Iterative Referencing for Improving the Interpretation of DNA Sequence Data

Alaa Ghanayim
Iterative Referencing for Improving the Interpretation of DNA Sequence Data

Research Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science

Alaa Ghanayim

Submitted to the Senate of the Technion - Israel Institute of Technology

Av 5773 Haifa May 25, 2014
The research thesis was done under the supervision of Prof. Dan Geiger in the Department of Computer Science.

Acknowledgment

I would like to express my deepest gratitude to Prof. Dan Geiger. His devoted guidance and encouragement made this work possible.

I am thankful for everything I learned from Dan. His advices, ideas and standard of work will surely assist me well beyond the scope of research.
Contents

Abstract 1

1 Introduction 3
   1.1 Related work 4
   1.2 Method and results 5

2 Mapping and Annotation Tools 6
   2.1 BWA 6
   2.2 BOWTIE 6
   2.3 SAMTOOLS 7
   2.4 GATK 7
   2.5 ANNOVAR 7
   2.6 CoNIFER 8

3 Iterative Referencing Algorithm 9
   3.1 IR Schematic Description 9

4 Expectation Maximization in IR 14
   4.1 EM in IR 15

5 Results 17
   5.1 Results on Simulated Data 17
   5.2 Results on Real Data 20

6 Conclusions, discussion, and directions for further research 32

Bibliography 33
List of Figures

3.1 Flow Chart for algorithm IR ..................................... 12
3.2 Mapping in two iterations. Figure (a) shows 6 reads and the variations between the sequenced DNA and the initial reference model. Figure (b) demonstrates that in the first iteration, IR maps the first 4 reads to the reference genome and discovers a mismatch of G and a deletion of T. In this example, reads are assumed to be mapped successfully when at most one variation exists wrt the current reference model. Then, IR generates a revised reference model. Figure (c) shows that all 6 reads are mapped perfectly and that the last revised reference model, in this example, becomes identical to the sequenced DNA. ............................................. 13

5.1 Number of true SNPs reported after 4 iterations as compared to BWA and BOWTIE2, with variation rate of 0.05. The total number of reported true SNPs increased by 3.5% .................. 19
5.2 Number of false SNPs reported after 4 iterations as compared to BWA and BOWTIE2, with variation rate of 0.05. The total number of reported false SNPs decreased by 1.5% ................. 20
5.3 Number of true and false INDELs reported after 4 iterations as compared to BWA, with variation rate of 0.05. Same results hold for IR_BOWTIE2 versus BOWTIE2 .................................. 21
5.4 The improvement spectrum of uniquely mapped reads as a function of the variation rate. This graph shows that the improvement of IR is maximal (6.8%) when the variation rate is between 0.045 and 0.07 .............................................................. 22
5.5 Lod scores of the sequenced family using the outputs of BWA and IR_BWA, respectively .............................................. 24
5.6 Pedigree segregating a recessive trait. Individuals 7,10,11 and 12 have been sequenced .............................................. 25
5.7 Pedigree segregating a dominant trait. Six individuals have been sequenced.

5.8 Lod score for chromosome 17. Only one suspected region was detected with LOD score above 1.

5.9 CoNIFER output. It shows different frequencies for the affected individuals in position 61.8 million bp.

5.10 Each entry in the table is the $\log_{10}$ likelihood of the data (for all analyzed pedigrees together and for each pedigree alone) for a specific Mode Of Inheritance (MOI) and specific penetrance probabilities. When comparing two lines, one is checking the likelihood of a dominant MOI versus a recessive MOI for a specific penetrance value. The highest value, namely when the negative number is closest to zero, is the more likely MOI.

5.11 Lod score for the first 12 chromosomes.

5.12 Lod score for the last 10 chromosomes from the whole genome. It was computed by summing the LOD scores of the three separated families. It is obvious from the figures below that only one range in chromosome 2 is suspicious to be linked to the CDD disease.
### List of Tables

<table>
<thead>
<tr>
<th></th>
<th>Results based on simulating seventy million 100bp reads from the Human genome, with induced variation rate 0.05 for different rates of INDELs. Each row shows the log likelihood of the seventy million reads for each IR iteration. The columns reflect increasing percentage of INDELs within the 5% variations used to simulate the reads.</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Results based on simulating seventy million 100 bp reads from the Human genome with average coverage of 4.82 folds, with induced variation rate 0.05, for different rates of INDELs. Each row shows the percentage of uniquely mapped reads for each IR iteration. The columns reflect increasing percentage of INDELs within the 5% variations used to simulate the reads.</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>On average, after one update by IR_BWA of the reference model additional 88556.8 reads among the 812457 (10.8%) unresolved reads were mapped uniquely to the initial reference genome and by the third update a total of 172389.4 reads (21.2%) of the unresolved reads were mapped uniquely.</td>
<td>24</td>
</tr>
</tbody>
</table>
Abstract

Next-Generation Sequencing (NGS) facilitates genetic studies to discover SNPs and indels associated with Mendelian and complex diseases. The measurement process, which generates millions of short reads, creates various data processing and interpretation challenges for which a multitude of software tools are being developed. A common framework used to date to discover variations in sequenced data includes the following steps in a pipeline: First, mapping the sequenced reads to some reference the genome. Second, local realignment to account for indels. Third, recalibration of the base quality score and discovering variations by using the GATK or SAMTOOLS packages.

Improving the accuracy of discovering true SNPs and indels in analyzing NGS data requires improved tools and capabilities in each step of this pipeline. The mapping accuracy is the bottleneck of this process because reads mapped incorrectly dramatically increase false positive rates of discovered variations and lower the rate of detected true positive variations. We present a revised approach, Iterative Referencing (IR), that increases the accuracy of mapping the sequence data by iteratively improving the reference genome via the Expectation maximization (EM) algorithm.

The idea is that if sufficient coverage in the sequenced data contain a specific homozygous SNP that is not seen in the reference genome then the reference genome should be altered to contain that SNP. Such a situation occurs when a reference genome is used that is not close enough to the population under study. In each step, the EM algorithm discovers the new homozygous variations in the whole genome, then alters the reference genome by replacing the reference bases by the alternative bases and builds a new reference genome that is more appropriate to the population under study. The results demonstrate that using the updated reference genome improves the alignment process up to 6%, increases the rate of true variations by up to 3%, decreases the rate of false variations up to 3%, all measured with respect to the original reference model.
## Abbreviations and Notations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum likelihood estimate</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>WES</td>
<td>Whole Exome Sequencing</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole Genome Sequencing</td>
</tr>
<tr>
<td>LOD</td>
<td>Logarithm of Odds</td>
</tr>
<tr>
<td>MOI</td>
<td>Mode of Inheritance</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>INDEL</td>
<td>Insertion or Deletion</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy Number variant</td>
</tr>
<tr>
<td>LOH</td>
<td>Loss of heterozygosity</td>
</tr>
<tr>
<td>GATK</td>
<td>Genome Analysis Toolkit</td>
</tr>
<tr>
<td>AML</td>
<td>Acute Myeloid Leukemia</td>
</tr>
<tr>
<td>QUAL</td>
<td>Variation quality</td>
</tr>
<tr>
<td>DP</td>
<td>Total depth over all reads</td>
</tr>
<tr>
<td>BWT</td>
<td>Burrows Wheeler Transform</td>
</tr>
<tr>
<td>VCF</td>
<td>Variant Call Format</td>
</tr>
<tr>
<td>SAM</td>
<td>Sequence Alignment/Map</td>
</tr>
<tr>
<td>BAM</td>
<td>Binary Alignment/Map</td>
</tr>
<tr>
<td>AHUS</td>
<td>Atypical Hemolytic Uremic</td>
</tr>
<tr>
<td>CCD</td>
<td>Congenital chloride diarrhea</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Next-Generation Sequencing (NGS) steadily replaces SNP array technology in genetic studies that search genomic variations associated with Mendelian and complex diseases. The NGS measurement process, which generates millions of short reads, creates various data processing and interpretation challenges for which a multitude of software tools are being developed. A common framework used to date to discover variations in sequenced data includes the following steps in a pipeline: First, mapping the sequenced reads to some reference genome. Second, local realignment to account for INDELs. Third, recalibration of the base quality score and discovering variations by using, for example, the GATK or SAMTOOLs packages. And finally, annotating the variations by using ANNOVAR and filtering the results by software such as VARSIFTER.

