THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Predicting Population Diversity from Unassembled Reads

Nicolas Fishman, Keylie Gibson and Matthew Bendall

Abstract

Estimating the genetic diversity of viruses with high rates of nucleotide substitution, as HIV-1 and Influenza is difficult with data generated by Next Generation Sequencing (NGS). There have been several successful types of approaches that have developed over the time the technology has been available.

The gold standard for assessing the genetic diversity within a viral population is conducting Single Genome Amplification (SGA), where after a series of dilutions, it is presumed that a single viral strain has been isolated. From these multiple viral strains are isolated and NGS is performed. Genetic diversity is then calculated from this population to be a representative sample (Maldarelli et al, 2013 and Gibbs et al, 2007). SGA is time consuming, expensive, and often does not accurately represent the entire viral population within a single host, but only the dominate population.

A significant amount of attention has been paid to creating methods that attempt to take advantage of the longer reads produced by some NGS techniques in order to generatively regress a population of haplotypes. Assuming the distribution is correctly estimated, these haplotypes will be representative of the diversity of the original population.

Other methods, like Tanden (Zukurov et al, 2016), focus on taking advantage of the shorter, more accurate, higher coverage reads other NGS methods, allowing the avoidance of difficulties associated with haplotype reconstruction in favor of a frequency analysis of specific sites of the virus genome.

Here we propose a different approach, circumventing the assembly process entirely. We propose using the raw reads to do a k-mer analysis, and then using various summaries of the k-mer counts, regressing to population diversity. This approach is extremely flexible, able to adapt to any sequencing technology, very robust to noise and fairly accurate for the naïve type of approach it represents.

The key questions we examine here all revolve around (a) whether this represents a valid and useful technique for estimating virus population diversity, (b) which precise regressors are the best predictors of population diversity, and (c) what are the failure cases of this approach, and how to address them.

