7-1-2013

Next-Generation Sequencing: Acquisition, Analysis, and Assembly

Antoine Ho

Antoine Ho

Candidate

Molecular Genetics and Microbiology

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Jeremy S. Edwards , Chairperson

Susan R. Atlas

David Peabody

Margaret Werner-Washburne

NEXT-GENERATION SEQUENCING:
ACQUISITION, ANALYSIS, AND ASSEMBLY

BY

Antoine Ho

B.S. Genetics, University of California, Davis, 2005

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biomedical Science

The University of New Mexico
Albuquerque, New Mexico

July, 2013

DEDICATION

I begin this dedication by thanking my parents who have supported me throughout my life, and sacrificed very much to make sure that my life would be better than their own lives. My parents who buy and mail vitamins to me. My parents who always ask me when they see me, "have you been eating enough?" My parents who still text me to make sure that I know that UNM is having a snow day. Thanks Mom and Dad, for always catching my fly ball.

To Erin, I know that I put you through a lot of turmoil and you were consistently strong, supportive, and encouraging throughout my studies even when you had your own graduate career to worry about. We conquered many things together, from cramming for exams to preparing last minute presentations. Most importantly, you were always there for me, and I know how lucky I was to have your strength supporting me. Thank you for everything, I owe you so much.

Alex, graduate school is finally over for both of us and it's time for us to see what's out there. I'm grateful to have you in my life, and it makes me happy to know that we'll be together. Let's go have an adventure.

ACKNOWLEDGMENTS

I would like to thank Maurice Murphy and Susan Wilson for helping me learn computational aspects of research. Not only were your contributions necessary for my publications, but you were also kind and patient when working with a snot-nosed graduate student brat.

Many thanks to my graduate committee who guided me and made a respectable scientist out of me. Thank you to Dr. Peabody, the easiest man to schedule events with. Additionally, I always respected your blunt and honest manner of talking, to tell me what I needed to hear. Thank you to Dr. Werner-Washburne, who exudes such an intense enthusiasm and love of learning that I found myself infected by this ideal.

To my co-advisor, Dr. Atlas, thank you. Your many efforts, dripping with variously colored pens, to make sure I could talk the talk and write the writing were painful, but invaluable. I also admire your passion for mentoring, and how you always seemed to genuinely care about my development as both a scientist and a person.

And finally to my advisor, Dr. Edwards, thank you for taking me on as a graduate student. I am very thankful for the opportunity to have studied in your lab. When we first started, you told me that my graduate career in your lab would be different from many other graduate careers, and I didn't understand what you meant until it was over. You instilled within me an appreciation for tackling interesting and difficult problems, and the triumphs of being able to conquer them.

NEXT-GENERATION SEQUENCING:
ACQUISTION, ANALYSIS, AND ASSEMBLY

by

Antoine Ho

B.S. Genetics, University of California, Davis, 2005
PhD, Biomedical Science, University of New Mexico, 2013

ABSTRACT

The process of sequencing a genome involves many steps, and accordingly, this project contains work from each of those steps. Genome sequencing begins with acquisition of sequence data, therefore, a novel biochemistry was utilized and optimized for the Sequencing By Ligation (SBL) process. A cyclic SBL protocol was created that could be utilized to extend sequencing reads in both the 5' and 3' directions, for an increase in read length and thru-put.

After sequence acquisition, there is the process of data analysis, and the focus shifted to creating software that could take sequence information and match up the individual reads to a reference genome with greater speed and efficiency than other commonly-used software. The Sequence Analysis Workbench Tool, SAWTooth, was written and shown to outperform contemporaries NOVOAlign and BOWTIE.

Finally, the last aspect of genome sequencing is *de novo* assembly, prompting a comparative analysis of three assemblers: CLC Genomics Workbench, Velvet Assembler, and MIRA. Results were generated using Mauve to assess the general effects of different sequencing platforms on the final assembly.

TABLE OF CONTENTS

Chapter 1

Introduction

Next-generation sequencing is broad in both its potential applications as well as the component parts, requiring a wide expanse of expertise and background to accomplish. Genome sequencing is a versatile tool, and its most fundamental use is to simply gain the sequence information of genes and genome, though even such a simple endeavor is still a complicated affair to achieve. The completion of the Human Genome Project did not signal the end or the apex of sequencing technology, but only its beginnings. Though it was a monumental achievement, it was not sufficient to categorize and pinpoint molecular pathways of disease (1,2). Certain diseases were easy enough, monogenic diseases such as cystic fibrosis, those determined by single alleles, and work to track those genes were already underway without using sequencing technology (3-5). Sequencing technology in these cases could give a very clear diagnosis by seeking and sequencing the allele in question. However, sequencing technology is ideal for the categorization and diagnosis of complicated, multi-genic diseases, and the identification of genetic markers for predisposition to diseases states (6-12).

One application of next-generation sequencing is to sequence cancer genomes with the intention to identify oncogenes, whether they were inherited genes that increased cancer susceptibility or somatic mutations that led to the development of tumor tissue. Cancer presents a different and interesting problem than other polygenic diseases, though they share many overlaps from the angle of genetic analysis. Polygenic diseases can include heart disease or Alzheimer's, and along with cancer, have potentially inherited genetic

traits that can affect predisposition to these disease states (3-5,13). BRCA1 and BRCA2 were both discovered to be inherited traits that increased breast cancer rates, whereas mutations in p53 or the Rb gene seemed to increase general cancer susceptibility. These inherited traits could be detected by sequencing germline genomes from individuals (14-16). Cancer provides the interesting case of somatic mutations which are as determinative in risk and susceptibility. This results in the need, in addition to having a reference human genome, to sequence multiple genomes from a single patient in order to provide a more complete assessment of disease risk (7,14,17-19). This complication is layered on top of the more usual genetic aspects of interest: mutated genes, specific SNPs, translocations, or copy number variations (19,20). All of these can have varying effects on disease states, which is further convoluted by environmental impacts and penetrance of the observed genetic abnormalities. This leads to constant demand for greater sequencing standards in sequencing speeds, fidelity, and read lengths (21-25).

In the acquisition of sequencing information, various biochemistries and methodologies are employed, that all have aspects that can be improved and further optimized (26-28). The sequencing medium can vary, utilizing beads that are either iron and therefore magnetic, or comprised of polystyrene, making them non-magnetic (29,30). These methods can be performed in real-time or in cycles (31,32). The genetic template can be amplified or not, relying on single-molecule sequencing (2,33,34). Detection methods can involve fluorophore-coupled oligomers, or the detection of liberated hydrogen ions during synthesis, or luciferase (32,35). Sequencing can involve different enzymes, such as ligases for Sequencing By Ligation (SBL), polymerases for Sequencing By

Synthesis (SBS), or can involve no enzymes, such as in Sequencing By Hybridization (SBH) (26,27,29,36).

SBL in particular, has a limitation in read-lengths, though it can provide bi-directional reads (30) unlike traditional SBL approaches. One project of this dissertation focused on increasing the read lengths of traditional SBL by utilizing a non-proprietary method that involved using deoxyinosine as both a universal base and a substrate to be recognized by Endonuclease V (37). This method could be performed with off-the-shelf reagents and increased read-lengths through cyclic digestion and re-ligation. Results increased traditional SBL results from approximately seven bases or so to thirteen contiguous bases in the 3' to 5' direction. Using mate-paired tags with this cyclic SBL variation would enable > 95% coverage of the genome. The potential gains of this cyclic SBL variation were calculated using in-house developed software, the Sequence Analysis Workbench Tool (SAWTooth) (38).

Once the sequence information has been generated and gathered, many steps of analysis must be performed, depending on the desired information. In the case of cancer genome sequencing, variations in the genome would yield the most interest; whether differences between patient germline and reference genomes for hereditary factors, or between germline and tumor tissue for somatic variations (14-16,39,40). One of the first steps is simply matching these tags to the reference genome, filtering out those that match perfectly and identifying SNPs and Indels (41-43). SAWTooth's capability in this regard, and ability to simulate potential gains from the cyclic SBL variation is not unique (44-47), however, SAWTooth was able to perform this task far more efficiently than other existing codes. This is important because the sizes of datasets being generated to adequately

sequence the human genome, which is three gigabases long, have become extremely large, and an efficient algorithm to perform this mapping was necessary (46,48). SAWTooth uses a pre-compiled hash-indexing algorithm to achieve faster mapping times, and even with the time required to compile and generate the hash indexes, still demonstrated faster match times for short mate-paired tags.

The last component of the dissertation is meta-analytical in nature, comparing a selection of *de novo* assemblers and assessing their ability to assemble a known genome given real and simulated mate-paired and single tag data. There is currently a large selection of *de novo* assemblers to choose from. Many assemblers support reference mapping as well, since it is a less complex computational task (25,49,50). There are commercial, free, and open-source assemblers and they each utilize different algorithms. Many of them have been used in the publication of draft genomes, though predicting accuracy of a draft genome that has been assembled *de novo* is difficult (41,49,51). There is no real verification, and often re-sequencing efforts reveal errors with rearrangements and translocations in the assembly. Therefore, it seemed prudent to analyze a few of the most used and popular assemblers in widespread use today.