This framework suggests mapping the reads against a reference genome to enable researches to annotate all known variations. Many methods are developed to improve the efficiency of each step due to the high error rate of the identified variations as a result of base calling and mapping errors. The alignment step is the key of the NGS analysis because mapping the reads to the wrong position produces false variations and may miss to detect others. The need for assigning sequenced reads to their genomic locations on the reference genome, a task called mapping, has led to the development of significant tools such as the state of the art software BWA, BOWTIE2, or CROSSBOW. The fundamental assumption when using a reference genome is that only minor variations exist between the sequenced data and the reference genome. A critical flaw of this approach is revealed when there is a sizable range in the sequenced genome that is condensed with variations, because the mapping tools are often unable to map the sequenced reads of this range. Therefore, the number of uniquely mapped...
reads is decreased and the quality of the detected variations is decreased as well.

Improving the accuracy of discovering true SNPs and INDELs in analyzing NGS data requires improved tools and capabilities in each step of the analysis pipeline. The mapping accuracy is the bottleneck in this process because reads mapped incorrectly dramatically increase false positive rates of discovered variations and lower the detection rate of true positive variations. Improving the mapping process of the framework to obtain accurate variations, raising the rate of uniquely mapped reads, and still allowing researchers to use any state of the art tools developed for individual steps has led us to suggest a new procedure. In this article, we present a revised approach, Iterative Referencing (IR), that increases the accuracy of mapping the sequence data by iteratively improving the reference genome via the Expectation Maximization (EM) algorithm. The idea is that if sufficient number of reads in the data set contain a specific homozygous variation that is not seen in the reference genome, such as SNP or INDEL, then the reference genome should be altered to contain that variation. Such a situation occurs when a reference genome is used that is not close enough to the population under study. In each step, the algorithm IR discovers the new homozygous variations in the whole genome, then alters the reference genome by replacing the reference bases by the alternative bases and builds a new reference genome that is more appropriate to the data under study.

1.1 Related work

The iterative approach we offer herein is novel and has not been discussed earlier. The closest work is perhaps by DePristo et al. [4]. Their work is dedicated to reducing the error rate of sequencers by using base quality score recalibration that assigns well-calibrated probability to each variant based on known variations provided to the algorithm, and integrates a local realignment algorithm that improves the detection of gaps and INDELs by constructing haplotypes from the reference model and from known INDELs. This framework shares some degree of similarity to our work in its goals, but does not iteratively generate a revised reference model. Furthermore, researchers can use IR with different reference model that better represents the population under study to save time and improves the quality without using the original reference model. Notably there are also studies that focus on improving the quality of the sequenced reads without mapping them to a reference genome such as recount [23]. Furthermore, several studies published recently [24], Novoalign (http://www.novocraft.com) and mosaik-aligner (http://code.google.com/p/mosaik-aligner/) improved the detection by integrating user defined SNPs. Our approach is different
because IR combines both SNPs and Indels to build a revised model, which is used to fix the mapping of misaligned reads, and allow researchers to use an updated reference as input model in future studies.

1.2 Method and results

We developed a new IR algorithm to improve the used pipeline for detecting variations in the sequenced data. It focused on building a new revised reference models that better represents the sequenced data and improved the alignment process. We tested the new algorithm on simulated data by using wgsim simulator, that can generates sequences data by using different rate of variations. Additionally, we tested this algorithm on four sets of real data. Three of the four examples are families located in north of Israel and affected in recessive disease. The last one is a Druzen family that affected in dominant disease. The results demonstrate that using the IR improves the alignment process by up to 6.8%, increases the detection rate of true variations by up to 3.5%, and decreases the rate of false variations up to 1.5%, in the simulated data. Also IR succeeded to detect novel variations in analyzing the real data.
Chapter 2
Mapping and Annotation Tools

The major focus of IR is to make the alignment tools, which are responsible for mapping short reads, like BWA and BOWTIE2 more efficient, because the alignment step is the slowest phase during the analysis of NGS data. Also, other tools and packages were used in developing and evaluating IR annotation tools, like ANNOVAR, and manipulating alignments in the SAM format, such as SAMTOOLS and GATK.

2.1 BWA

A software package that is used for mapping paired and single end reads against a reference model. It requires building index file for the model by using Burrows-Wheeler Transform (BWT) before running the package which is time consuming, but it is essential because it drastically reduces the required time for finding the accurate position for the reads in the model. BWA includes three mapping algorithms BWA-BACKTRACK, BWA-MEM and BWA-SW, where each one is designed for different sequence length. BWA-MEM, the latest developed algorithm, is more accurate and faster than the other algorithms. The implementation of these algorithms is based on dynamic programming, prefix trie and string matching, and support mismatches and gapes between the reads and reference model. Recently, many state-of-the-art frameworks that upgrade the analysis process of NGS data used BWA as the main alignment tool.

2.2 BOWTIE

This alignment tool is intended for mapping short sequenced reads against long reference genome. The Burrows-Wheeler Transform (BWT) is used also in
BOWTIE for indexing the reference genome to make the alignment process efficient. Two versions of BOWTIE were developed recently, BOWTIE-1 and the new version BOWTIE-2. BOWTIE-2 is more flexible and faster than the old version, consumes less memory and appropriate for long reads that generated by several sequencers such as Ion Torrent instruments, Roche 454 and Illumina HiSeq 2000. Additionally, It supports gapped alignment by performing two steps: First, splitting the read into smaller substrings called seed reads and aligning them against the reference by disabling gaps. Second, the aligned seed reads are extended by enabling gaps and performing dynamic algorithm until the whole read is aligned entirely.

2.3 SAMTOOLS

A software package that adjusts for analyzing aligned data, which is generated by mapping tools in SAM or BAM format. It provides a diverse set of services like discovering variations, sorting aligned data, indexing sorted files, merging multiple alignment files and converting SAM to BAM format. The main focus of SAMTOOLS is to simplify the analysis process of NGS data, and to make researchers able to process the data with less difficulties. WGSIM, which is provided in this package, is responsible for generating simulated sequenced reads for the whole genome. This utility provides several adjustable parameters, such as rate of mutations, fraction of indels, reads number, and length of reads.

