Inclusion of Sequencing Error in Substitution Models for more Realistic Genetic Analyses

Justin Bricker Advisor: Dr. Peter Beerli

Department of Scientific Computing, Tallahassee, FL

Abstract: Next generation sequencing can rapidly analyze entire genomes in just hours. However, due to the nature of the sequencing process, errors may arise which limit the accuracy of the reads obtained. Luckily, modern sequencing technologies associate with their reads, a quality score, derived from the sequencing procedures, which represents the rate of error for each nucleotide in the sequence. Currently, these quality scores are used as a criteria for the removal or modification of reads in the data set. These methods result in the loss of information contained in those sequences and rely on parameters that are somewhat arbitrary; this may lead to a biased sample. We propose an alternative method for incorporating the error of the sequences without discarding poor quality reads by including the error probabilities of the reads in the substitution models used for sequence analysis. While this method will result in analyses with less defined results, these results will be more grounded in reality as we take into account the uncertainty that we have in our sequenced samples.

BACKGROUND

How error estimates are made in sequenced reads:

In genetic sequencing, base calls are made by analyzing the traces that are produced in the sequencing reaction. These traces are represented as sets of time-dependent light-intensities (peaks) of different wavelengths (representing the 4 possible nucleotides). The appropriate nucleotide base is attributed to a particular peak (known as "base calling") by considering the order, spacing, width(among other attributes) of the light signals in the trace [1].