All these topics are separated by both subject and field, but are all necessary to collectively form the steps of next-generation sequencing. Each of these steps is obviously complex, and this dissertation only explores one small facet of each, whether dealing with sequencing biochemistry, cancer biology, reference mapping or *de novo* assembly.

References

1. Nebert, D.W., Zhang, G. and Vesell, E.S. (2008) From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug metabolism reviews*, **40**, 187-224.
2. Bennett, S.T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics*, **6**, 373-382.
3. Gaudet, M.M., Kirchhoff, T., Green, T., Vijai, J., Korn, J.M., Guiducci, C., Segre, A.V., McGee, K., McGuffog, L., Kartsonaki, C. *et al.* (2010) Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. *PLoS genetics*, **6**, e1001183.
4. Walsh, T., Lee, M.K., Casadei, S., Thornton, A.M., Stray, S.M., Pennil, C., Nord, A.S., Mandell, J.B., Swisher, E.M. and King, M.C. (2010) Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 12629-12633.
5. Welcsh, P.L. and King, M.C. (2001) BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Human molecular genetics*, **10**, 705-713.
6. Hobert, O. (2010) The impact of whole genome sequencing on model system genetics: get ready for the ride. *Genetics*, **184**, 317-319.
7. Stratton, M.R. (2011) Exploring the genomes of cancer cells: progress and promise. *Science (New York, N.Y.)*, **331**, 1553-1558.
8. Dudley, J.T., Chen, R. and Butte, A.J. (2011) Matching cancer genomes to established cell lines for personalized oncology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 243-252.
9. Chin, L., Hahn, W.C., Getz, G. and Meyerson, M. (2011) Making sense of cancer genomic data. *Genes & development*, **25**, 534-555.
10. Katsios, C., Zoras, O. and Roukos, D.H. (2010) Cancer genome sequencing and potential application in oncology. *Future oncology (London, England)*, **6**, 1527-1531.
11. Goya, R., Sun, M.G., Morin, R.D., Leung, G., Ha, G., Wiegand, K.C., Senz, J., Crisan, A., Marra, M.A., Hirst, M. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics (Oxford, England)*, **26**, 730-736.
12. Futreal, P.A. (2007) Backseat drivers take the wheel. *Cancer Cell*, **12**, 493-494.
13. Perez-Losada, J., Castellanos-Martin, A. and Mao, J.H. (2011) Cancer evolution and individual susceptibility. *Integrative biology : quantitative biosciences from nano to macro*, **3**, 316-328.
14. Timmermann, B., Kerick, M., Roehr, C., Fischer, A., Isau, M., Boerno, S.T., Wunderlich, A., Barmeyer, C., Seemann, P., Koenig, J. *et al.* (2010) Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PloS one*, **5**, e15661.
15. Rothenberg, S.M. and Settleman, J. (2010) Discovering tumor suppressor genes through genome-wide copy number analysis. *Current genomics*, **11**, 297-310.
16. Bonifaci, N., Gorski, B., Masojc, B., Wokolorczyk, D., Jakubowska, A., Debniak, T., Berenguer, A., Serra Musach, J., Brunet, J., Dopazo, J. *et al.* (2010) Exploring

the link between germline and somatic genetic alterations in breast carcinogenesis. *PloS one*, **5**, e14078.
17. Gonzalez-Bosquet, J., Calcei, J., Wei, J.S., Garcia-Closas, M., Sherman, M.E., Hewitt, S., Vockley, J., Lissowska, J., Yang, H.P., Khan, J. *et al.* (2011) Detection of somatic mutations by high-resolution DNA melting (HRM) analysis in multiple cancers. *PloS one*, **6**, e14522.
18. Parmigiani, G., Boca, S., Lin, J., Kinzler, K.W., Velculescu, V. and Vogelstein, B. (2009) Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics*, **93**, 17-21.
19. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153-158.
20. Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)*, **318**, 1108-1113.
21. Yang, M.Q., Athey, B.D., Arabnia, H.R., Sung, A.H., Liu, Q., Yang, J.Y., Mao, J. and Deng, Y. (2009) High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. *BMC genomics*, **10 Suppl 1**, I1.
22. Johansson, M.L. (2009) Next generation sequencing in nonmodel organisms: has the future arrived? *The Journal of heredity*, **100**, 807.
23. Teer, J.K. and Mullikin, J.C. (2010) Exome sequencing: the sweet spot before whole genomes. *Human molecular genetics*, **19**, R145-151.
24. Ocana, A. and Pandiella, A. (2010) Personalized therapies in the cancer "omics" era. *Molecular cancer*, **9**, 202.
25. Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics*, **11**, 31-46.
26. Gao, L. and Lu, Z. (2009) The removal of fluorescence in sequencing-by-synthesis. *Biochemical and biophysical research communications*, **387**, 421-424.
27. Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C. *et al.* (2009) The challenges of sequencing by synthesis. *Nature biotechnology*, **27**, 1013-1023.
28. Housby, J.N. and Southern, E.M. (1998) Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic acids research*, **26**, 4259-4266.
29. Porreca, G.J., Shendure, J. and Church, G.M. (2006) Polony DNA sequencing. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, **Chapter 7**, Unit 7 8.
30. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, **309**, 1728-1732.
31. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, **7**, 461-465.

32. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, **323**, 133-138.
33. Efcavitch, J.W. and Thompson, J.F. (2010) Single-molecule DNA analysis. *Annual review of analytical chemistry (Palo Alto, Calif.)*, **3**, 109-128.
34. Milos, P. (2008) Helicos BioSciences. *Pharmacogenomics*, **9**, 477-480.
35. Ghasemi, Y., Peymani, P. and Afifi, S. (2009) Quantum dot: magic nanoparticle for imaging, detection and targeting. *Acta bio-medica : Atenei Parmensis*, **80**, 156-165.
36. Drmanac, R., Drmanac, S., Chui, G., Diaz, R., Hou, A., Jin, H., Jin, P., Kwon, S., Lacy, S., Moeur, B. *et al.* (2002) Sequencing by hybridization (SBH): advantages, achievements, and opportunities. *Advances in biochemical engineering/biotechnology*, **77**, 75-101.
37. Ho, A., Murphy, M., Wilson, S., Atlas, S.R. and Edwards, J.S. (2011) Sequencing by ligation variation with endonuclease V digestion and deoxyinosine-containing query oligonucleotides. *BMC genomics*, **12**, 598.
38. Murphy M., W.S.M., Ho A., and Edwards J.S., A.S.R. (2012 (To be submitted)) The Sequence Analysis Workbench: A Framework for Fast, Parallel Genomic Sequence Mapping. *PLoS computational biology*.
39. Ostrovnaya, I., Nanjangud, G. and Olshen, A.B. (2010) A classification model for distinguishing copy number variants from cancer-related alterations. *BMC bioinformatics*, **11**, 297.
40. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, **40**, 722-729.
41. Young, A.L., Abaan, H.O., Zerbino, D., Mullikin, J.C., Birney, E. and Margulies, E.H. (2010) A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome research*, **20**, 249-256.
42. Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, **26**, 873-881.
43. Costantini, M. and Bernardi, G. (2009) Mapping insertions, deletions and SNPs on Venter's chromosomes. *PloS one*, **4**, e5972.
44. Shankar, R. (2011) The bioinformatics of next generation sequencing: a meeting report. *Journal of molecular cell biology*, **3**, 147-150.
45. Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A. *et al.* (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome research*, **20**, 273-280.
46. Trapnell, C. and Salzberg, S.L. (2009) How to map billions of short reads onto genomes. *Nature biotechnology*, **27**, 455-457.
47. McPherson, J.D. (2009) Next-generation gap. *Nature methods*, **6**, S2-5.
48. Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. and Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome biology*, **10**, R98.

49. Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome research*, **20**, 1165-1173.
50. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, **20**, 265-272.
51. Li, Y., Hu, Y., Bolund, L. and Wang, J. (2010) State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human genomics*, **4**, 271-277.

Chapter 2

Lessons from Cancer Genome Sequencing

by

Antoine Ho [1] and Jeremy S. Edwards [1,2,3]

[1] Molecular Genetics and Microbiology

[2] Cancer Research and Treatment Center

[3] Chemical and Nuclear Engineering

University of New Mexico Health Sciences Center

I. Introduction

The Human Genome Project (HGP) was one the greatest achievements of the 20th Century, and the publication of the full human genome sequence in 2001 ushered in the new century by starting the post-genome era in human biology. The great success of the HGP has paved the way to many future discoveries. The human genome sequence represents just the beginning of the payoffs for the biomedical community, and many future benefits are promised and expected in the near future. Specifically, the HGP has enabled the rapid sequencing of more genomes, such as cancer genomes, and this holds the potential to transform cancer research and treatment. Therefore, it is more appropriate to look at the completion of the human genome as the end-of-the-beginning, rather than the beginning-of-the-end of the era of human genome sequencing. "Next generation" sequencing technologies are providing fast, cheap and high quality sequence. As these technologies become less expensive and easier to operate, they will become more widely available. However, the bottleneck in the process will quickly shift to the analysis phases. In other words, making sense of the vast amount of sequence data will be a challenging task, and it will require bioinformatics and systems biology. The analysis of sequencing data will likely have a tremendous impact on many areas of medicine and biomedical research.