2.4 GATK

A highly generic software package developed at the Broad Institute for handling data from any NGS platform. It includes a wide range of utilities, which are used for annotation, alignment and analyzing VCF data, and has a range of capabilities to discover unique variations. Furthermore, It improves the analysis process of the data by using computational infrastructure and engine. In contrast to other tools, GATK provides end-to-end methodology for processing the data, while other tools are focused on a specific part of the analysis pipeline.

2.5 ANNOVAR

A functional annotation software tool that is freely available at [http://www.openbioinformatics.org/annovar/](http://www.openbioinformatics.org/annovar/) It makes use of updated information to annotate chromosomal variations. The standard input is a text file, that includes SNPs, INDELs or CNVs. ANNOVAR makes use of F-SNP [11] and Seattle Seq, which is used to annotate novel variations, against other tools which are limited to known variations.
2.6 CoNIFER

The standard tools like SAMTOOLS or GATK are targeted for detecting short deletion and insertion, while CoNIFER main task is to detect copy number variation(CNV) or loss of heterozygosity(LOH). CoNIFER is a statistical method, implemented in python, used to handle WES data and different than other WGS tools [1,2,7] that assume continuity and normal distribution on the genome. It uses depth of coverage for each site in the alignment of the short reads (BAM files) to detect CNV or LOH.
Chapter 3

Iterative Referencing Algorithm

The key of the proposed iterative referencing algorithm is the process of mapping the raw sequence data against the current reference model and using the most appropriate aligned reads to revise the reference model. The revised model is used as the input reference model for the next iteration. The inputs of the algorithm are the raw sequence data, an initial reference model, the quality (QUAL) and coverage (DP) threshold, and the desired number of iterations. The number of iterations can also be set automatically. Normally less than 5 iterations are needed.

3.1 IR Schematic Description

Figure 3.1 provides a flow chart of the algorithm. The figure demonstrates that the first mapping of the raw data is executed using a given initial reference model. Then genotyping homozygous variations takes place in 4 steps as suggested by DePristo et al [4]. Then high quality homozygous variations are saved and a new model is generated.

Figure 3.2 illustrates the iterative adjustment of the reference model. In this example, the first iteration maps 4 reads and rejects the other 2, due to one deletion, one insertion and one mismatch wrt the initial reference model. In the next iteration the alignment tools map these two reads successfully to the revised reference mode. Consequently, new variations in the sequenced DNA are discovered. Note that the algorithm then updates the positions of the variations wrt the initial reference model.

The theoretical justification of algorithm IR is the approximation of the maximum likelihood of the data using the same ideas underlying the Expectation Maximization (EM) algorithm [3]. The maximization step (M-step) uses the current
Algorithm Iterative Referencing (IR)

**Input:** List of reads $R = \{r_1, r_2, \ldots, r_n\}$ (reads taken from one or many individuals)
- Reference Model $M_0$
- Number of iterations $I$ (default can be set dynamically by the algorithm)

**Output:**
- The position for each variation (INDELs or SNPs) wrt reference model $M_0$
- The most likely position for each read $r_j$ wrt reference model $M_0$
- A revised reference model $M_I$ that better represents the input reads

for $i = 1 \ldots I$ do

(M-step). Map all reads $r_j$ in $R$ using the model $M_{i-1}$ (via the BWA or BOWTIE2 software)
Save all reads $r_j$ from $R$ in $R^*$ except those mapped to multiple locations in $M_{i-1}$
Detect all homozygous variations using $R^*$ and $M_{i-1}$ (via GATK or SAMTOOLS software)
Compute the average of quality (QUAL) and coverage (DP) of all homozygous variations
Select homozygous variations with quality and coverage higher than the computed average
Save the selected homozygous variations in $F_i$

(E-step). Insert the variations in $F_i$ into $M_{i-1}$ to build a revised reference model $M_i$

end for

Detect all variations using $R^*$ and $M_I$ (via GATK or SAMTOOLS software)
Save the variations in $F_0$

for $i = 0 \ldots I$ do
- Adjust the position wrt $M_0$ of each variation in $F_i$

end for
genome reference as the true model and computes the most likely true position $x_j$ of each read $r_j$. Denoting by $\Theta$ the current reference model, the M-step maximizes the likelihood of the observed variables, the position $x_j$ of the uniquely mapped reads $r_j$, given the hidden parameters, namely the current reference model $\Theta$,

$$logL(Data) = \sum_{j=1}^{N} logP(x_j|\Theta)$$

assuming all reads are mapped independently of each other. The probability of a correct mapping $P(x_j|\Theta)$ is a function of the number of mismatches $m_j$ wrt $\Theta$ and the length $l_j$ of the read $r_j$. This probability is calculated by each mapping tool, as detailed in [10, 12].

After the current most likely positions of all reads are determined, a list of uniquely mapped homozygous variations are listed. The E-step computes the expected genomic reference model given these variations by replacing the current bases at positions of uniquely mapped homozygous variations with the alternative bases.

The step of inserting homozygous variations into the current reference model to build a revised model is critical for the algorithm’s accuracy. In particular, committing erroneously to false variations in any iteration will deteriorate the mapping quality of the next iteration. Therefore the threshold of including homozygous variations is determined dynamically in each iteration of the algorithm. IR calculates the average Quality and DP of all the homozygous variations observed in the current iteration and accepts as true variations only those that are supported by higher than average QUAL and higher than average DP. In the unlikely event that the average quality and coverage are lower than some default predetermined constants, which are based on the specifications of current measurement devises, new variations are introduced into the revised reference model only if they are also higher than the predetermined constants. In our implementation and experiments, the default thresholds for QUAL is 15 and for DP is 6, but these can be changed by the user.

For simplicity of our schematic description of IR the number of iterations has been fixed as input. However, the normal mode of operation of the implemented algorithm is that the dynamic criterion for stopping the iterations is that the difference between the number of uniquely mapped reads between two consecutive iterations drops below a certain threshold (of 0.3%).
Fig. 3.1: Flow Chart for algorithm IR
Fig. 3.2: Mapping in two iterations. Figure (a) shows 6 reads and the variations between the sequenced DNA and the initial reference model. Figure (b) demonstrates that in the first iteration, IR maps the first 4 reads to the reference genome and discovers a mismatch of G and a deletion of T. In this example, reads are assumed to be mapped successfully when at most one variation exists wrt the current reference model. Then, IR generates a revised reference model. Figure (c) shows that all 6 reads are mapped perfectly and that the last revised reference model, in this example, becomes identical to the sequenced DNA.
Chapter 4

Expectation Maximization in IR

In this chapter we demonstrate that the IR algorithm can be described via the EM method, a well-known method in the domain of statistical models [3]. Recall that statistical model is the set of probability distributions on the sample space [16]. A model assigns a probability distribution function \( P_\theta \) for each parameter point \( \theta \). In general statistical model methods used to resolve the problem of estimating unknown parameters vector \( \Theta \) based on viewing incomplete data from the sample space \( X \). In most cases there is no direct method to determine the value of the missing parameters, and several methods are used like maximum likelihood to evaluate them.