The Problem with Assembly

What is Assembly

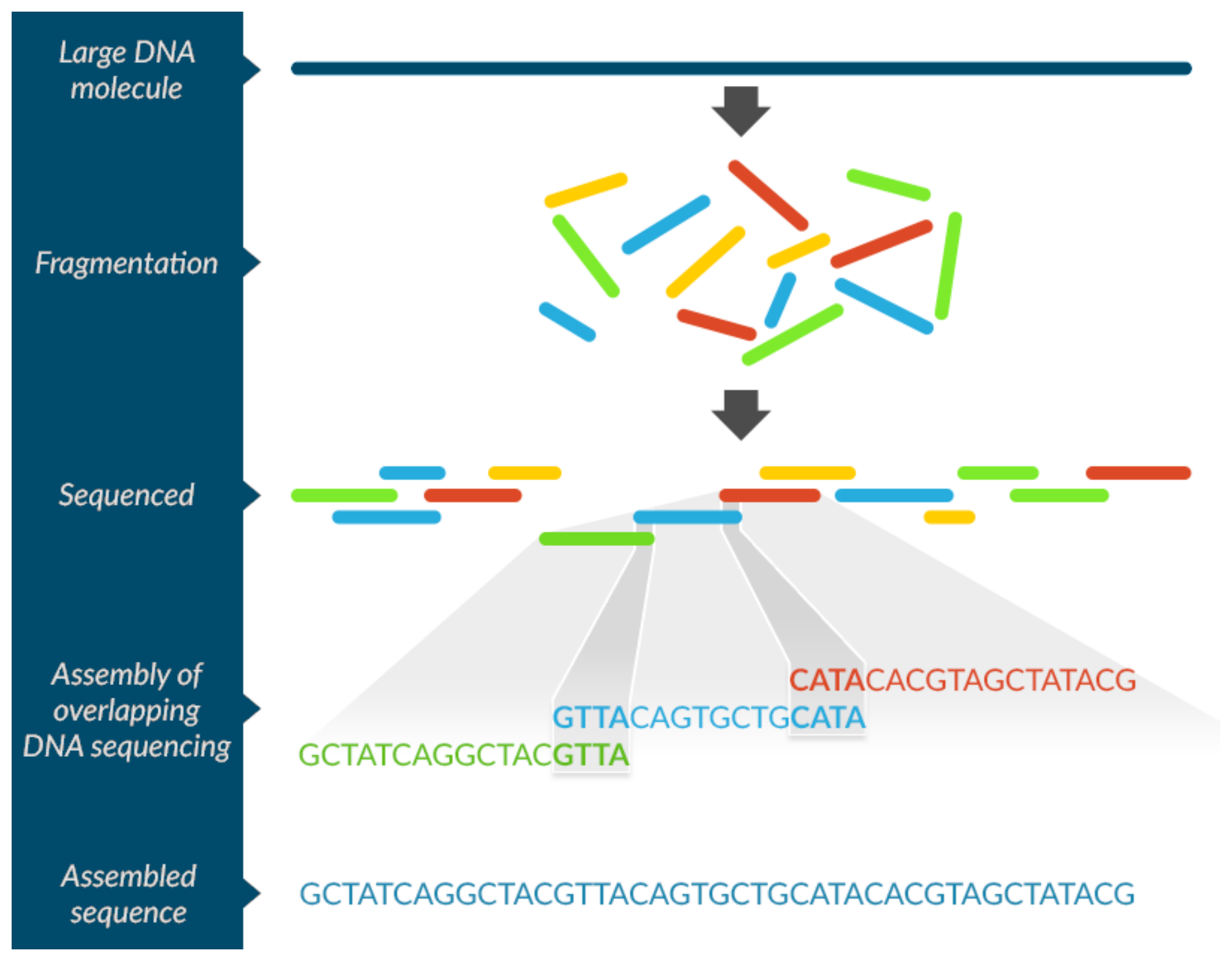
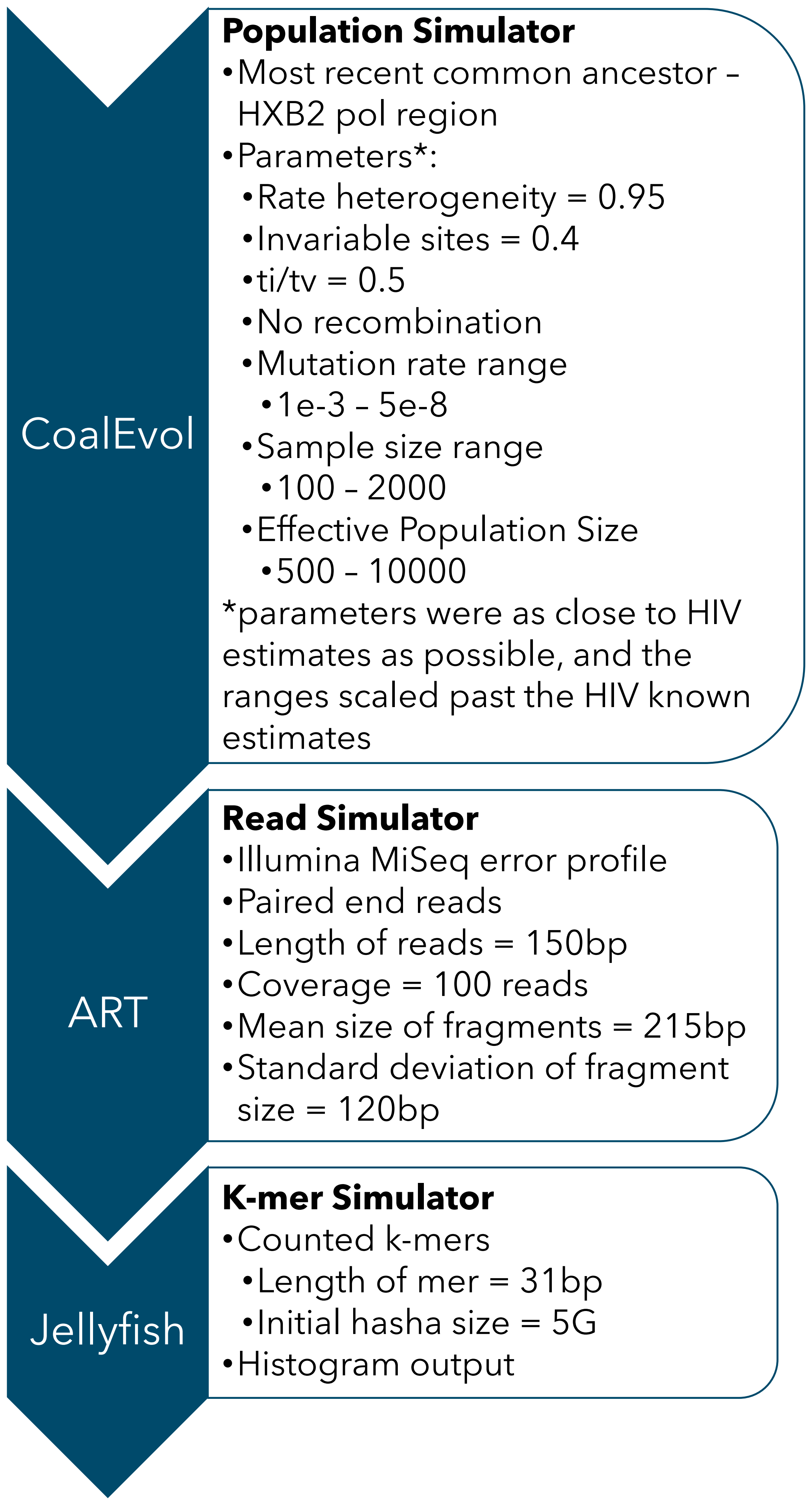