The sequencing and publication of the human genome was performed simultaneously by two competing groups, one was publicly funded and the other was privately funded. The publicly funded sequencing project was led by Dr. Francis Collins and was performed in the classical clone-by-clone approach using traditional Sanger sequencing. The private sequencing project was based at Celera and was led by Dr. J. Craig Venter. The Celera group sequenced the human genome using the shotgun

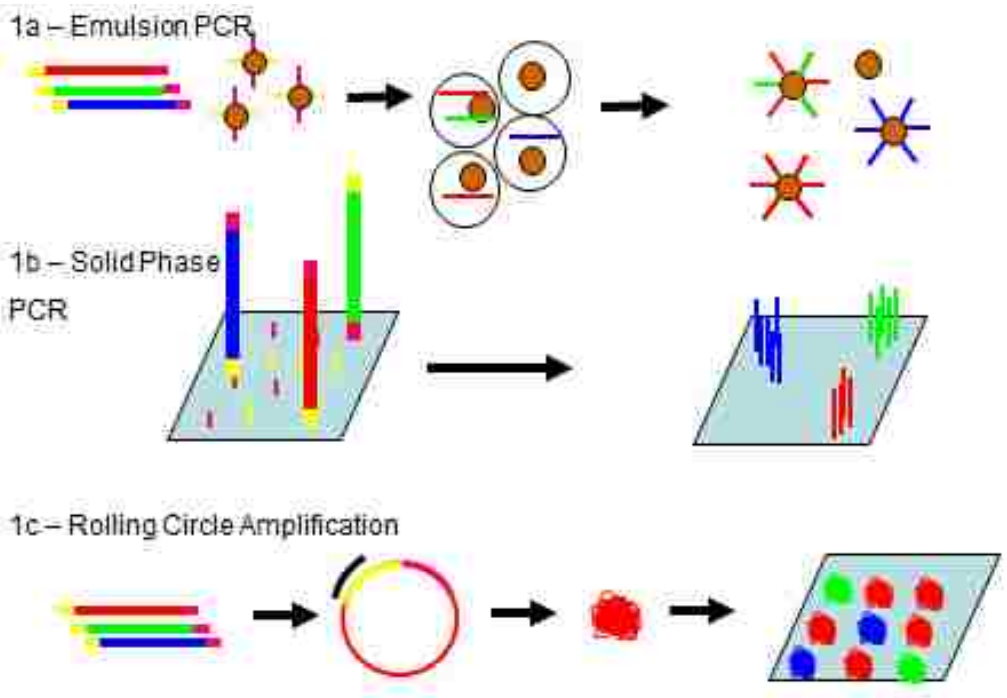
sequencing approach, which was made possible for three main reasons: (a) they developed novel assembly algorithms, (b) they utilized data from the public project, and (c) they sequenced a very homogeneous sample, as opposed to a sample representative of a large number of individuals. (1)

The HGP's impact on future human genome sequencing has two broad implications. First, the HGP has now established a reference human genome sequence, allowing for relatively rapid sequencing of future genomes while using the reference sequence to align reads. Additionally, a major impact of the HGP has been spin-off technologies and bioinformatics tools, which have led to what is now known as "next-generation" sequencing technology. (2)

II. Next Generation Sequencing Technologies

During the HGP, a number of technologies were developed with the goal of increasing sequencing throughput to allow for cheap and rapid human genome sequencing. The first phases of the improvements were essentially advances in instrumentation and miniaturization of the traditional Sanger sequencing approach. However, a number of true next generation technologies were also developed and have become widely available.

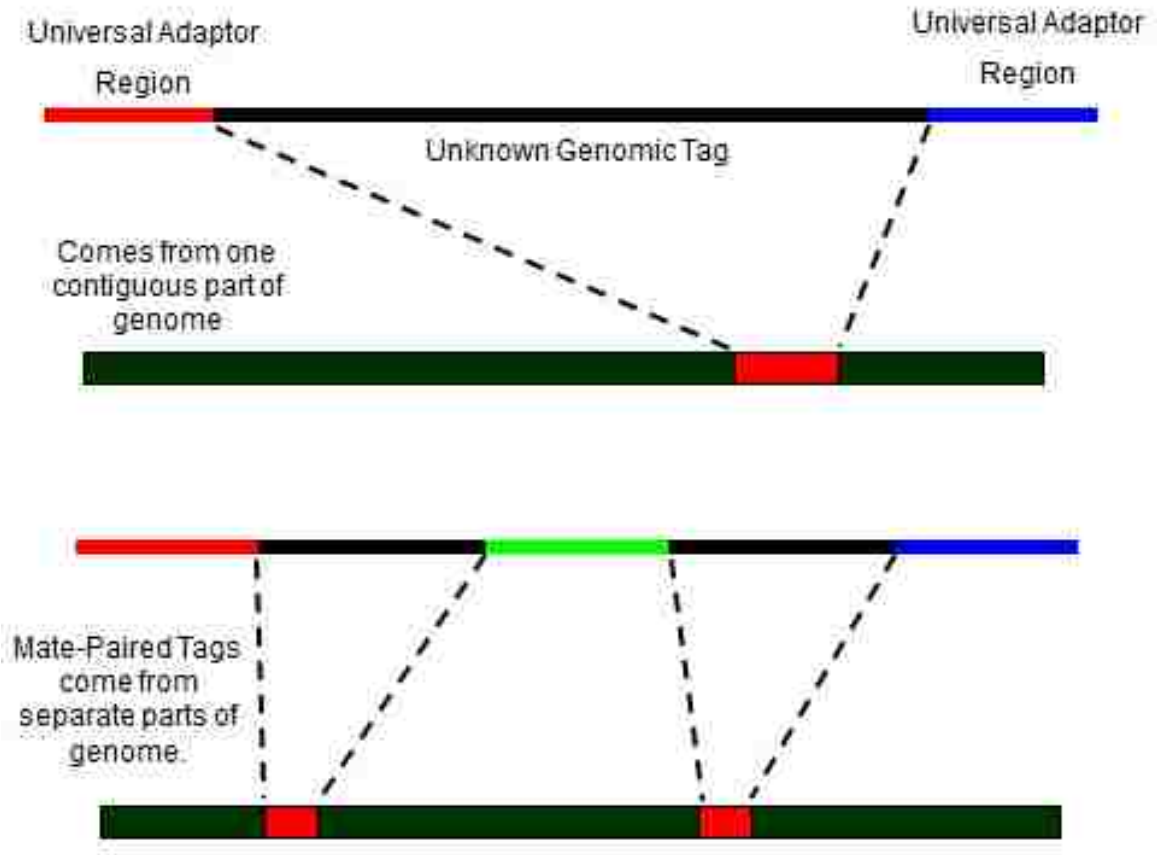
Sequencing Template Preparation



The first step of the next generation sequencing pipeline is the construction of the sequencing library. The library preparation step essentially takes a genomic DNA sample, and converts it into DNA molecules that can be sequenced by a given sequencing technology (see Figure 1). For example, sequencing using the Illumina system, fragments the genomic DNA into ~300 bp fragments, amplifies these fragments via PCR and ligates sequencing primer sites to the ends of the fragments.(3-5) These protocols vary in complexity depending on the sequencing platform.