The likelihood function is a function that defines the relation between observed data and the unknown parameters. For observed data \( X = X_1, X_2, ..., X_n \) and parameter vector \( \Theta = \theta_1, \theta_2, ..., \theta_n \), the likelihood function is defined as:

\[
L(\theta_1, \theta_2, ..., \theta_n | X_1, X_2, ..., X_n) = P(X_1, X_2, ..., X_n | \theta_1, \theta_2, ..., \theta_n)
\]

Which means that the likelihood function is equal to the probability of the observed data given unknown parameters.

Maximum likelihood estimate (MLE) is a method used to evaluate the unknown parameters \( \Theta \) of incomplete model. MLE estimates the unknown parameter vector \( \Theta \), and maximizes the likelihood function \( P(X | \Theta) \) over the observed data. It is equivalent to maximize the log likelihood function rather than the regular one, because it is an increasing function and appears to be more convenient. In many cases the estimation of the MLE value is computationally hard and the EM method is used for this task.
EM is an iterative method used to estimate the MLE values for $\Theta$ in statistical models. The algorithm starts with an initial guess for $\Theta$ and alternates through the following steps: First, E-step which estimates the log likelihood function using the observed data and the current value of $\Theta$. Second, M-step which estimates new model parameters that maximize the log likelihood function.

### 4.1 EM in IR

It is useful to consider the alignment of the sequenced data problem as an approximation of the MLE, using the same ideas underlying the Expectation Maximization (EM) algorithm [3]. Let $x_j$ denote the uniquely mapped reads $r_j$, $\Theta$ is the current reference model, $N_1$ denotes the number of uniquely mapped reads and $N_2$ denotes the total number of unmapped and multiple aligned reads. Since we assume the reads are mapped independently of each other, then the log probability of the entire event is given by:

$$\log L(Data) = \sum_{j=1}^{N_1} \log P(x_j|\Theta) + \sum_{j=1}^{N_2} \log \pi =$$

$$\sum_{j=1}^{N_1} \log P(x_j|\Theta) + N_2 \log \pi$$

where $\pi$ stands for the probability of non uniquely mapped reads.

The maximization step (M-step) uses the current genome reference as the true model and computes the most likely true position $x_j$ of each read $r_j$. Altering the position of the detected homozygous variations, and generating the new reference model $\Theta_1$ is maximizing the log likelihood function because $P(x_j|\Theta)$ will be a function of decreased number of mismatches and indels.

The essential benefit of the EM algorithm is that the likelihood function is not decreasing after each step. The probability for mapping each read $P(x_j|\Theta)$ depends on several parameters: the matching score between the two bases in the reference model and the short read, the penalty for mismatching score and the deletion/insertion score for each read. These parameters are set to default values by the mapping tools and could be changed by the users.

Table 4.1 demonstrates that IR_BWA and IR_BOWTIE2 increases the logarithmic likelihood function after each iteration, and the main cause for that is related to more reads are uniquely mapped to single position in the reference genome. Additionally, it should be noticed that the first iteration for both tools increases
\[ \log L(\text{Data}) \times 10^{-9} \]

<table>
<thead>
<tr>
<th></th>
<th>BOWTIE2</th>
<th></th>
<th>BWA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>1</td>
<td>-31.7</td>
<td>-55.7</td>
<td>-721</td>
<td>-758</td>
<td>-4880</td>
</tr>
<tr>
<td>2</td>
<td>-29.1</td>
<td>-53.6</td>
<td>-447</td>
<td>-694</td>
<td>-3910</td>
</tr>
<tr>
<td>3</td>
<td>-28.7</td>
<td>-52.8</td>
<td>-423</td>
<td>-673</td>
<td>-3750</td>
</tr>
<tr>
<td>4</td>
<td>-28.6</td>
<td>-51.9</td>
<td>-412</td>
<td>-661</td>
<td>-3720</td>
</tr>
</tbody>
</table>

Tab. 4.1: Results based on simulating seventy million 100bp reads from the Human genome, with induced variation rate 0.05 for different rates of INDELs. Each row shows the log likelihood of the seventy million reads for each IR iteration. The columns reflect increasing percentage of INDELs within the 5% variations used to simulate the reads.

the logarithmic likelihood function drastically, and the improvement is smaller for the later iterations. Therefore, it is always recommended to run IR for at least one iteration.
Chapter 5

Results

We tested the Algorithm on simulated and real data. In the simulated data, we generated several data sets that each represents a different variation rate and length of the sequenced data. While we used the sequenced data of three different families in Israel to test the algorithm performance on real data.

5.1 Results on Simulated Data

The algorithm IR uses in the M-step of each iteration one of the standard tools for mapping: BWA and BOWTIE2. We test the performance of IR wrt each of these software packages using simulated and real data. There are several available whole genome simulators to generate NGS reads such as WGSIM from SAMTOOLS[13], MAQ[14] and ART[6]. We used WGSIM because it also simulates a quality score of the measured bases, often referred to in the literature as dummy values. We simulated a baseline data set of seventy million 100 bp reads from the human genome with variations rate 0.05 of which 20% are INDELs. The read size was chosen to match the output of the Illumina Hiseq 2000 sequencer which generates 100 bp reads and was used in our real data experiments. The simulation was repeated for different variation rates. We denote by IR_BWA the iterative algorithm in which each iteration uses BWA, and by IR_BOWTIE2 the iterative algorithm in which each iteration uses BOWTIE2. We ran these four programs on the simulated data sets to generate reports of INDELs and SNPs. In each experiment the input consisted of the reads of one simulated genome, with an average coverage of approximately 5.

All our results show a clear cut dominance of the proposed iterative algorithms in all accuracy aspects. Figure 5.1 shows that the percentage of reported true SNPs from the overall SNPs detected by IR_BWA and IR_BOWTIE2 is increased
<table>
<thead>
<tr>
<th>Iteration</th>
<th>BOWTIE2 10%</th>
<th>BOWTIE2 20%</th>
<th>BOWTIE2 30%</th>
<th>BWA 10%</th>
<th>BWA 20%</th>
<th>BWA 30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85.15</td>
<td>85.18</td>
<td>85.16</td>
<td>72.80</td>
<td>69.52</td>
<td>65.52</td>
</tr>
<tr>
<td>2</td>
<td>88.75</td>
<td>88.38</td>
<td>87.89</td>
<td>77.34</td>
<td>73.36</td>
<td>68.62</td>
</tr>
<tr>
<td>3</td>
<td>89.13</td>
<td>88.74</td>
<td>88.22</td>
<td>78.01</td>
<td>73.91</td>
<td>69.05</td>
</tr>
<tr>
<td>4</td>
<td>89.18</td>
<td>88.8</td>
<td>88.28</td>
<td>78.13</td>
<td>74.21</td>
<td>69.34</td>
</tr>
</tbody>
</table>

Tab. 5.1: Results based on simulating seventy million 100 bp reads from the Human genome with average coverage of 4.82 folds, with induced variation rate 0.05, for different rates of INDELs. Each row shows the percentage of uniquely mapped reads for each IR iteration. The columns reflect increasing percentage of INDELs within the 5% variations used to simulate the reads.

wrt BWA and BOWTIE2, respectively. Higher quality is awarded by the iterative algorithms to true variations. Despite the vast improvement in detecting true SNPs with far higher confidence, the rate of false positive SNPs is still improved as well. Figure 5.2 demonstrates that IR_BWA and IR_BOWTIE2 report less false SNPs than BWA and BOWTIE2, respectively. Furthermore, IR removed all the false SNPs with low quality, and reduced the quality awarded to remaining false SNPs. The same results regarding the identification of true and false INDELs is shown in Figure 5.3. In summary, the iterative referencing algorithm, within less than 5 iterations, improves the precision and recall of both state-of-the-art programs wrt SNP and INDELs detection.