Fig. 1: A basic primer on assembly. Modified from Micron

Assembly is biased

1) Using a reference – introduces an inherent bias to that reference. Reads may be assembled in the wrong position. It's possible to be missing differences (SNPs) because of this assembling.
2) Phase uncertainty – there is error in read pairing, resulting in the incorrect reads being paired. This ultimately leads to missing SNPs as well. All of which presents issues for finding haplotypes, which sit at the core of understanding the genetic diversity of virus populations.

Data Generation



CoalEvol

Population Simulator
- Most recent common ancestor – HXB2 pol region
- Parameters*:
  - Rate heterogeneity = 0.95
  - Invariable sites = 0.4
  - ti/tv = 0.5
  - No recombination
  - Mutation rate range
    - 1e-3 – 5e-8
  - Sample size range
    - 100 – 2000
  - Effective Population Size
    - 500 – 10000

*parameters were as close to HIV estimates as possible, and the ranges scaled past the HIV known estimates

ART

Read Simulator
- Illumina MiSeq error profile
- Paired end reads
- Length of reads = 150bp
- Coverage = 100 reads
- Mean size of fragments = 215bp
- Standard deviation of fragment size = 120bp

Jellyfish

K-mer Simulator
- Counted k-mers
  - Length of mer = 31bp
  - Initial hasha size = 5G
- Histogram output

Fig. 2: Pipeline used to generate the simulated data to train the various regressors

Data Processing

Problem

- Jellyfish produces 11,690 k-mer histogram files
- Each k-mer histogram has a different number of bins
- Regression requires a fixed size matrix

Solution Requirements

- Represent heterogeneous histograms in a consistent way
- Representation must preserve seperability

Solution

Attempt several means to represent the data:

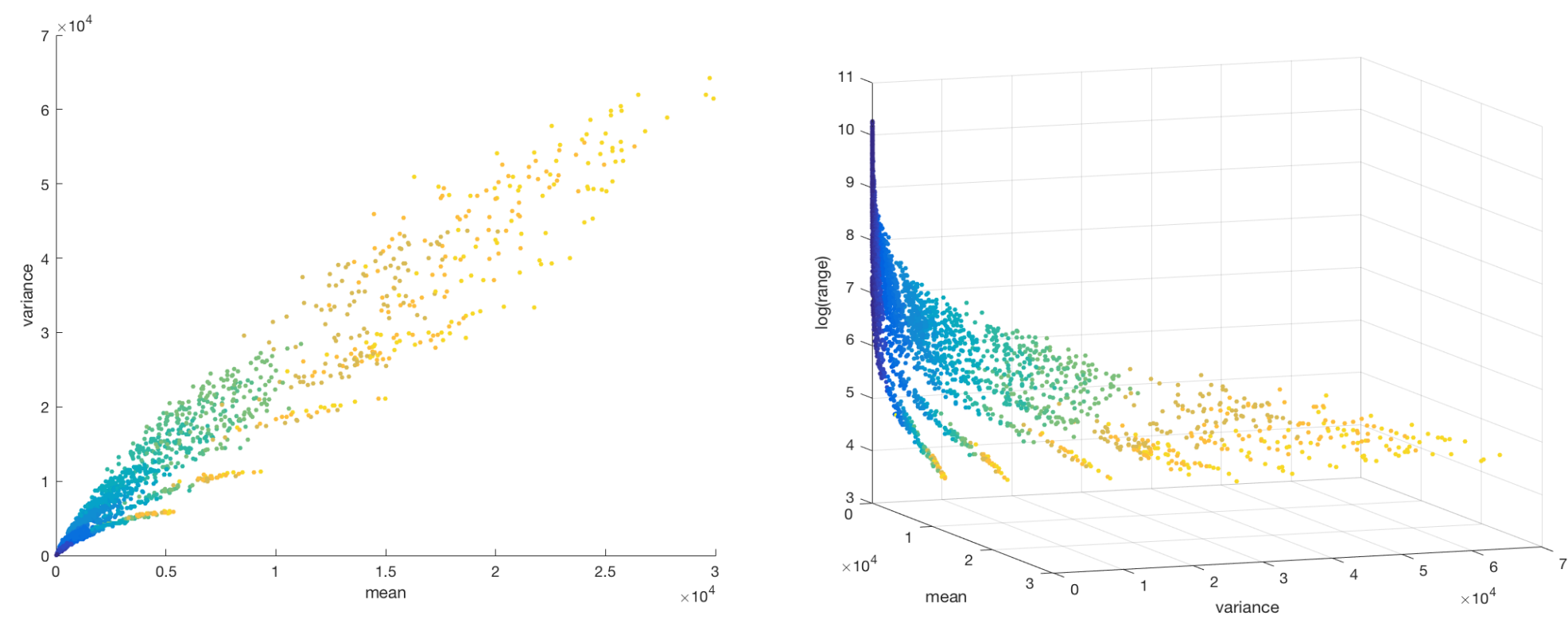


Fig. 3: Statistical tests applied to each k-mer analysis

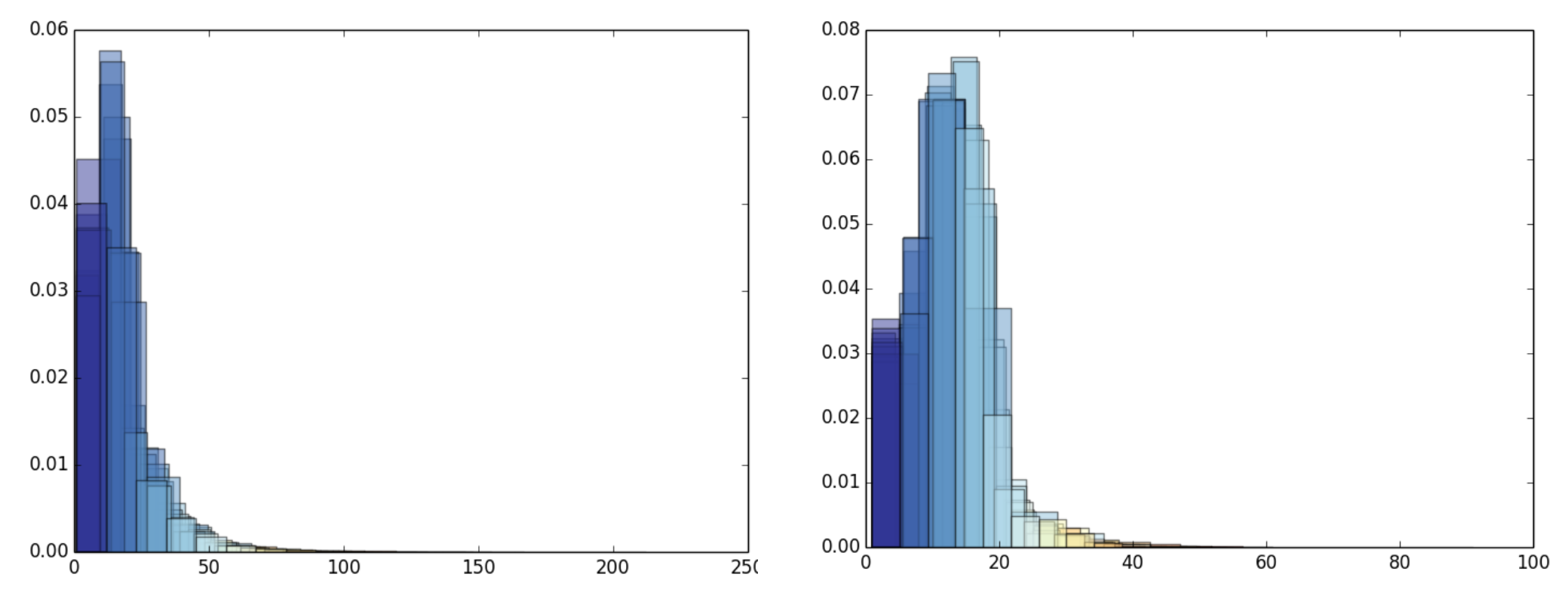


Fig. 4: Histogram representations generated from the k-mer counts

Run several regressors for all representations and take the best

Regressor Training

Regressor: Ridge Regression
Params: alpha, solver

Regressor: Lasso Regression
Params: alpha

Regressor: LassoLARS Regression
Params: alpha

Regressor: Bayesian Ridge Regression
Params: alpha_1, alpha_2, lambda_1, lambda_2

Regressor: K-Nearest Neigbor Regression
Params: algorithm, neighbors, weights

Regressor: Gradient Boosting Regression
Params: loss, learning rate, estimators, max depth