Additionally, genome libraries can be constructed to contain mate-pair sequences. This means that the genome tags will be adjacent in the library molecule, but will have a kilobase or more separation in the genome. The mate-pair approach complicates library preparation, but assists in genome assembly/mapping, especially when dealing with very

short read lengths, as is typical in most next generation sequencing technologies (see Figure 2). (3-5)



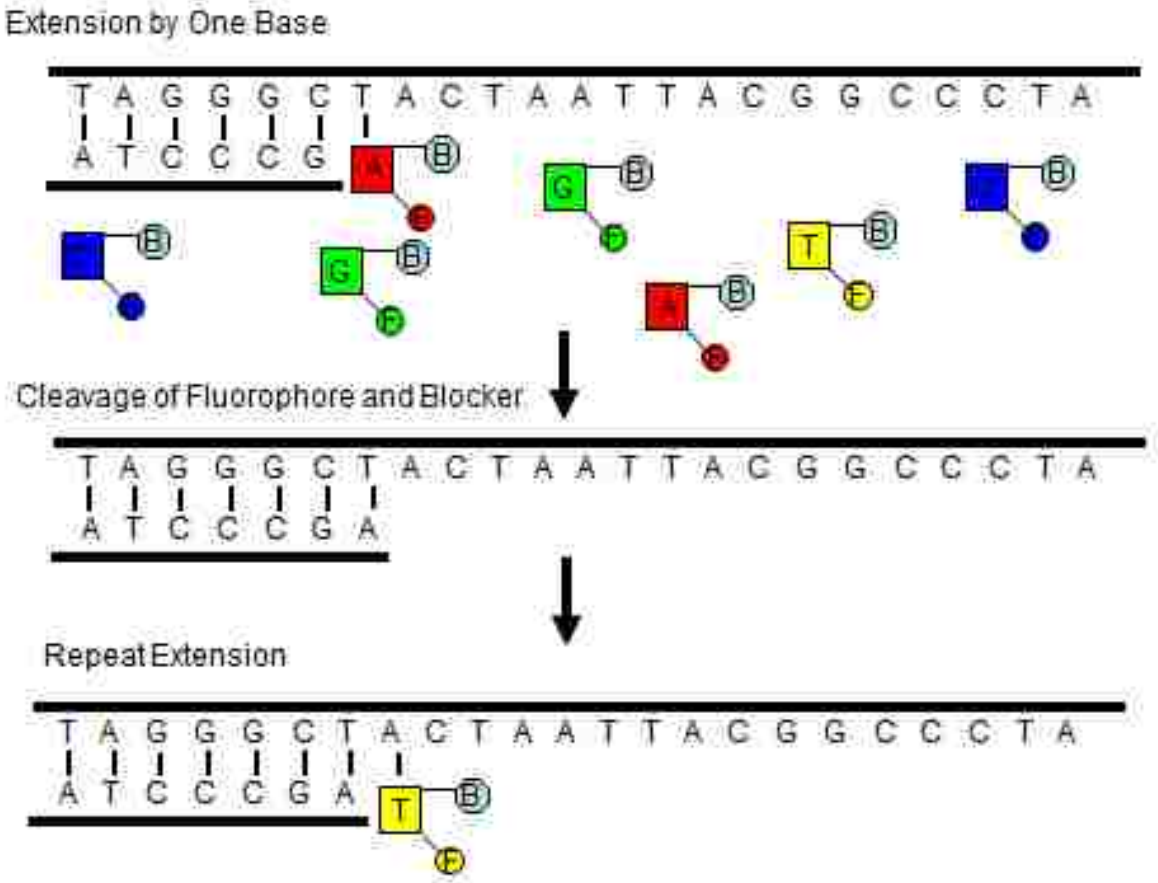
There are many ways to sequence DNA, and because of this, there are many ways in which to prepare the DNA libraries for sequencing. First, the template can be clonally amplified unless sequencing can be performed on single molecules without the need for amplification. Methods that do not rely on an amplification step are known as single-molecule sequencing methods. Amplification is necessary for many sequencing approaches because a signal, whether it is light or electrical, must be amplified or would be too weak to identify otherwise. This amplification can occur through an emulsion PCR (ePCR)(6) step or through solid phase PCR as in the Illumina Inc. system. Additionally,

rolling circle amplification (RCA) can be utilized to amplify the DNA into a ball, which may itself be coupled to an array (see Figure 1). (7) Clonal amplification may make certain sequencing approaches possible, however, when clonal amplicons are being sequenced, the issue of phasing arises. For example, when a clonal population of DNA molecules is being sequenced, the initial signals for sequencing each base are near identical for all molecules. However, as sequencing progresses, inefficiencies in biochemistry, enzymatic activity, chemical cleavage steps, or incomplete washing causes the signal to become noisy and may contain an earlier (lag phasing) or later (lead phasing) position.

Single-molecule sequencing template preparation is greatly simplified, as there is no need for amplification, and there are no amplification biases that may occur. Some single-molecule sequencing methods also make real-time sequencing possible, though there are obstacles to single-molecule sequencing that methods must take into account, such as being able to recognize the signal of a single molecule, which requires more expensive and larger sequencing equipment. (8)

Sequencing By Synthesis

Fluorescent Methods



The most popular next generation sequencing approach is known as Sequencing-By-Synthesis (SBS). In SBS, a DNA polymerase is used to extend a primer on the template strand (see Figure 3). (3-5) The DNA template to be sequenced must contain a known region at its 3' end to hybridize a primer. Once hybridized, synthesis is allowed to occur under controlled conditions with specific reagents. The goal is to allow only the incorporation of a single nucleotide onto this growing strand and to visualize the base that was incorporated. The key is to modify (block) the nucleotides in some fashion that not only allows termination of synthesis once incorporated, but also can be reversible. These can, for example, involve a blocking group on the 3' OH of the growing DNA strand that can be removed enzymatically or by a chemical cleavage reaction.(3-5) The second

element is to attach unique fluorophores onto each of the four different nucleotides to allow visualization. After imaging, and storing this data, the termination must be reversed by removing this blocking group, to allow the addition of another single nucleotide, and then the fluorophores must be cleaved to visualize the signal of the newly incorporated nucleotide. This process is repeated to sequencing up to ~150 bases. SBS can be performed on clonal amplicons from an amplification step (i.e. sequencing being carried out on beads or a clonal cluster of DNA), or SBS can be performed on a single molecule. (3-5,7)

SBS can also be performed in real-time with single-molecule visualization. Real-time SBS methods are faster, but constrained to the viewing area limitations of a camera mounted microscope. Real-time sequencing approaches will likely have a significant impact on cancer systems biology. This is because real-time sequencing has the potential for very long reads, requires a very simple library preparation, and can readout epigenetic markers, such as methylation and hydroxymethylation (9).

Non-fluorescent Methods

In addition to using fluorophores to identify incorporated bases, there are other methods to measure and quantify DNA polymerase extension, such as detecting the H^+ or pyrophosphate released during polymerase extension. Since all bases give the same pyrophosphate or H^+ signal this sequencing approach requires cycles of extending with each of the individual nucleotides. This sequencing approach has the advantage of using natural nucleotides; however, this introduces the homopolymer repeat problem. Namely, these types of sequencing methods must record the intensity of such a signal to deduce how many bases of the same type were incorporated in homopolymer repeats. While

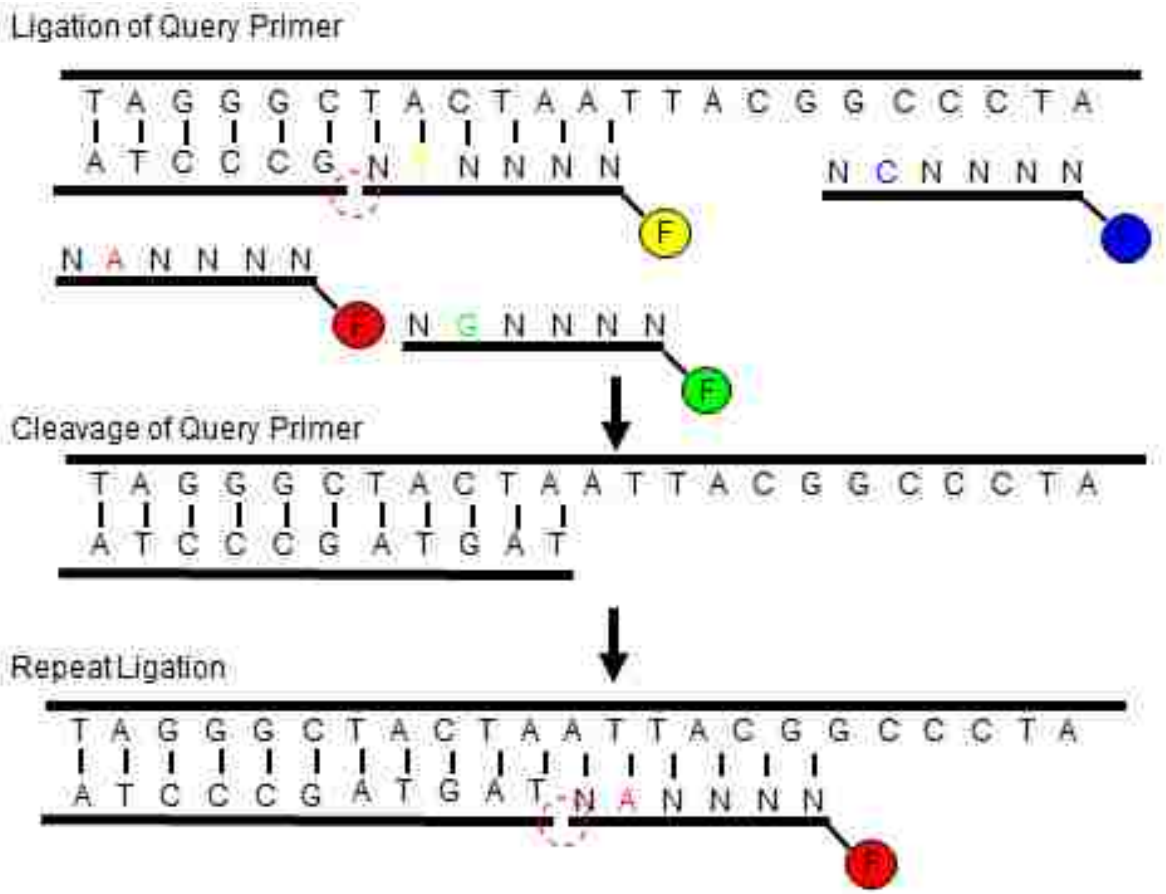
distinguishing the difference in signal between single or double nucleotide incorporation events is straight-forward, it is harder to discern the difference between five or six incorporated nucleotides in a homopolymer repeat. (3-5)

The measured signal can be pH changes, as induced by the release of hydrogen atoms when incorporating a nucleotide during synthesis, or there can be other enzymes involved such as luciferase and sulphurylase that create a flash of light when a phosphate is released during the same process. Due to the nature of this method, sequencing is performed in real-time, and tends to have lower throughput than fluorescent, sequential array methods.