Another important criterion in mapping quality is the number of uniquely mapped reads because normally these are the reads used to detect variations, while reads that are mapped to several loci are often not taken into account. Table 5.1 demonstrates that IR_BWA and IR_BOWTIE2 increase the number of uniquely mapped reads, because after each iteration more multiple aligned reads and unmapped reads are uniquely mapped to a single position in the new generated reference model. The table shows that the improvement becomes more prominent as variation rates grow higher and that IR increases the number of the uniquely mapped reads after each iteration. Already after one iteration a significant improvement is obtained but the accuracy of mapping reads improves for several more iterations.

In all experiments above we selected 5% variations as a baseline. This selection was done because in mapping projects when there is a condensed area with 5% variations, existing methods fail, while our iterative mapping is still able to perform well, generating 3.5% more variations with higher quality. When the
Fig. 5.1: Number of true SNPs reported after 4 iterations as compared to BWA and BOWTIE2, with variation rate of 0.05. The total number of reported true SNPs increased by 3.5% variation rates drops to 0.0015, a rate acceptable as the average genome variation, the improvement of IR drops to approximately 0.1% but only one update of the reference model is needed. Also for very high variation rates such as 0.1, which are not often seen in real data, the improvement of IR is still noticeable. Figure 5.4 provides the improvement spectrum of IR as a function of the variation rate, and it can be noticed that variation rate of 0.06 provides the maximum improvement (6.8%).

In a final experiment we showed how to reduce the run time of IR using a time saving heuristic. The heuristic is to simply remove uniquely mapped reads from further iterations. The experiment shows a reduction of 77% in the total run time of this heuristic IR algorithm when two iterations are used on the simulated data. For three iterations the reduction is 48%, and for five iterations the reduction of the total run time is 33%. The potential problem with this heuristic IR algorithm is that a variation may be erroneously determined at an early iteration and as a result not being considered in later iterations while the IR algorithm, as described
Fig. 5.2: Number of false SNPs reported after 4 iterations as compared to BWA and BOWTIE2, with variation rate of 0.05. The total number of reported false SNPs decreased by 1.5%

in Figure 3.1 may overcome such local maximum phenomena. Despite our theoretical concern regarding accuracy, the results on simulated data show that the rate of false variations is as low as reported by IR. Furthermore, the heuristic IR algorithm still moderately improves the reported true variations versus non iterative algorithms but finds 66% less true additional variations compared to IR. Therefore, when run time becomes critical due to number of reads, one may consider to use this time saving heuristic as a first step of analysis.

5.2 Results on Real Data

An important area for which the iterative referencing algorithm has a major impact is the study of rare diseases in isolated populations. In such genetics studies the number of patients is often small and the reference model does not necessarily reflect well the common variations in the studied population. Furthermore, the DNA measurements are quickly moving towards NGS data posing the problems of recall and precision of variations in the collected data.
In this section we provide three real examples for which IR found new variations that escaped discovery by the standard mapping algorithms. The data was provided by three medical research groups that have not found a mutation via the current tools. Hence there is a higher probability that new mutations identified by the iterative algorithm will include the causative mutations. These genetics studies are still on going, sifting the newly identified mutations. In these three examples, all individuals were genotyped by Whole Exome Sequencing (WES) using the Illumina Hiseq 2000 sequencer which generates approximately 100bp reads. In the three examples the baseline method that was used to map the WES reads against the whole genome was the BWA software and the reference model used was HG19 (build37). In each study, the input reads for IR have been the combined pool of measured reads of all genotyped pedigree members.

A first example is an Ashkenazi Jewish nuclear family with 3 children from mid Israel segregating a familial recessive version of Angelman syndrome, not caused by known mutations. Three affected siblings and both parents have been
Fig. 5.4: The improvement spectrum of uniquely mapped reads as a function of the variation rate. This graph shows that the improvement of IR is maximal (6.8%) when the variation rate is between 0.045 and 0.07.

genotyped by WES. The parents are assumed healthy carriers. The total number of reads for each individual was 161 millions pairs of reads. The average number of reads that were mapped perfectly against a unique locus in the reference model by BWA, with default parameters, was 155.8 millions (96.77%). All averages reported for this example are taken over the data of the 5 sequenced siblings. Reads that are mapped uniquely and perfectly were mapped to the same locus by every mapping method. The difference between methods is in mapping the remaining reads. In our example there were 5.2 million such remaining reads. Only 2.72 million of these reads were mapped to the reference genome of which 1.91 million were mapped uniquely (but not perfectly), and 812457 were not found a unique position. Our iterative referencing algorithm aims to decrease further the set of 812457 reads that are not mapped uniquely and by that, as we later show, find variations that remained undetected. Table 5.2 shows that after one update by IR of the reference model, on average, additional 88556.8 reads among the 812457 (10.8%) were mapped uniquely to the initial reference genome and by the third update a total of additional 172389.4 reads (21.2%) were mapped uniquely.

These newly mapped reads allow the identification of new variations in the dataset. Since the trait is assumed recessive and the population is isolated we
concentrated on sizable homozygous areas. We examined homozygous areas on the genome of length larger than 1Mb. The number of sporadic heterozygous loci in the 1Mb homozygous stretches dropped from 132 to 76. This drop should be attributed practically fully to a noticeable drop in mapping errors. Furthermore, BWA found 1074 homozygous variations in the homozygous areas while IR found 1252 such variations. This increment stems from two sources: 48 heterozygous variations became homozygous, and the remaining variations are due to increased coverage of the 1Mb homozygous stretches. The length of the covered stretches increased by 5% indicating that the coverage increased likewise.

To better fine tune our search for a causative variation, we used the program SAMTOOLS to generate a list of SNPs genome-wise from the output of BWA and IR_BWA and used the linkage program SUPERLINK ONLINE to compute linkage between the disease gene and loci on the genome in order to list suspicious ranges. Figure 5.5 shows how the output of IR_BWA produced much smoother LOD scores than that produced when using the output of BWA, attributed to the fact that less erroneous SNPs are produced as input to the linkage program. By using the linkage program on the outputs of IR_BWA, allowing only LOD scores above 1, we reduced the number of unknown, non-synonymous, within exons, variations from 640 new variations to just 14 new variations. These 14 variations are promising targets for investigation that escaped analysis by standard usage of BWA and revealed by the iterative algorithm IR_BWA.