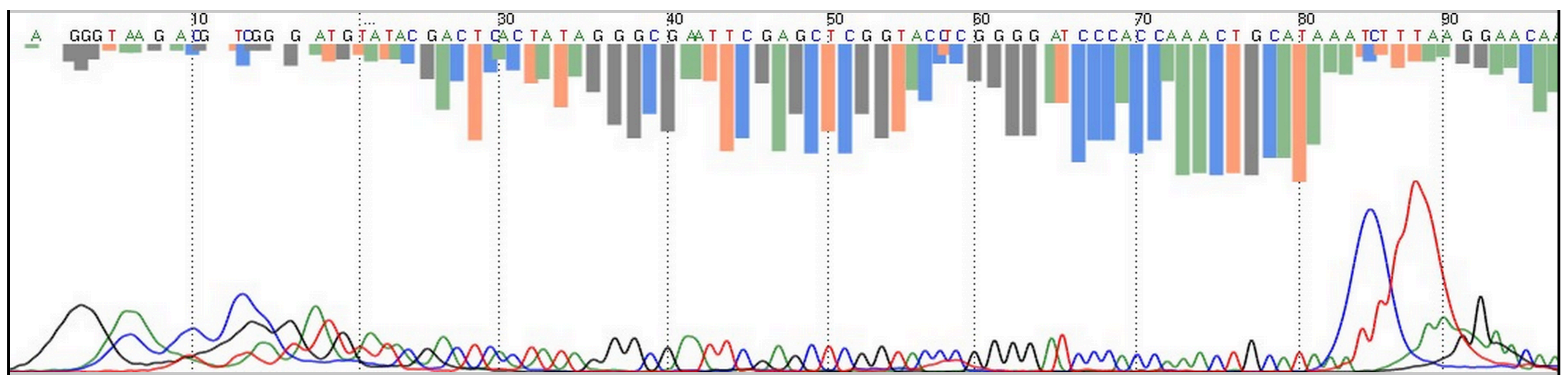


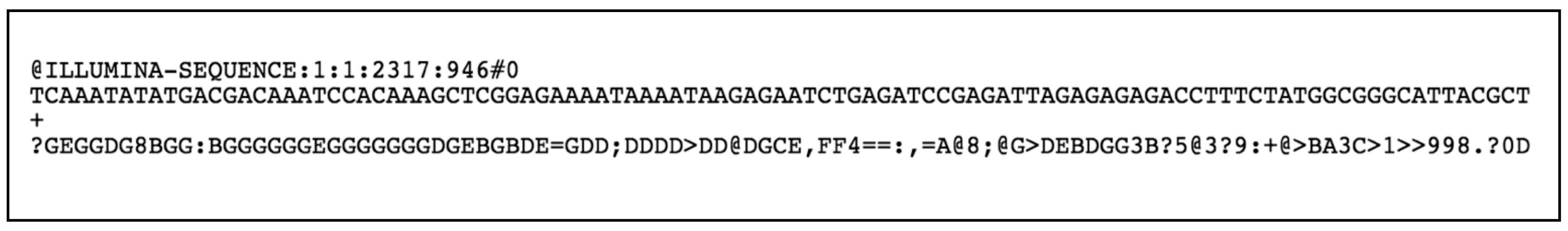
Figure 1: *DNA Sequence Trace Chromatograph used to determine base calls and associated quality scores*

This base-calling process is not without error and, as a result, bases observed in the resulting read sequences may not represent the true sequence. However, we can predict the rate of error (also represented as a quality score) of these base calls by sequencing a known data set and modeling the error based on parameters of the peaks in the trace [2]. This produces a lookup table of parameters and their predicted error/quality scores, allowing us to associate new base calls with a measure of uncertainty with reasonable accuracy.

QUAL Scores and FASTQ Files

Sequences of nucleotide bases produced using modern sequencing techniques are most often paired with an associated sequence of quality scores which have been mapped to a set of single ASCII characters[3]. For example, the Illumina Sequencing Platform uses ASCII character codes between 33 and 74. These codes map to the quality scores 0-41, representing error probabilities between 1 (an incorrect base call) and $10^{-4.1}$ (an extremely accurate base call), according to the formula Q = -10 log(P), where *P* is the error probability and *Q* is the quality score for a particular base.

The quality score information is packaged with the base call information in the form of the FASTQ data file type. An example of sequence data in this format is seen in the figure below, in which the first two lines are identical to the lines that would be seen in the standard FASTA format, the third line begins with the '+' symbol (which can optionally be followed by the identical title of the first line, and the last line contains the quality information, as mapped to the appropriate ASCII characters. Its important to note, that FASTQ formats between modern sequencing technologies are most often incompatible [3] and thus, care must be taken to convert the ASCII characters in a given FASTQ file to the appropriate quality score depending on the machine used to sequence the reads.



```
@ILLUMINA-SEQUENCE:1:1:2317:946#0
TCAAATATATGACGACAAATCCACAAAGCTCGGAGAAAATAAAATAAGAGAATCTGAGATCCGAGATTAGAGAGAGACCTTTCTATGGCGGGCATTACGCT
+
?GEGGDG8BGG:BGGGGGGEGGGGGGGDGEBGBDE=GDD;DDDD>DD@DGCE,FF4==:,=A@8;@G>DEBDGG3B?5@3?9:+@>BA3C>1>>998.?0D
```

Figure 2: *Example FASTQ File*

Sequence preprocessing using quality scores

Ignoring the quality scores of reads can be detrimental to the accuracy of downstream analyses. Numerous errors can occur in the sequencing and base-calling procedures, such as hairpinning and noise in sequence traces, which result in substitutions, insertions, and deletions in the sequenced reads. However, these errors may be recognized from the quality data associated with the reads.

Multiple preprocessing algorithms are available to handle reads with high error probability, such as read trimming [4], read correction [5], duplicate removal [6], contaminant sequence filtering [7], and adapter removal [8]. While these methods improve the accuracy and decrease computational time of downstream analyses, they often do so by removing poor quality reads (losing information from the sample) and using program parameters that are determined through trial and error (lacking "universability").