Sequencing By Ligation

Sequencing By Ligation (SBL) uses a ligase and a series of query primers to sequence a template strand. The template DNA to be sequenced will contain the unknown genomic tag, flanked by a known region. The main disadvantage of this sequencing approach is that the read lengths are very short. Therefore, to obtain a reasonable read length, a complicated library preparation is required. The sequencing strategy is to hybridize an anchor primer onto a known region, and ligate a query primer to the anchor primer to sequence the unknown genomic tag. Ligation is determined by hybridization of that query primer next to the anchor primer, meaning it must be complementary for the unknown tag region. The query primers are degenerate, a mix that contains all possible combinations for every position except for one. For example, when determining the identity of the first base next to the anchor primer, the query primer set will be degenerate for all positions, however, in this set, all query primers that have an adenine in that first

position will have a specific fluorophore attached to the other end of the query primer (see Figure 4). (3-5) The clonal features (i.e. beads) will then be imaged, in a manner similarly to fluorescent SBS, and each specific fluorescent signal corresponds with specific bases. This is repeated to obtain the identity of the second base, however, now the fluorophores are specifically linked to bases in the second position of that query primer. This is repeated, generally to a length of 7 nucleotides.



The reason for this read-length limitation is that base pairing is specific closer to the site of ligation and less so further out. To get longer reads, cleavage of the ligated query primer is performed, resulting in loss of the fluorophore and effectively, extending the anchor primer into unknown regions of the genomic tag (see Figure 4). For example,

after sequencing the 2nd base of a tag, and imaging the array, the signal and query primer can be cleaved after the 5th base. The anchor primer will be extended five bases. Now when using same query primer that sequenced the 2nd base, it will now sequence the 7th base. Repeat the process, extend the anchor primer by another five bases, and sequence the 12th base with the same query primer. This is repeated to get longer reads. When signal becomes too weak to continue, the growing ligated template is removed and the sequencing can be repeated to sequence further, for instance, the 3rd, 8th, 13th … positions. This is repeated overall to obtain a contiguous sequence for the genomic tag.

The read-lengths of SBL are shorter than those obtainable from SBS, however, SBL can be performed in both 5' to 3' as well as 3' to 5' directions, whereas SBS must be performed in the direction of DNA synthesis. Similarly to SBS, SBL can be performed on DNA amplified on beads or DNA clusters, and many different types of enzymes or chemicals can be used to cleave the query primer. SBL can suffer from phasing errors as well through repeated ligation and cleavage, but the problem is reduced through changing anchor primers, the entire array is "reset" by washing with a buffer that strips and single-strands the DNA.

Sequencing Through DNA Observation

In addition to methods that involve sequencing a template strand by building a complementary sequencing strand through SBS or SBL, there are emerging methods of sequencing that focus on observing certain traits of the DNA itself. These methods are always single-molecule sequencing methods, and as of this writing, are not commercially available. (10)

A key to observing DNA is to make the DNA single-stranded and pulling this single-stranded DNA through a detector or a nanopore. As the DNA passes through the nanopore, a detector must measure the electrical current which is different for each of the individual nucleotides as they pass through the nanopore. (11)

In addition there is also a method to sequence DNA by directly visualizing it using electron microscopy. This involves stretching the DNA on a surface and visualizing the DNA by conjugating metal ions to specific nucleotides, which are read out in the respective order using electron microscopy. (12)

III. Analysis of Sequencing Information

Sequencing the genome was a monumental task in of itself, but deciphering the data is critical and complicated. The human genome is three billion base pairs long, and humans are diploid, and thus each individual carries two homologous chromosomes. Furthermore the genome is not simply a random arrangement of the four bases. If it were random, sequencing it would be a lot easier. However, when Mother Nature finds a motif or a protein shape that functions well, she will use it again and again. While this conservation of form and function is elegant and pragmatic, it makes sequencing difficult. These motifs, and regions of similarities may span hundreds of bases and may be located far apart. There are also regions of extreme redundancy called microsatellites, where short patterns, one to six bases in length, will repeat over and over again. These traits make the genome difficult to sequence, but there are sequencing methods to mitigate these obstacles. (13)

Various sequencing technologies have varying read lengths and the longer the read length, the easier it is to sequence redundant regions of the genome since many sequencing

reads will contain part of the redundant region as well as more uniquely identifying adjacent regions. Long read length assists greatly in allowing one to align, or put together, the sequences obtained from a sequencing run. Another important trait to consider is how many reads one can obtain from the genome The number of reads multiplied by the average read length gives the total number of bases sequenced, and this product divided by the genome's size (three billion for humans) gives us the coverage. Coverage is important for identifying Single Nucleotide Variations (SNVs), since an altered base pair will not align to a reference genome, it is necessary to re-sequence that difference to gain confidence. (13) It is estimated that to identify a large percentage of SNVs would require a coverage of 30x, or at least 100 gigabases of sequence. Lastly, the raw accuracy of the sequencing method must be taken into account. Most current next generation sequencing methods can generate sequence with 98-99% raw accuracy.

These factors impact the ability to assemble the sequencing information into a genome. It takes much more information with longer reads to assemble a genome without a reference, or *de novo* sequencing. When a reference sequence is available, shorter reads can be tolerated, since these reads can be aligned to a completed reference genome. This is the most common method for human genome sequencing today, however, information is lost with this approach, namely, information regarding structural variation cannot be resolved from these sequencing studies. Additionally, phasing the SNVs are also not determined, in other words, which of the two homologous chromosomes contain which variant cannot be determined.

The goal of genome sequencing is to ultimately use this information and improve medical treatment for various disease states that are influenced by genetic factors, such as

heart disease and of course, cancer. (14,15) The strategy is to catalogue genetic differences that had led to the development of cancer, as well as use this information to engineer specifically targeted therapeutic measures. The sequence information and what can be inferred varies on the nature of the information, how the sequence was obtained, and what it was compared to.

The practice of associating disease states with specific genome information is Genome-Wide Association Studies (GWAS). GWAS were initially performed with microarrays that targeted specific candidate genes and known SNPs across the genome. However, in cancer the problem is much more difficult. Namely, it is unlikely that a single SNV that may be the direct cause of a disease, such as a single base difference in a chloride ion channel that leads to Cystic Fibrosis. There are a myriad of genes that contribute and protect against tumor progression, all of which interact in a manifold of ways. GWAS therefore require significant sample sizes, and detailed genomic information to determine the nature of the SNVs as they pertain to cancer. Each SNV confers a small percentage of increased or decreased protection to cancer, whether they act in DNA repair pathways, cell growth, or metastasis. However, as genome sequencing technology has advanced, it's not simply a matter of categorizing SNVs in patient samples and determining novel cancer genes, but also genome rearrangements or copy number arrangements. (16,17)

Single Nucleotide Variations

Complete genome sequencing can reveal information about SNVs, which in turn, can provide information about the resulting protein after translation if the SNV resides in an exon. Even if a SNV is not located within an exon, changes to promoter regions for

example, may impact the transcription of a gene and the subsequent amount of protein product which may then affect cancer development. (18)

These SNVs can be substitutions from one base pair to another, which may result in the usual gamut of synonymous, non-synonymous, or non-sense mutations, which may or may not change the amino acid and the protein produced. In addition, there could be insertions or deletions (sometimes collectively referred to as indels), which can also result in a frame-shift that completely alters the protein product made.

Cancer sequencing requires a high coverage to accurately detect SNVs. Therefore, high coverage, or repeatedly sequencing the same SNV containing region many times will allow the SNV to be called with confidence. Without high coverage, the sequence information may simply be thrown out, incorrectly labeled an inaccuracy in the sequence acquisition itself.