Finally, we now demonstrate that IR also provides better quality scores. As stated, the number of sporadic heterozygous loci in the 1Mb homozygous stretches dropped from 132 to 76. The average quality score of the remaining heterozygous loci dropped significantly from an average coverage (DP) of 8.32 to 6.67, identifying these loci as more likely to be mismapped. Similarly the 1074 homozygous variations in the homozygous stretches received an average score of 68.12 when IR_BWA is used versus only an average score of 52.43 when BWA is used.

A second example is a large consanguineous Muslim family in North Israel segregating a highly penetrating familial recessive version of Atypical Hemolytic-Uremic (AHUS) syndrome, not caused by known mutations. The pedigree consists of a double first cousin marriages, as shown in Figure 5.6. Each couple has one affected child, and individuals 7,10,11 and 12 have been genotyped by WES. A LOD score analysis using the output of IR_BWA revealed two homozygous area of LOD score higher than 2 which were the only suspect regions found. The total length of these regions was 13 million bp. IR_BWA succeeded to map additional 1.92 million reads for the affected son and 315671 reads for the parents wrt the reference genome. The main reason for this difference is the existence of many
New uniquely mapped reads found by IR_BWA

<table>
<thead>
<tr>
<th>Sibling</th>
<th>IR_BWA(1)</th>
<th>IR_BWA(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81245</td>
<td>171458</td>
</tr>
<tr>
<td>2</td>
<td>85412</td>
<td>170125</td>
</tr>
<tr>
<td>3</td>
<td>96254</td>
<td>168524</td>
</tr>
<tr>
<td>4</td>
<td>90124</td>
<td>174258</td>
</tr>
<tr>
<td>5</td>
<td>89749</td>
<td>177582</td>
</tr>
<tr>
<td>Average</td>
<td>88556.8</td>
<td>172389.4</td>
</tr>
</tbody>
</table>

Tab. 5.2: On average, after one update by IR_BWA of the reference model additional 88556.8 reads among the 812457 (10.8%) unresolved reads were mapped uniquely to the initial reference genome and by the third update a total of 172389.4 reads (21.2%) of the unresolved reads were mapped uniquely.

![Lod scores of the sequenced family using the outputs of BWA and IR_BWA](image)

Fig. 5.5: Lod scores of the sequenced family using the outputs of BWA and IR_BWA, respectively.
Fig. 5.6: Pedigree segregating a recessive trait. Individuals 7,10,11 and 12 have been sequenced.

homozygous regions in the affected son and IR_BWA utilize it to map more reads. The number of sporadic heterozygous loci in these homozygous stretches in the affected individuals dropped from 312 to 278. This drop is due to a noticeable drop in mapping errors. Furthermore, BWA found 1622 homozygous variations in the homozygous areas while IR_BWA found 1798 such variations. This increment stems from two sources: heterozygous variations became homozygous, and the remaining variations are due to increased coverage of these homozygous stretches. The length of the covered stretches increased by 2.1% indicating that the coverage increased likewise. We reduced the number of unknown, non-synonymous, within exoms, variations from 11 new variations to just 4 new variations. These 4 variations are promising targets for investigation that escaped analysis by standard usage of BWA and revealed by the iterative algorithm IR_BWA. One of these variations turned out to be the causative one, and is currently under functional analysis.

The third example is a large consanguineous Druze family in North Israel segregating a moderately penetrating familial dominant version of Acute Myeloid Leukemia (AML), not caused by known mutations. This family consists of six
affected individuals as shown in Figure 5.7. The assumption behind the gene search in this project is that a single gene is causing this familial disease and that a living person can carry only a single copy of the mutation because a double carrier cannot survive pregnancy. Individuals 14, 16, 46, 44, 12 and 13 have been genotyped by WES. A LOD score analysis using the SNP array data revealed one heterozygous area of the maximal attainable LOD score of 1.25 which was the only suspect region as can be seen in figure 5.8. The total length of these regions was 15 million bp. IR\textsubscript{BWA} increased the average number of heterozygous variations in the affected individuals in the suspected region by 103 (3.5%) after running IR for four iterations, and succeeded to map an additional 48.2 thousand reads on average to the reference genome. We observed that the addition of uniquely mapped homozygous stretches found by IR\textsubscript{BWA} also improve the recall and precision of the heterozygous variations and allowed us to identify 11 unknown, non-synonymous, within exons, new candidate variations for AML. IR succeeded to detect the suspected region that is shared for all affected individuals as shown in figure 5.8 and to reveal new suspected variations in chromosomes 17 that were verified by using other 100 controls from the same town in Israel.

Additionally, researchers suspected that the main cause of the disease is related to CNV [21] or LOH [8]. Therefore, CoNIFER [19] was used to analyze the aligned short reads. CoNIFER succeeded to reveal new suspicious regions by using the BAM files from the IR package. Figure 5.9 shows a specific range (around 61.8 million bp) with different frequencies between the affected and the healthy individuals.

The final example is related to three consanguinity Arab families from the same village in Israel, segregating a version of Congenital Chloride Diarrhea (CCD), which immediately causes loss of body water after birth. These three families comprise three affected children and healthy carriers parents. We used segregation analysis to determine the mode of inheritance (MOI) for the disease. As shown in Figure 5.10 the recessive disease mode is drastically more fitting than the dominant mode for the combined pedigrees.

The researcher team used the microsatellite markers technology to sample the DNA of nine samples (three affected children and their parents), SUPERLINK online was used to compute the linkage between the disease gene and loci on the chromosomes. The Lod score was computed by summing the LOD score of the three separated families, and the results show one range on chromosomes 2 with LOD scores above 3 between 36-40 million base pairs. Next, Illumina HiSeq 2000 was used for sequencing two affected children and one carrier parent. Only three individuals were sampled due to budget considerations. The researchers did
Fig. 5.7: Pedigree segregating a dominant trait. Six individuals have been sequenced.
Fig. 5.8: Lod score for chromosome 17. Only one suspected region was detected with LOD score above 1.

Fig. 5.9: CoNIFER output. It shows different frequencies for the affected individuals in position 61.8 million bp.
Fig. 5.10: Each entry in the table is the $\log_{10}$ likelihood of the data (for all analyzed pedigrees together and for each pedigree alone) for a specific Mode Of Inheritance (MOI) and specific penetrance probabilities. When comparing two lines, one is checking the likelihood of a dominant MOI versus a recessive MOI for a specific penetrance value. The highest value, namely when the negative number is closest to zero, is the more likely MOI.