Grid Search

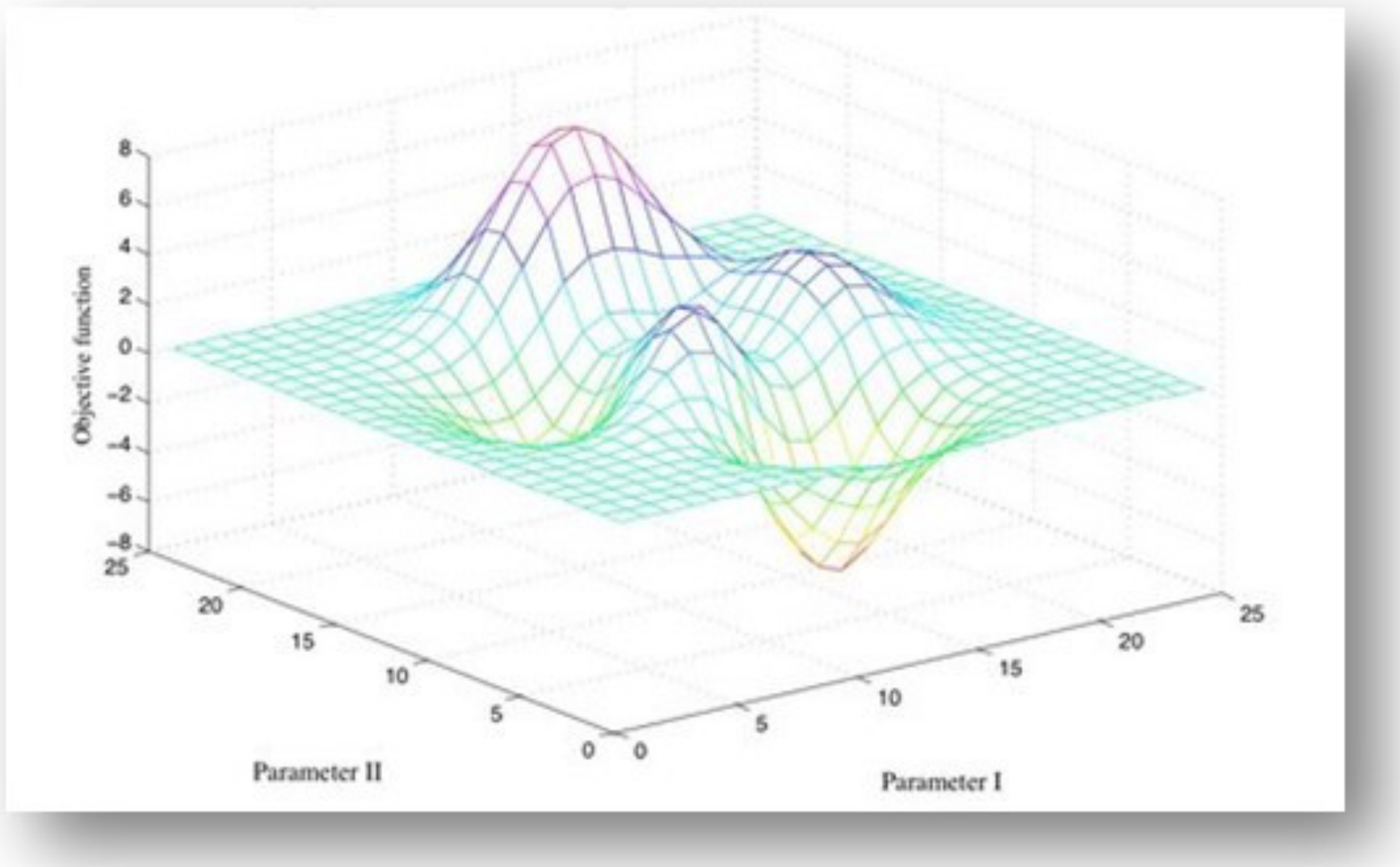


Fig. 5: Grid search is the process we used to automate the optimization of hyper parameters, minimizing the error of the objective function by the given parameters.

Results

	Ridge	LASSO	LassoLARS	Bayesian	KNN	Gradient
histogram	2302.86	2309.78	342.83	3173.64	0.58	6.70
count histogram	1934.52	959.25	393.32	833.01	0.26	1.25
mean, var	1107.55	329.83	102.06	1117.09	0.95	8.36
mean, var, skew	1145.73	404.38	147.05	946.34	0.79	19.49
mean, var, kurt	1106.13	413.95	147.05	1098.87	2.65	21.82
mean, var, len	895.29	402.00	147.05	960.10	0.98	3.10
mean, var, lglen	756.13	329.84	127.76	637.71	0.32	3.00

Table 1: The |E| error for each of the regressors on different representations of the data. The histogram and count histograms are two different histogram represntations. Mean, variance, skew, kurtosis and length and the log(length) are all summary statistics which were attempted to find a good subset of statistics for regression.

K-Nearest Neighbor and Count Histogram

- A learning algorithm based on the most similar datapoints to the query point
- A representation preserving as much of the original histogram information as possible

These two techniques compliment each other, the information maintained in the count histogram contributes to the similarity KNN uses.

Analysis

- An error rate of 26% represents a success, showing that regression from unassembled reads is a valid approach
- The error is only as high as it is because on edge cases, when the regressor is wrong, it is very wrong. This is in large part an issue with the largely non-continuous training set.

Future Work

- Simulate more data to improve the accuracy of the KNN regressor
- Attempt a deep learning classifier, properly tuned, as a true regressor may be more useful than a KNN
- Test on true data and on more continuous simulated data

Works Cited

- Maldarelli, F., Kearney, M., Palmer, S., Stephens, R., Mican, J., Polis, M. A., ... & Metcalf, J. A. (2013). HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *Journal of virology*, *87*(18), 10313-10323.
- Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., ... & Batzer, M. A. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *science*, *316*(5822), 222-234.
- Zukurov, J. P., Nascimento-Brito, S., Volpini, A. C., Oliveira, G. C., Janini, L. M. R., & Antoneli, F. (2016). Estimation of genetic diversity in viral populations from next generation sequencing data with extremely deep coverage. *Algorithms for Molecular Biology*, *11*(1), 2.
- Bioinformatics. Bioinformatics. Micron, n.d. Web. 05 May 2017.
- Arenas M, Posada D. (2014). "Simulation of Genome-wide Evolution under Heterogeneous Substitution models and Complex Multispecies Coalescent Histories." Molecular Biology and Evolution 31 (5):1295-1301.
- Huang, W., Li, L., Myers, J.R., Marth, G.T. (2012) "ART: a next-generation sequencing read simulator". Bioinformatics. 28 (4):593-594.doi:10.1093/bioinformatics/btr708.
- Melsted, Pall, Jonathan K. Pritchard (2011) "Efficient counting of k-mers in DNA sequences using a bloom filter." BMC bioinformatics. 12.1: 333.