Structural Variations

Chromosomal rearrangements may be caused by a number of factors, and there is a range in consequences for these events. Even between healthy individuals, genome structure will vary without observable detrimental effects. However, it is also clear that rearrangements can have effects on disease states. (19-21)

To obtain information about rearrangements, translocations, insertions, and deletions genome sequence over a wide range must be obtained, even if that entire range isn't sequenced directly. In other words, mate-pair sequencing is crucial for discerning structural variations. In mate-pair sequencing, two short reads are obtained, but in addition to the sequence, the relative position of these two reads is known. This knowledge of how

these two reads are connected is critical for uncovering structural variation. For example, if one cannot map the two short reads to an area in the reference genome, but find that the mate-pairs map too close or too far, it is possible to make inferences about whether a large indel is involved or if that region was rearranged completely. The key is having a library that is constructed with the mate pair design, as well as having an adequate coverage to increase the confidence of found structural changes.

Copy Number Variations (CNVs) are another type of structural variation which is similar to indels, involving either the deletion or duplication of large parts of the genome, which results in increased or decreased, or even deleted copies of genes. (22,23) CNVs can effectively result in the under expression of key tumor suppressors or over expression of oncogenes, resulting in cancer development. CNVs are obtainable from genome sequencing, although there are optimized protocols to specifically identify these. Identification of CNVs with genome sequencing can be difficult, and special attention during the sequencing and library preparation must be made if this information is desired.

Somatic Mutations and Inheritance

A cancer genome will contain more sequence variants than a "normal" germline genome. Specifically, in addition to the natural SNPs in the individual, the tumor will also contain a number of somatic mutations and structural changes. (24-26) Therefore, sequencing a genome that comes from a cancer patient's tumor will identify many more alterations in the genome than sequencing a genome from non-tumor tissue. It is assumed that an individual develops cancer due to mutations occurring in cells that results in those cells being positively selected for in terms of growth. There are many key areas in cell

growth and regulation that need to be perturbed to allow tumor development; DNA repair pathways, cell growth and division, apoptosis, etc. Therefore, the differences between the tumor genome and the germline genome are considered somatic mutations. These somatic mutations are considered important because a subset of these mutations gave rise to tumorigenesis.

This gives researchers options when comparing cancer genomes in order to obtain the information they consider relevant. Comparison of a patient's germline genome with reference genomes will assist in finding inherited genes that may have contributed or increased a patient's risk for cancer. On the other hand, comparison of a patient's germline and tumor genome will reveal a list of somatic mutations that may have lead to the development of cancer. There is a risk however, in identifying somatic mutations because one of the hallmarks of cancer development is lax DNA repair and reduced apoptosis. Therefore, a cancer genome will have many mutations that have nothing to do with cancer development because pathways that would normally stop further mutations have already been damaged.

Drivers and Passengers

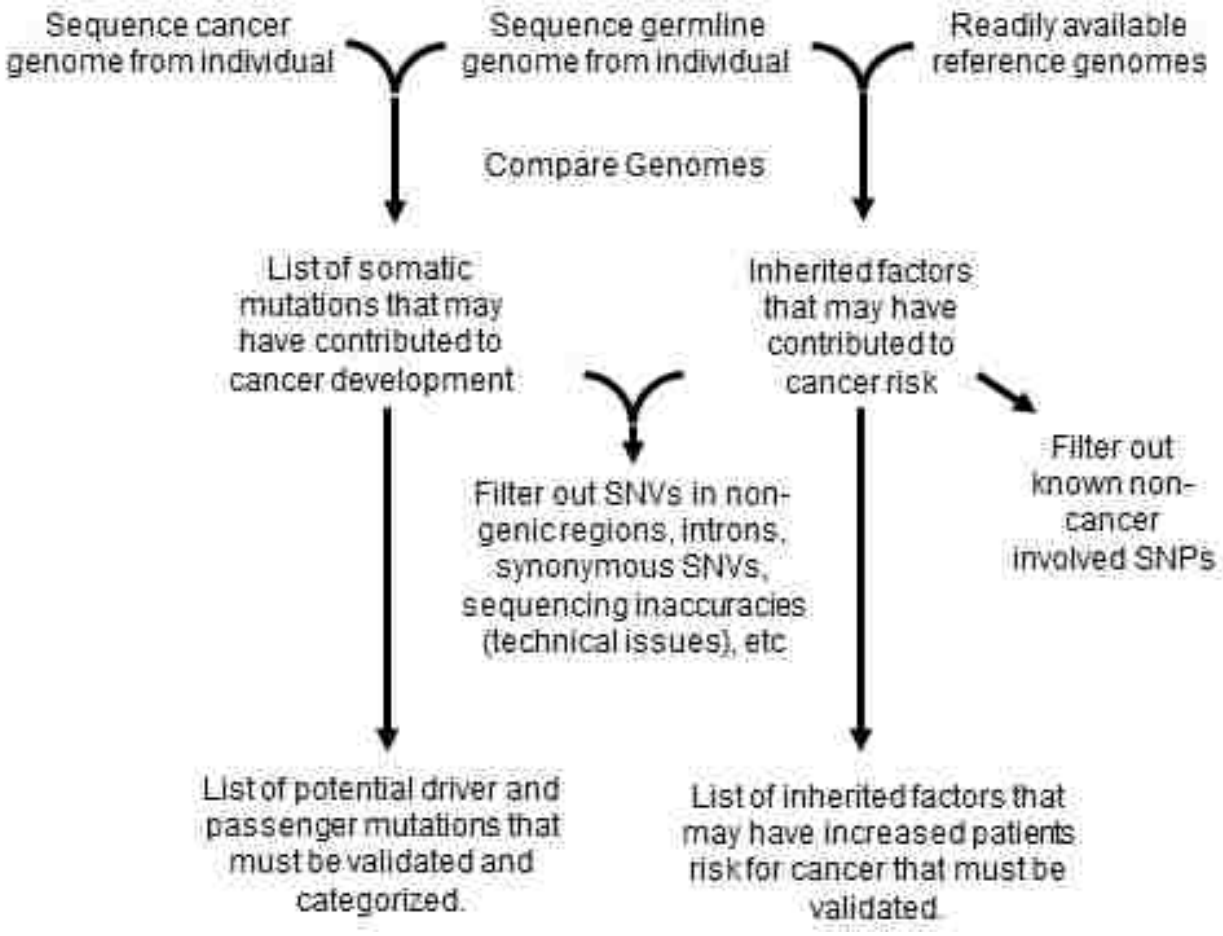
The difference between a mutation that leads to cancer development and those mutations that are merely the result of a cancerous cell allowing other mutations to randomly arise are the difference between so-called driver and passenger mutations. Drivers are present due to selection during cancer development, whereas passengers have been mutated and have no functional consequence. (27,28) Consequently, on top of analyzing data and statistically determining what mutations are even real, one must

determine what mutations are important. Experimental verification of a potential driver mutation would be time consuming, requiring careful bench science experiments with observations of knockouts and knockdowns of the candidate genes. Depending on the organism used, results may or may not even be relevant. Experimental verification would also run counter to how data from genome sequencing is generated, which is a discovery-based approach to research. There are other potential methods reliant upon pre-existing knowledge about genes and their function, where mutation driver or passenger status can be verified with a literature search. However, this still falls in the same trap of requiring time-consuming experimental verification.

To discover and classify driver and passenger mutations and genes through genome sequencing alone would require a much larger sample size. Only through a large database of high-quality genome sequences will true driver mutations be made evident. Different cancer types most likely have different somatic evolution, creating a need for a large sample size of human genomes, but also for patient genome information for each specific cancer. (27,28)

There are computational approaches that have been developed to look at the complete set of somatic mutations and identify putative cancer genes, or basically separate out the driver and passenger mutations (see Figure 5). To identify somatic mutations, complete genome sequencing on a patient's germline and tumor tissue must be carried out. Comparisons between the two will yield a list of differences that must be processed thoroughly. SNVs that, for example are in introns or are synonymous, are eliminated and classified as passenger somatic mutations. Once driver and passenger mutations are

identified, a validation must ultimately be performed to confirm whether these somatic mutations had an effect on cancer development.



IV. Cancer Genome Sequencing Strategies

SNP Profiling

Single Nucleotide Polymorphism (SNP) profiling is not actually sequencing, however, it is a useful, and relatively low cost, genotyping tool for analyzing a large sample size. (29-32) In fact, it is the ability to perform a study with a large sample size that is SNP profiling's greatest strength. SNP profiling is performed with SNP arrays, which have been following similar trajectories as next-generation sequencing in terms of throughput,

increased number of SNPs investigated per array, etc. The SNPs on the array may not have anything to do with cancer, but with large sample sizes, regions of the genome can be identified and these regions can be studied in great detail using targeted resequencing strategies on a very large sample size (a sample size much too large for full genome sequencing). In addition, SNP profiling can provide CNV information that will also be very useful in tracking down cancer causing genes.