<table>
<thead>
<tr>
<th>penetrance \ population prevalence</th>
<th>0.45</th>
<th>0.4</th>
<th>0.3</th>
<th>0.25</th>
<th>0.2</th>
<th>0.15</th>
<th>0.1</th>
<th>0.05</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>dominant 0 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>recessive 0 0 1</td>
<td>-6.4</td>
<td>-6.1</td>
<td>-5.5</td>
<td>-5.3</td>
<td>-5.1</td>
<td>-4.9</td>
<td>-4.9</td>
<td>-5.1</td>
<td>-6.1</td>
<td>-6.6</td>
<td>-7.9</td>
</tr>
<tr>
<td>dominant 0.99 0.99</td>
<td>-15.3</td>
<td>-15.2</td>
<td>-15.1</td>
<td>-15.1</td>
<td>-15.2</td>
<td>-15.4</td>
<td>-15.7</td>
<td>-16.4</td>
<td>-18.4</td>
<td>-19.3</td>
<td>-21.4</td>
</tr>
<tr>
<td>recessive 0.0 0.99</td>
<td>-6.4</td>
<td>-6.1</td>
<td>-5.5</td>
<td>-5.3</td>
<td>-5.1</td>
<td>-4.9</td>
<td>-4.9</td>
<td>-5.1</td>
<td>-6.1</td>
<td>-6.6</td>
<td>-7.9</td>
</tr>
<tr>
<td>dominant 0.95 0.95</td>
<td>-11.1</td>
<td>-11.0</td>
<td>-10.9</td>
<td>-10.9</td>
<td>-11.0</td>
<td>-11.2</td>
<td>-11.5</td>
<td>-12.2</td>
<td>-14.2</td>
<td>-15.1</td>
<td>-17.2</td>
</tr>
<tr>
<td>recessive 0 0 0.95</td>
<td>-6.4</td>
<td>-6.0</td>
<td>-5.5</td>
<td>-5.2</td>
<td>-5.0</td>
<td>-4.9</td>
<td>-4.9</td>
<td>-5.1</td>
<td>-6.1</td>
<td>-6.6</td>
<td>-7.9</td>
</tr>
<tr>
<td>dominant 0.0 0.9</td>
<td>-9.4</td>
<td>-9.2</td>
<td>-9.1</td>
<td>-9.1</td>
<td>-9.2</td>
<td>-9.4</td>
<td>-9.7</td>
<td>-10.4</td>
<td>-12.4</td>
<td>-13.3</td>
<td>-15.4</td>
</tr>
<tr>
<td>recessive 0 0 0.9</td>
<td>-6.3</td>
<td>-6.0</td>
<td>-5.4</td>
<td>-5.2</td>
<td>-5.0</td>
<td>-4.9</td>
<td>-4.8</td>
<td>-5.0</td>
<td>-6.1</td>
<td>-6.6</td>
<td>-7.9</td>
</tr>
<tr>
<td>dominant 0.8 0.8</td>
<td>-7.7</td>
<td>-7.5</td>
<td>-7.3</td>
<td>-7.3</td>
<td>-7.4</td>
<td>-7.5</td>
<td>-7.9</td>
<td>-8.6</td>
<td>-10.6</td>
<td>-11.4</td>
<td>-13.5</td>
</tr>
<tr>
<td>recessive 0 0 0.8</td>
<td>-6.2</td>
<td>-5.9</td>
<td>-5.3</td>
<td>-5.1</td>
<td>-4.9</td>
<td>-4.8</td>
<td>-4.8</td>
<td>-5.0</td>
<td>-6.1</td>
<td>-6.6</td>
<td>-7.9</td>
</tr>
<tr>
<td>dominant 0.7 0.7</td>
<td>-6.8</td>
<td>-6.6</td>
<td>-6.3</td>
<td>-6.3</td>
<td>-6.3</td>
<td>-6.5</td>
<td>-6.8</td>
<td>-7.5</td>
<td>-9.5</td>
<td>-10.4</td>
<td>-12.4</td>
</tr>
<tr>
<td>recessive 0 0 0.7</td>
<td>-6.1</td>
<td>-5.7</td>
<td>-5.1</td>
<td>-4.9</td>
<td>-4.8</td>
<td>-4.7</td>
<td>-4.7</td>
<td>-4.9</td>
<td>-6.1</td>
<td>-6.7</td>
<td>-8.0</td>
</tr>
<tr>
<td>dominant 0.6 0.6</td>
<td>-6.2</td>
<td>-5.9</td>
<td>-5.6</td>
<td>-5.6</td>
<td>-5.6</td>
<td>-5.7</td>
<td>-6.0</td>
<td>-6.7</td>
<td>-8.7</td>
<td>-9.6</td>
<td>-11.7</td>
</tr>
<tr>
<td>recessive 0 0 0.6</td>
<td>-6.0</td>
<td>-5.6</td>
<td>-5.0</td>
<td>-4.8</td>
<td>-4.6</td>
<td>-4.5</td>
<td>-4.6</td>
<td>-4.9</td>
<td>-6.1</td>
<td>-6.7</td>
<td>-8.0</td>
</tr>
<tr>
<td>dominant 0.5 0.5</td>
<td>-5.8</td>
<td>-5.5</td>
<td>-5.1</td>
<td>-5.0</td>
<td>-5.0</td>
<td>-5.1</td>
<td>-5.4</td>
<td>-6.1</td>
<td>-8.1</td>
<td>-9.0</td>
<td>-11.0</td>
</tr>
<tr>
<td>recessive 0 0 0.5</td>
<td>-5.8</td>
<td>-5.4</td>
<td>-4.8</td>
<td>-4.6</td>
<td>-4.5</td>
<td>-4.4</td>
<td>-4.5</td>
<td>-4.8</td>
<td>-6.1</td>
<td>-6.7</td>
<td>-8.1</td>
</tr>
</tbody>
</table>

not succeed to find any variation that might be linked to the disease.

Failing to find suspicious variations led the researchers to use SNP array technology to compute the linkage. Again. For the second time only one suspected range was detected in the whole genome in chromosome 2 as demonstrated in figures 5.12, but the range was extended to be from 36 to 56 million base pairs. Furthermore, IR was used to discover the variations in the new suspected range and succeeded to find a new novel variation in the new range. This variation was not detected by the traditional tools and the researchers started to test it in 100 controls from the same village.
Fig. 5.11: Lod score for the first 12 chromosomes.

(a) Chromosome 1  
(b) Chromosome 2  
(c) Chromosome 3  
(d) Chromosome 4  
(e) Chromosome 5  
(f) Chromosome 6  
(g) Chromosome 7  
(h) Chromosome 8  
(i) Chromosome 9  
(j) Chromosome 10  
(k) Chromosome 11  
(l) Chromosome 12
Fig. 5.12: Lod score for the last 10 chromosomes from the whole genome. It was computed by summing the LOD scores of the three separated families. It is obvious from the figures below that only one range in chromosome 2 is suspicious to be linked to the CDD disease.

(a) Chromosome 13  
(b) Chromosome 14  
(c) Chromosome 15  
(d) Chromosome 16  
(e) Chromosome 17  
(f) Chromosome 18  
(g) Chromosome 19  
(h) Chromosome 20  
(i) Chromosome 21  
(j) Chromosome 22
Chapter 6
Conclusions, discussion, and directions for further research

We presented an Iterative Referencing (IR) algorithm that increases the accuracy of mapping sequence data by iteratively improving the reference genome to fit the studied population. The algorithm IR can work hand in hand with any mapping tools as BWA and BOWTIE2 as its mapping procedure within each iteration. The novelty is the idea that each iteration improves the reference model and allows improved mapping in the next iteration, until convergence. The results demonstrate that IR improves the alignment process, increases the rate of true variations, and decreases the rate of false variations, all measured with respect to the original reference genome.

