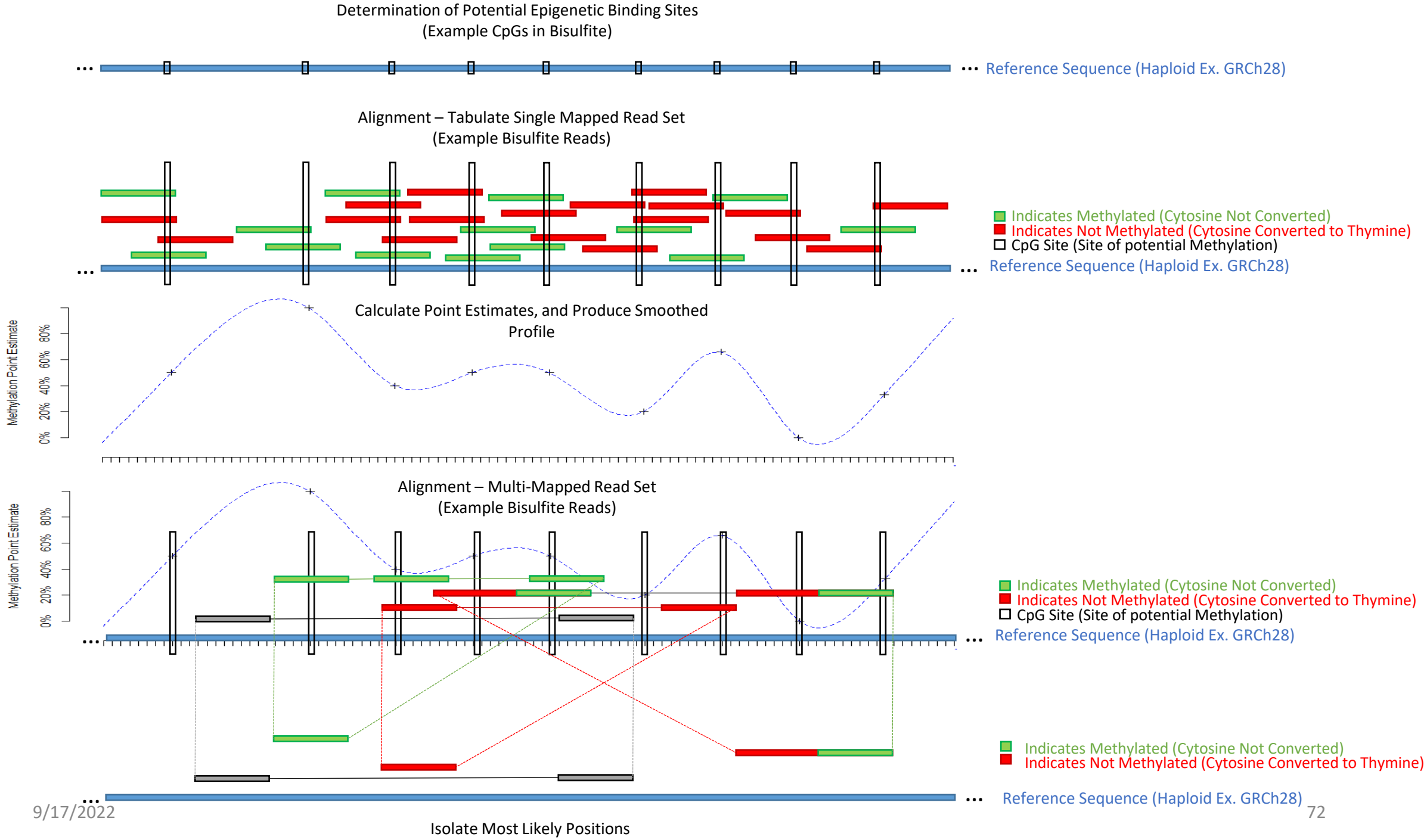
High Five Variance Only (GRP)

- whole - 50%, fab - 43%, fc - 48%, and combined 55%

	True Class	Predicted Class				
		ATB/ND	*ATB/DM*	*LTB/ND*	*LTB/DM*	*Neg/ND*
Whole Glycan Only	*ATB/ND*	1	7	9	1	1
	ATB/DM	0	14	1	1	2
	LTB/ND	0	2	10	2	3
	LTB/DM	0	4	1	1	1
	Neg/ND	0	0	1	1	11
Fab Glycan Only	*ATB/ND*	1	11	8	4	0
	ATB/DM	0	11	0	5	0
	LTB/ND	1	1	12	2	0
	LTB/DM	0	0	2	5	0
	Neg/ND	1	2	8	3	0
Fc Glycan Only	*ATB/ND*	9	6	6	1	1
	ATB/DM	3	12	1	1	0
	LTB/ND	4	0	12	1	0
	LTB/DM	2	1	3	1	0
	Neg/ND	3	5	2	0	3
All Glycan compositions	*ATB/ND*	2	6	6	1	3
	ATB/DM	1	12	0	2	1
	LTB/ND	2	0	12	2	0
	LTB/DM	0	1	2	4	0
	Neg/ND	0	1	2	1	8


Nonparametric Distribution of PPD_ADCP by Class (Gaussian KDE)
Density
0.0
0.1
0.2
0.3
0.4
10
15
20
25
30
35
40
45
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
N = 83 Bandwidth = 2.046


Nonparametric Distribution of Flu_ADCP by Class (Gaussian KDE)
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
15
20
25
30
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
N = 83 Bandwidth = 0.9699


Determination of Potential Epigenetic Binding Sites
(Example CpGs in Bisulfite)
Reference Sequence (Haploid Ex. GRCh28)
Alignment – Tabulate Single Mapped Read Set
(Example Bisulfite Reads)
Indicates Methylated (Cytosine Not Converted)
Indicates Not Methylated (Cytosine Converted to Thymine)
CpG Site (Site of potential Methylation)
Reference Sequence (Haploid Ex. GRCh28)
Calculate Point Estimates, and Produce Smoothed
Profile
Methylation Point Estimate
0%
20%
40%
60%
80%
Alignment – Multi-Mapped Read Set
(Example Bisulfite Reads)
Indicates Methylated (Cytosine Not Converted)
Indicates Not Methylated (Cytosine Converted to Thymine)
CpG Site (Site of potential Methylation)
Reference Sequence (Haploid Ex. GRCh28)
Indicates Methylated (Cytosine Not Converted)
Indicates Not Methylated (Cytosine Converted to Thymine)
Reference Sequence (Haploid Ex. GRCh28)
Isolate Most Likely Positions

Some Filter Charts