Paired-End Mapping

Paired-End mapping is a type of genome sequencing strategy that can more effectively provide information about genome structure and variation (33,34). Variations in structure can cause varying expression in cancer developmental pathways by altering expression. Similar to SNPs, even healthy individuals will differ greatly in terms of genome structure, (35) but there are obviously variations that can lead to an increased risk of cancer. Genomic structural changes may also impact other factors, such as by disrupting exon and intron organization, leading to altered proteins. Additionally, CNV may be affected as well as gene synteny or order.

Paired-end mapping can be performed on many different sequencing methods, whether it is various SBS or SBL methods. Paired-End mapping requires a library that has been mate-paired, where two reads are separated by a known distance. For the specific applications of pair-end mapping, a larger separation distance is often required, due to the fact that indels may be several kilobases in length. Genome structural variations may be investigated with arrays, but next generation sequencing methods allow higher resolution mapping.

Targeted Resequencing

The human genome is very large, and researchers may be more interested in obtaining more focused information, such as focusing purely on the exome (all exons), (36-38) or even epigenetic changes such as the methylome (all methylated genes) (39,40) or the kinome (all kinased genes). (38,41) For example, the exome, in addition to being about 2% of the entire genome, is focused on only the expressed regions of the genome where many (if not most) of the important somatic driver mutations will lie. Additionally, targeted resequencing could focus on the transcriptome, where sequencing the transcriptome provides information about variation in the expressed exons as well as important information regarding the gene expression level and splice variants. Splice variants as well as expression information on these variants can provide valuable insight into how genomic sequences translate into protein products.

Exome sequencing is the targeted sequencing of all known exons. Exome sequencing has advantages and disadvantages with respect to transcriptome sequencing. First, the advantages of exon sequencing are that all exons are equally represented so the coverage is essentially equal, minus stochastic effects, across all exons, whereas in transcriptome sequencing the highly expressed exons are present in large excess and hence, over sampled and the lowly expressed exons are often not adequately covered. Also, information is gathered about all exons, not only the expressed exons as in transcriptome sequencing. As in transcriptome sequencing, findings are consolidated into areas of the genome that are translated. (42,43) The disadvantage to exon sequencing is the complexity associated with isolating all exons from the genome, however, there are currently "kits"

available to enrich for all exons and these approaches are becoming easier and cheaper. The most common methods for exon enrichment are PCR and capture-based approaches.

The first complete exon sequencing study was undertaken by Sjoblom et al (44) in a study focused on colorectal and breast cancer. A total of eleven colorectal cancer samples, along with eleven breast cancer samples, and their corresponding normal tissues were sequenced. The entire exome was sequenced with over thirteen-thousand genes. The identified variations were narrowed down to identify cancer related mutations by eliminating synonymous mutations, as well as SNVs that were present in the germline normal. This approach has the added benefit of consolidating their findings into exons, which focuses the found changes into actual translated sequence.

Whole Genome Sequencing

Whole genome sequencing is self-explanatory, sequencing is performed on the entire genome in its entirety with its introns, exons, non-coding regions, repetitive regions, telomeric regions, etc. (45-48) Everything is obtained and in effect, will provide all the information that the above methods can give and more, with the exception of expression based data. In addition to SNVs, no matter whether they reside in introns, exons, and uniquely, non-coding regions will be discovered, as well as structural changes and copy number variations. These differences can all be discovered using just whole genome sequencing, as opposed to performing different sequencing methods to get various information. (3,13) Additionally, chromosomal rearrangements are detectable through whole genome sequencing, as opposed to other methods, which seek to parse down the information. The drawback of whole genome sequencing is the massive and redundant

human genome, making whole genome sequencing expensive and laborious. This often results in greatly reduced sample sizes, making statistically significant observations difficult. Not only is the acquisition of sequence data more stringent in its requirement, but the alignment and assembly of information, even with the aide of a reference, is still problematic. This significant obstacle must be tackled from a computation angle. Regardless, the high quality information of whole genome sequencing is the most detailed, and therefore has the most potential to be useful.

The first full cancer genome sequencing study was performed by Ley et al. (49) They used the Illumina sequencing platform, which is SBS based. They were able to identify a complete set of somatic mutations that resulted during tumor progression, and were able to identify ten potential cancer genes with acquired mutations, only two of which were previously described. Large parsing of the data was necessary to find the cancer genes, as the original analysis found 2,647,695 single nucleotide variations (SNVs) after quality control checks. 2,584,418 were also found in the patients' germline which had to be eliminated. Of the remaining 63,277 genetic variations, 31,645 were previously described in SNP databases, and 20,440 were in the intra-genic regions. This left a total of 11,192 variants. 10,735 were found were in introns, and 216 were in untranslated regions. This left 241 variants, 60 of which were synonymous. The final 181 variants were non-synonymous mutations, which were then actually investigated further using traditional PCR and Sanger methods. Further extension of this vigorous elimination process yielded ten genes with mutations, eight of which were present in nearly all tumor tissue, but whose functions had not previously been described.

Since this first foray into whole cancer genome sequencing, next-generation sequencing methods have continued to be improved and have become even cheaper, allowing more groups to utilize this methodology. The Ley et al paper, for example, focused primarily on somatic mutations, and categorizing SNVs, insertions, and deletions as passengers and drivers. Cancer genomes sequenced today can also be investigated for chromosomal rearrangements, translocations, and copy number variations. In the near future, we hope to focus on the functional characteristics in the non-coding regions of the genome and the role somatic mutations in these regions have on cancer.

Conclusion

The ability to cheaply generate genome sequences very rapidly will undoubtedly have many medical implications. Ultimately, the value of next generation sequencing technologies will be in the sequencing of large numbers of samples. For example, the ability to sequence hundreds of tumor samples will provide important information toward understanding the microscale evolution that leads to tumor development and will be used to design treatment protocols in the future. Furthermore, sequencing technology is rapidly evolving and will soon allow for large scale sequencing projects to study thousands of human genomes. Currently, having a personal genome project may be of minimal medical value; however, once many genomes are available, we will have a very powerful tool for uncovering the associations between the genotype and the cancer.

Figure Legends

Figure 1a. Emulsion PCR – Template DNA and beads are mixed and then put into an emulsion mixture consisting of an oil phase and an aqueous phase of PCR reagents. These

beads have primers complementary to the ends of the template strands coupled to them, allowing the PCR reaction to extend these primers and cover the bead in copies of the template DNA. Template DNA is diluted to maximize the number of emulsions having exactly one template strand and one bead. Proceed with PCR temperature cycling. Sequencing is performed on beads with only clones of a single template DNA, as beads with no DNA and beads with more than one template DNA do not provide usable data. These beads can then be fixed onto an array for sequencing and imaging.

Figure 1b. Solid Phase PCR – Very similar to ePCR, but without beads. Template DNA is diluted and then added to a slide with primers complementary to end regions of the template DNA coupled to the slide, which allows hybridization and priming. Through a series of PCR temperature cycling, a slide is covered in clonal patches of DNA to be sequenced.

Figure 1c. Rolling Circle Amplification – A piece of linear DNA is circularized enzymatically. Once circularized, RCA is performed with a polymerase that has displacement activity. This results in a ball of clonal DNA, effectively amplifying the DNA but without the need for emulsions or beads. These balls of DNA are then coupled to an array and sequenced.

Figure 2. Mate-Paired Libraries – Mate-paired libraries can provide alignment information that is very valuable, especially when trying to sequence large redundant regions with short reads. The most ideal way to sequence a large redundant region is to simply get a single contiguous read of the entire region, however that may not be technologically possible,

which is why this mate-paired strategy is key. Because the mate-paired reads come from two different regions, a set distance apart, it is possible, even with short reads, that one half of the mate-pair will be in a uniquely identifiable region, and even though the other will be in the redundant, difficult to map region, that read will still provide useful alignment data.

Figure 3. Sequencing By Synthesis with Fluorophores – A primer is hybridized onto the template DNA onto a universal region to allow extension by a polymerase. A single nucleotide will incorporate due to a blocking group on the nucleotides, and the DNA will be able to be visualized by the fluorophores attached to each nucleotide type. If there is a saturation step, as is often the case when dealing with amplified DNA template, it would be performed following the first extension step (not shown). A saturation step is identical to the first step except that there is no fluorophores, though there are still blockers on the nucleotides, and the nucleotides at usually at a very high concentration to saturate. The fluorophores are then cleaved chemically, and then the blocking group is removed so extension can continue another base. This cycle then repeats.

Figure 4. Sequencing By Ligation – A template strand of DNA is exposed to a population of query primers after hybridizing an anchor primer onto a universal region. These are degenerate for all positions except for the position of interest (2nd shown). The nucleotide in the position of interest will determine what fluorophore is attached to this query primer. The query primer will ligate on, allowing imaging to decode the base at the position of interest. This query primer is then cleaved, either enzymatically or chemically, releasing the fluorophores and exposing a new ligation site. Ligation is repeated to obtain further positions.