About 5 iterations are normally needed for convergence which means the running time is 5 times higher. However, in these studies the primary concern is the accuracy of the results which we improved rather than the running time. Nevertheless, a lot of effort has been invested recently to reduce the run time without compromising accuracy by parallelizing existing algorithms such as PBWA that speed up BWA [18] or the parallel alignment tool SOAP3 [15] that is used for short reads. Our iterative Referencing algorithm with PBWA in each iteration will provide a running time that is better than BWA.

We demonstrated IR on simulated human genomes and on real measured genomes for rare Mendelian diseases and showed the practical benefit achievable. An interesting potential usage of the iterative approach is when mapping new species with no reference model. Instead of overcoming de novo assembly difficulties, one may choose a related mapped organism as an initial reference model and allow the algorithm to converge to a better model. Our experiments show that IR
succeeded to improve by 6.8% the mapping of sequenced DNA with variation rate 0.06, which may suffice in some DNA mapping projects.
Bibliography


אלגוריתמים הפניה נשנים
ליניות נתונים ריצוע
גנצי

עלאת גנאים
אלגוריתמים הפניה ושניית ליניותaptors

גבתי

תיבמור על מחקר

לשם מדליית חלקי של הדירישות לקבלי התותאה
מגייסר למдесяים במועד ייחודי המחבר

עלאת גבאיים

מגשים לסנט הטכניון — מכון טכנולוגי לישראל
اهرة תשע”ד — חיפה

פברואר 2014
המחק נעשה בהנחייה פרופ' דנ גייגר bppקולטה למדעי המחשב

אני מודה לאספנים על התמיכת הכספית הנדרשת בהשטלה זו.
In the search, multiplicity of truth, sequence, is required in order to allow the researchers mapping to position against but these are the reality of the accuracy BWA, within skipping those found machines is mapping the reality to the round on the genome he is mapping the accuracy of the bases on the sequence) is important in the genome, in the genome mutations and the use of CROSSBOW he measured to increase the errors in the process works of him who implements this is the commonly used VARSIFTER in the algorithm values of mutations of NGS in the genome and this is the source of the improvement and is provided in the database with the annotated GATK and SAMTOOLS (and all) with the mutation detection and the genome.

VARSIFTER. Out, the tool for the mutation tool for the_bases is BANNOVAR and the genome is annotated.

The match against the toolkits for the results and the results in the genome or more genome, this is achieved in the sequence and the results achieved. Shown below the flowchart is the result of the flow of the information in the base sequence. In the genome, the reality is mapped to the sequence of the calculation and the correction of the accuracy of the round on the genome he is mapping the reality of the accuracy of the bases on the sequence) is important in the genome, in the genome mutations and the use of CROSSBOW he measured to increase the errors in the process works of him who implements this is the commonly used VARSIFTER in the algorithm values of mutations of NGS in the genome and this is the source of the improvement and is provided in the database with the database (and all) with the mutation detection and the genome.

CROSSBOW.

The tool for the mutation tool for the bases is BANNOVAR and the genome is annotated.

Philip in the mutation was achieved in the sequence, and the results achieved. Shown below the flowchart is the result of the flow of the information in the base sequence. In the genome, the reality is mapped to the sequence of the calculation and the correction of the accuracy of the round on the genome he is mapping the reality of the accuracy of the bases on the sequence) is important in the genome, in the genome mutations and the use of CROSSBOW he measured to increase the errors in the process works of him who implements this is the commonly used VARSIFTER in the algorithm values of mutations of NGS in the genome and this is the source of the improvement and is provided in the database with the database (and all) with the mutation detection and the genome.

Technion - Computer Science Department - M.Sc. Thesis MSC-2014-09 - 2014
In the beginning, the differences among the GDNs available for use in addition to the direct measurement of IR in the population enable the analysis of new GDNs. The analysis will reach and improve the runtime of the algorithm:

Algorithm IR—Description:

Input: \( \{r_1, r_2, ..., r_n\} \) (a set of GDNs extracted from all reads)

Output: GDN

Method:

1. Sort the GDNs in ascending order of the measured (Qual) and depth (DP) values for each position and the corresponding GDNs.
2. Select the top positions with the highest Qual and DP values.
3. For each selected position, find the GDN with the highest Qual and DP values.
4. For each selected position, find the GDN with the highest Qual and DP values.
5. Repeat steps 2-4 until all positions are selected.

The algorithm runs in a dynamic fashion, ending when the Qual and DP values reach a certain threshold, and continues to improve the runtime of the algorithm:

\[ M_{k-1} \cup R_u \]

Output: GDN

Method:

1. Select the top positions with the highest Qual and DP values.
2. For each selected position, find the GDN with the highest Qual and DP values.
3. For each selected position, find the GDN with the highest Qual and DP values.
4. Repeat steps 2-3 until all positions are selected.

The algorithm runs in a dynamic fashion, ending when the Qual and DP values reach a certain threshold, and continues to improve the runtime of the algorithm:

\[ M_i \cup R_u \]

Output: GDN

Method:

1. Select the top positions with the highest Qual and DP values.
2. For each selected position, find the GDN with the highest Qual and DP values.
3. For each selected position, find the GDN with the highest Qual and DP values.
4. Repeat steps 2-3 until all positions are selected.

The algorithm runs in a dynamic fashion, ending when the Qual and DP values reach a certain threshold, and continues to improve the runtime of the algorithm:

\[ M_i \cup R_u \]

Output: GDN

Method:

1. Select the top positions with the highest Qual and DP values.
2. For each selected position, find the GDN with the highest Qual and DP values.
3. For each selected position, find the GDN with the highest Qual and DP values.
4. Repeat steps 2-3 until all positions are selected.

The algorithm runs in a dynamic fashion, ending when the Qual and DP values reach a certain threshold, and continues to improve the runtime of the algorithm:
אלגוריתם IR הביקע על סדרות נתונים באימוג ימי סימולטור "wgsim" לייצר רצףを持って בישוע המוטציות של התווך התווך והשכלה הפנימית, מציאת campaigned גודל רצפים בברק ולהיבדיל גודל הסוכנות של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרцファン בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, كافة המוטציות של הרצפים בשיעור של נוביטים עד נוביטים, כל המוטציות של הרצפים לשיעור של נוביטים עד נוביטים, בනוסף, האלגוריתם הביא על סדרות של תואם אמיצים. שלושה מטרות לארבע שדרות התואם נוקמים. מ cenaו האלגוריתם בברק להיבדיל גודל הסוכנות של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך התווך בשיעור של 0.001-0.002 ו- 0.003 שבעים מילון רצפים עם 0.05 שיעור מוטציות של 100 רצפים והשכלה הפנימית של 100 בסיסים. השוואת האלגוריתם IR שלת בו BOWTIE 1 BWA. האלגוריתם שלת את התווך התווך בשיעור של נוביטים, כל המוטציות של הרצפים של התווך ושתווך בשיעור של נוביטים עד נוביטים, בברק, המוטציות של אמיצים בברק, ארבע רצפים של תואם אמיצים. שלושה מטרות לארבע שדרות התואם נוקמים.