The Extraction of Single Nucleotide Polymorphisms and Understanding of Current Sequencing Tools

Stephen Tetreault

Department of Mathematics and Computer Science
Rhode Island College
Email: Stephen Tetreault – tetreault1@gmail.com

Abstract

Single nucleotide polymorphisms (or SNPs, pronounced as “snips”) are DNA sequence variations that occur when one nucleotide in the genome sequence has been altered. SNPs can occur in coding and noncoding regions of the genome and many actually have no effect on cell function, but could predispose people to disease or alter their reaction to agents such as medication. Many are working to discover and catalog more SNPs to create a more detailed and medically useful picture of human genetic variation, the 1000 Genomes Project is one such consortium and is also where the genomes were obtained from. To work with such genomic data many alignment programs are now widely, the main program that was used was SAMtools. SAMtools was used to receive raw data by printing the alignment in the pileup format and in calling the consensus sequence using the MAQ consensus model. In an attempt to extract only higher quality SNPs a minimum quality score of 20 is used which means the base call accuracy will be 99% and the probability of an incorrect base call is 1 in 100. For each participant genome used variant frequency is calculated and to make the task of finding which regions have more variation easier, the occurrences of SNPs by region are laid out in graphs.

Keywords: Single nucleotide polymorphism, SAMtools, alignment, genotyping

1 Background

Single Nucleotide Polymorphisms, or SNPs, make up the majority of all human genetic variation. The number of SNPs in the human genome so far is around 1.4 million with two haploid genomes differing at 1 nucleotide per 1,331 bp. And despite any two human genomes being around 99.9% identical, that would result in millions of differences among the 3.2 billion base pairs [4]. Since most DNA does not code for proteins, the great majority of SNPs have no effects on function and are known as “silent variation.” However, for some SNPs one allele can cause major health problems. These sequences can impact how humans respond to disease, environmental factors, chemicals, drugs, etc. Finding these SNPs are of high interest and figuring out a SNP’s genotype for an individual participant is important. SNPs are also thought to be crucial in the effort to personalize medicine.
Collaboration has been widespread to work on furthering the current knowledge of human genetic variation. The 1000 Genomes Project is one such project that plans to sequence the genomes of at least one thousand anonymous participants from a number of ethnic groups utilizing newer, less expensive technologies (e.g. Solexa, 454, SOLiD) which prove to be faster. The project is carried out in three steps. The first, the detailed scanning of six individual participants. This will then be followed by a less detailed genome scan of 180 anonymous participants, and lastly followed by partial scans of 1000 additional people [1]. The participant data used within this project was of the partial scans of participant data utilizing only the data from chromosome 1.

With the advent of novel sequencing technologies such as Illumina/Solexa, AB/SOLiD and Roche/454 a variety of new alignment tools have been designed to carry out efficient read mapping against large reference sequences, including the human genome [2]. The primary focus of the project has turned into a study of the current state of sequencing tools and learning how to utilize them and understanding the raw data that is output by them. Then writing a program to automate the process of going through the raw data and extracting where SNPs occur and the corresponding information to those SNPs, to calculate simple variant frequencies. And also to lay the number of occurrences out in a graph to easily see which regions have more occurrences of variation within the chromosome. Further advanced analysis into genotyping will be discussed within the proposed future works section.

2 Discussion

2.1 Data and Tools

A list of the resources and tools used is given below.

- 1000 Genomes Project data
- SAMtools 0.1.5
  http://sourceforge.net/projects/samtools/files/
- MAQ 0.7.1
  http://sourceforge.net/projects/maq/files/

2.2 Sequencing

The first efforts to align and sequence data were attempted with the program MAQ (Mapping and Assembly with Qualities) which maps short reads to references and calls the genotypes from the alignment. MAQ maps a read to the position where the sum of quality values of mismatched nucleotides is minimum. If there is a read with several equally best positions then MAQ maps it randomly and MAQ gives each read alignment a mapping quality [4, 8]. Due to the large amount of time MAQ would take to run it’s alignment and the limited computing power the sequencing program SAMtools was used.

SAMtools is a set of utilities that allows for manipulation to be carried out on a BAM file (Binary SAM file) [9] which were the prealigned files in the participant da-
to avoid the long alignment times experienced with MAQ. This project will utilize the SAMtools pileup which describes the base-pair information at each chromosomal position. Using this format one can carry out SNP calling accompanied by brief alignment viewing. There is the option of generating the consensus sequence with the model that is implemented within MAQ. With the data that is output, the consensus quality is the Phred-scaled probability that the consensus is incorrect [6]. The SNP quality score given is the Phred-scaled probability that the consensus is different from the reference. For the task of SNP calling the SNP quality is of more importance. The high accuracy of Phred quality scores makes them a useful tool in assessing the quality of sequences.

<table>
<thead>
<tr>
<th>Quality of Phred Score</th>
<th>Probability of Incorrect Base Call</th>
<th>Base Call Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

Table 1. Phred quality scores and how they relate to accuracy of the base called.