The Genetics "*LINGO*"

- ***DNA-*** *genetic material which contains information regarding the characteristics and functions of living organisms.*
- ***Nucleotide-****the molecular building blocks of DNA, represented as 4 characters:* ***A, T, C, and G.***
- ***Next-Generation Sequencing-*** *determination of nucleotides in DNA using modern, cheaper, high-throughput techniques which are more error prone.*
- ***Substitution Model-*** *a mathematical representation of the nucleotides in DNA change over time.*
- ***Substitution Matrix-*** *a matrix used in substitution models, which describes the relative rates of transition between the nucleotide types.*
- ***Genetic Analysis****- any kind of analysis done post-sequencing for the purpose of extracting information from the DNA.*

My project seeks to address these issues by introducing a scheme to take into account the uncertainty in sequenced reads using a statistical framework and without discarding the information contained in those reads.

METHODS

Standard Substitution Model

Assume, for simplicity, that we have 2 states, *U* and *V*. Then the probability of transitioning from the state *U* to the state *V* in time *t* is denoted by $P_{U,V}(t)$, an element of the matrix *P(t), calculated as:*

$$P(t) = \begin{pmatrix} P_{U,U}(t) & P_{U,V}(t) \\ P_{V,U}(t) & P_{V,V}(t) \end{pmatrix} = e^{t(R-I)_{2x2}} = e^{t\begin{pmatrix} r_{U,U}-1 & r_{U,V} \\ r_{V,U} & r_{V,V}-1 \end{pmatrix}}$$

Where *R* is a 2x2 matrix containing the relative rates at which a state is substituted to another (or the same) state when a substitution occurs.

Our Method

Again, assume that we have 2 states, *U* and *V*. However now let ε denote the probability that the first state is not actually U and let δ denote the probability that the second state is not actually V. Then we could write our transition probability as a weighted sum of the various combinations of possible states and their respective transition probabilities (we'll denote this as $P^*_{U,V}(t)$) *:*

$$P^*_{U,V}(t) = (1-\epsilon)P_{U,V}(t)(1-\delta) + (\epsilon)P_{V,V}(t)(1-\delta) + (1-\epsilon)P_{U,U}(t)(\delta) + (\epsilon)P_{V,U}(t)(\delta)$$

This can be formulated more succinctly as a matrix formula, replacing a single substitution matrix with the product of three matrices (the middle matrix being the original substitution matrix and the outer matrices describing the uncertainty weights) as follows:

$$P^*(t) = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} \begin{pmatrix} P_{U,U}(t) & P_{U,V}(t) \\ P_{V,U}(t) & P_{V,V}(t) \end{pmatrix} \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}$$

This produces a new substitution matrix from which we can obtain appropriate transition information between sites of two sequences given the quality score at those two sites.This matrix, using the standard 4 nucleotide states becomes:

$$P^*(t) = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} \begin{pmatrix} P_{A,A}(t) & P_{A,T}(t) & P_{A,C}(t) & P_{A,G}(t) \\ P_{T,A}(t) & P_{T,T}(t) & P_{T,C}(t) & P_{T,G}(t) \\ P_{C,A}(t) & P_{C,T}(t) & P_{C,C}(t) & P_{C,G}(t) \\ P_{G,A}(t) & P_{G,T}(t) & P_{G,C}(t) & P_{G,G}(t) \end{pmatrix} \begin{pmatrix} 1-3\delta & \delta & \delta & \delta \\ \delta & 1-3\delta & \delta & \delta \\ \delta & \delta & 1-3\delta & \delta \\ \delta & \delta & \delta & 1-3\delta \end{pmatrix}$$

DISCUSSION

It is important to note that this method will not improve or sharpen the results of downstream analyses- quite the opposite, in fact. Correction with the uncertainty matrices dampens the transition probability matrix by a factor proportional to the reads' accuracy (Figure 3). Therefore, results from genetic analyses will appear less significant than they might if we ignore error altogether or throw out reads with low quality scores. However, by including low quality reads and taking into account the error probabilities of all reads, we may obtain results that, while not better, are more grounded in reality.