3 Finding and Analyzing SNPs

3.1 Project Data

With the use of SAMtools raw data that was output in text format is shown below:

<table>
<thead>
<tr>
<th>1</th>
<th>60</th>
<th>T</th>
<th>T</th>
<th>66</th>
<th>0</th>
<th>99</th>
<th>13</th>
<th></th>
<th>;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61</td>
<td>G</td>
<td>C</td>
<td>26</td>
<td>26</td>
<td>99</td>
<td>15</td>
<td></td>
<td>;</td>
</tr>
</tbody>
</table>

Table 2. Data received from SAMtools Pileup and Consensus calling.

The raw data received through SAMtools pileup and consensus calling contains the following: chromosome, position, reference base, consensus base, consensus quality score, SNP quality score, maximum mapping quality score, number of reads mapped, read bases, and base qualities [9].

3.2 Finding SNPs of Higher Quality

Methods of finding higher quality SNPs [3]:

- Look at the number of reads covering the SNPs. Discard those covered by three or fewer reads.
While the consensus quality is of importance, when calling SNPs it is most important to make sure to discard SNPs with a SNP quality less than 20. According to the Phred quality scores, this guarantees that the probability of being incorrect is 1 in 100. The key to reliable detection of SNPs is the ability to discern true allelic variation from sequencing error [7].

3.3 A Program for Extracting SNPs

It was necessary to write a program to automate the process of finding SNPs within the raw data from SAMtools pileup. Written in Java, the program was created to do the following:

1. Read in the raw data (example shown in table 2) per column, line by line. And per line test to see if there is a SNP at that position and one of high quality by comparing the reference and consensus base, then making sure the SNP quality score is above or equal to 20 which ensures a base call accuracy of 99%. If it proves to be a SNP of high quality it is then read in as an object into an array list, also being stored in the order of their positions and written out to a text file.

   Chromosome: 1
   SNP Position: 1125804
   Reference Base: T
   Consensus Base: C
   Consensus Quality: 36
   SNP Quality: 36

   Fig. 1. An example SNP from the project data as output from step 1 of the program.

2. When a SNP is found, counts kept for finding the variant frequency (for example, the frequency of bases being changed from A to C) are updated and are printed at the end of the program.

3. When all SNPs are found for chromosome 1 from a participant genome, the program then finds the number of occurrences of SNPs per 100,000 bases to find the amount of variation per region throughout chromosome 1. This information is later laid out in a graph.

4. Advanced analysis is to be added to the program for SNP genotyping, which will be proposed later in the future works section.

4 Results

After running genomes through SAMtools pileup and consensus calling to obtain the raw data, and running that raw data through the variant frequencies were calculated. From those results it is shown that the base change A to G and the base change T to C are the most frequently occurring variations while the base change of C to G was the
least frequent. As discussed in the third step of the program, the number of occurrences of SNPs per 100,000 bases were laid out graphically to make comparisons of the number of occurrences per region between participant genomes easier, and to also quickly see which region of chromosome 1 possesses more variation.

Fig. 2. The number of SNPs per 100,000 bases for chromosome 1 for participant genome NA07048 from the 1000 Genomes Project.
Fig. 3. The number of SNPs per 100,000 bases for chromosome 1 for participant genome NA12273 from the 1000 Genomes Project. SNPs appear more clustered together throughout the chromosome.

5 Conclusion

This project was met with some initial complications, one of which was issues in accessing the genomic data from the 1000 Genomes Project for about the first week and a half. And then it was necessary to switch from MAQ to SAMtools because running the data through MAQ took very long and needed more computing power than was available. With SAMtools it was possible to achieve the SNP calling much faster with the prealigned BAM files. But despite the complications that arose and the fact that the summer program was only 8 weeks long, it proved very useful for gaining a more thorough understanding of an area of bioinformatics, and helpful to now have the knowledge of how to use these sequencing tools. This was an experience that was rewarding and will be encouraging in plans to attend graduate school.

6 Future Work

6.1 fastPHASE

FastPHASE is used for estimating missing genotypes and and for reconstruction of haplotypes using unphased SNP genotype data from unrelated individuals [5].

6.2 Obtaining Haplotypes
More advanced analysis that is planned for the future is to work with the genotypes by calling the genotypes from the reads and running fastPHASE to obtain corresponding haplotypes. What has already been done, as discussed above, is to find out which positions are varying within the population. We look at chromosome 1 for one of the individual participants, and we look at the reads mapped covering the varying position and look at what the bases are for that position and determine whether it is heterozygous or homozygous.

Acknowledgments

ST performed this research as part of the Summer REU program “Bio-Grid Initiatives for Interdisciplinary Research and Education” at the Computer Science and Engineering Department, University of Connecticut, supported by NSF Award CCF-0755373 under the mentorship of Yufeng Wu and with help from graduate student Jin Zhang.

References