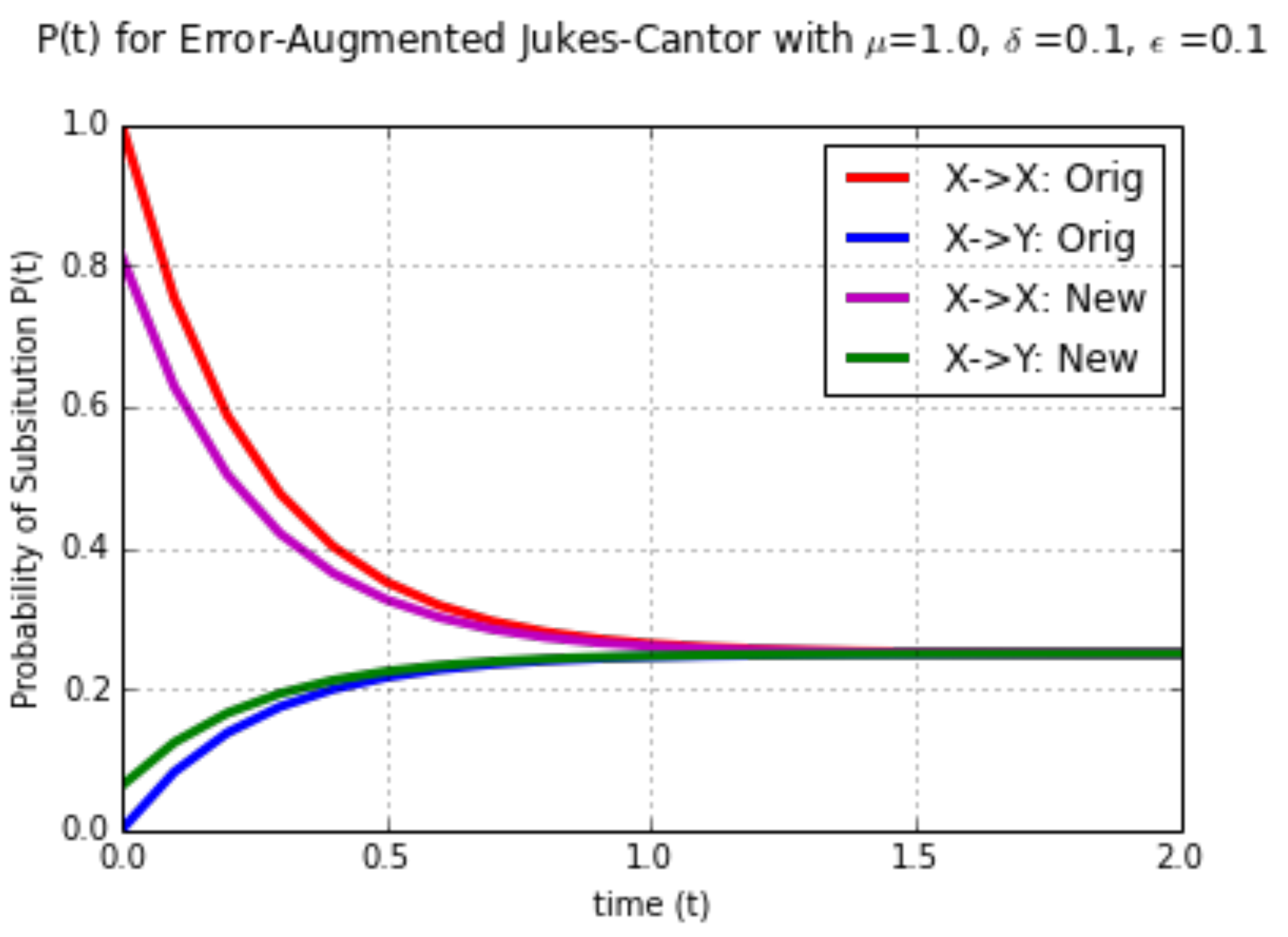


Figure 3: *Transition Probabilities over time for the standard substition model (red, blue) and error augmented substitution model (pink, green)*