Figure 5 – Processing SNVs and filtering into somatic mutations – Comparison between germline and tumor genomes provide somatic mutations, whereas comparison between germline and reference genomes can offer information on inherited factors that may have been involved in cancer risk. SNVs that are in non-genic regions, introns, etc. are filtered out, since they either have no effect or don't alter protein function. Validation must follow a highly processed and shortened list of SNVs.

References

1. Collins, F.S., Morgan, M. and Patrinos, A. (2003) The Human Genome Project: lessons from large-scale biology. *Science (New York, N.Y.)*, **300**, 286-290.
2. Diamandis, E.P. (2009) Next-generation sequencing: a new revolution in molecular diagnostics? *Clinical chemistry*, **55**, 2088-2092.
3. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.
4. Reis-Filho, J.S. (2009) Next-generation sequencing. *Breast cancer research : BCR*, **11 Suppl 3**, S12.
5. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nature biotechnology*, **26**, 1135-1145.
6. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.
7. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78-81.
8. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, **323**, 133-138.
9. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*, **7**, 461-465.
10. Efcavitch, J.W. and Thompson, J.F. Single-molecule DNA analysis. *Annu Rev Anal Chem (Palo Alto Calif)*, **3**, 109-128.
11. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nature biotechnology*, **26**, 1146-1153.
12. Xu, M., Fujita, D. and Hanagata, N. (2009) Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small (Weinheim an der Bergstrasse, Germany)*, **5**, 2638-2649.
13. Meyerson, M., Gabriel, S. and Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, **11**, 685-696.
14. Katsios, C., Zoras, O. and Roukos, D.H. Cancer genome sequencing and potential application in oncology. *Future Oncol*, **6**, 1527-1531.
15. Mardis, E.R. and Wilson, R.K. (2009) Cancer genome sequencing: a review. *Human molecular genetics*, **18**, R163-168.
16. Chin, L., Hahn, W.C., Getz, G. and Meyerson, M. Making sense of cancer genomic data. *Genes Dev*, **25**, 534-555.
17. Cui, Q. A network of cancer genes with co-occurring and anti-co-occurring mutations. *PLoS One*, **5**.
18. Goya, R., Sun, M.G., Morin, R.D., Leung, G., Ha, G., Wiegand, K.C., Senz, J., Crisan, A., Marra, M.A., Hirst, M. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730-736.

19. Bignell, G.R., Santarius, T., Pole, J.C., Butler, A.P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S. *et al.* (2007) Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome research*, **17**, 1296-1303.
20. Miller, C.A., Hampton, O., Coarfa, C. and Milosavljevic, A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**, e16327.
21. Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol*, **11**, R52.
22. Ostrovnaya, I., Nanjangud, G. and Olshen, A.B. A classification model for distinguishing copy number variants from cancer-related alterations. *BMC Bioinformatics*, **11**, 297.
23. Rothenberg, S.M. and Settleman, J. Discovering tumor suppressor genes through genome-wide copy number analysis. *Curr Genomics*, **11**, 297-310.
24. Beck, J., Urnovitz, H.B., Mitchell, W.M. and Schutz, E. Next generation sequencing of serum circulating nucleic acids from patients with invasive ductal breast cancer reveals differences to healthy and nonmalignant controls. *Mol Cancer Res*, **8**, 335-342.
25. Bonifaci, N., Gorski, B., Masojc, B., Wokolorczyk, D., Jakubowska, A., Debniak, T., Berenguer, A., Serra Musach, J., Brunet, J., Dopazo, J. *et al.* Exploring the link between germline and somatic genetic alterations in breast carcinogenesis. *PLoS One*, **5**, e14078.
26. Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A. *et al.* COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res*, **38**, D652-657.
27. Frohling, S., Scholl, C., Levine, R.L., Loriaux, M., Boggon, T.J., Bernard, O.A., Berger, R., Dohner, H., Dohner, K., Ebert, B.L. *et al.* (2007) Identification of driver and passenger mutations of FLT3 by high-throughput DNA sequence analysis and functional assessment of candidate alleles. *Cancer cell*, **12**, 501-513.
28. Youn, A. and Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, **27**, 175-181.
29. Amato, R., Pinelli, M., Monticelli, A., Marino, D., Miele, G. and Cocozza, S. (2009) Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. *PloS one*, **4**, e7927.
30. Bae, J.S., Cheong, H.S., Park, B.L., Kim, L.H., Han, C.S., Park, T.J., Kim, J.Y., Pasaje, C.F., Lee, J.S. and Shin, H.D. Genome-wide profiling of structural genomic variations in Korean HapMap individuals. *PLoS One*, **5**, e11417.
31. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R. *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087-1093.

32. Gamazon, E.R., Zhang, W., Dolan, M.E. and Cox, N.J. Comprehensive survey of SNPs in the Affymetrix exon array using the 1000 Genomes dataset. *PLoS One*, **5**, e9366.
33. Bashir, A., Volik, S., Collins, C., Bafna, V. and Raphael, B.J. (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS computational biology*, **4**, e1000051.
34. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, **40**, 722-729.
35. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, **318**, 420-426.
36. Hedges, D.J., Burges, D., Powell, E., Almonte, C., Huang, J., Young, S., Boese, B., Schmidt, M., Pericak-Vance, M.A., Martin, E. *et al.* (2009) Exome sequencing of a multigenerational human pedigree. *PloS one*, **4**, e8232.
37. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272-276.
38. Timmermann, B., Kerick, M., Roehr, C., Fischer, A., Isau, M., Boerno, S.T., Wunderlich, A., Barmeyer, C., Seemann, P., Koenig, J. *et al.* Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PLoS One*, **5**, e15661.
39. Cheung, H.H., Lee, T.L., Rennert, O.M. and Chan, W.Y. (2009) DNA methylation of cancer genome. *Birth defects research. Part C, Embryo today : reviews*, **87**, 335-350.
40. Sun, Z., Asmann, Y.W., Kalari, K.R., Bot, B., Eckel-Passow, J.E., Baker, T.R., Carr, J.M., Khrebtukova, I., Luo, S., Zhang, L. *et al.* Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, **6**, e17490.
41. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153-158.
42. Forrest, A.R. and Carninci, P. (2009) Whole genome transcriptome analysis. *RNA biology*, **6**, 107-112.
43. Levin, J.Z., Berger, M.F., Adiconis, X., Rogov, P., Melnikov, A., Fennell, T., Nusbaum, C., Garraway, L.A. and Gnirke, A. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome biology*, **10**, R115.
44. Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science (New York, N.Y.)*, **314**, 268-274.

45. Ahn, S.M., Kim, T.H., Lee, S., Kim, D., Ghang, H., Kim, D.S., Kim, B.C., Kim, S.Y., Kim, W.Y., Kim, C. *et al.* (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome research*, **19**, 1622-1629.
46. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, **20**, 265-272.
47. Pelak, K., Shianna, K.V., Ge, D., Maia, J.M., Zhu, M., Smith, J.P., Cirulli, E.T., Fellay, J., Dickson, S.P., Gumbs, C.E. *et al.* The characterization of twenty sequenced human genomes. *PLoS Genet*, **6**.
48. Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, **463**, 184-190.
49. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66-72.

Chapter 3

Sequencing By Ligation Variation with Endonuclease V Digestion and Deoxyinosine-Containing Query Oligos

by

Antoine Ho [1], Maurice Murphy [2, 3], Susan Wilson [2,3] , Susan R. Atlas [2, 3, 4], Jeremy Edwards [1, 2, 5]

[1] UNM Department of Molecular Genetics and Microbiology, University of New Mexico

[2] UNM Cancer Center, University of New Mexico

[3] Center for Advanced Research Computing, University of New Mexico\

[4] UNM Department of Physics and Astronomy, University of New Mexico

[5] Department of Chemical and Nuclear Engineering, University of New Mexico

Abstract

Background

Sequencing-by-ligation (SBL) is one of several next-generation sequencing methods that has been developed for massive sequencing of DNA immobilized on arrayed beads (or other clonal amplicons). SBL has the advantage of being easy to implement and accessible to all because it can be performed with off-the-shelf reagents. However, SBL has the limitation of very short read lengths.

Results

To overcome the read length limitation, research groups have developed complex library preparation processes, which can be time-consuming, difficult, and result in low complexity libraries. Herein we describe a variation on traditional SBL protocols that extends the number of sequential bases that can be sequenced by using Endonuclease V to nick a query primer, thus leaving a ligatable end extended into the unknown sequence for further SBL cycles. To demonstrate the protocol, we constructed a known DNA sequence and utilized our SBL variation, cyclic SBL (cSBL), to resequence this region. Using our method, we were able to read thirteen contiguous bases in the 3' - 5' direction.

Conclusions

Combining this read length with sequencing in the 5' - 3' direction would allow a read length of over twenty bases on a single tag. Implementing mate-paired tags and this SBL variation could enable > 95% coverage of the genome.

Background

Following the completion of the human genome project it is anticipated that genome sequencing of an individual will be an aspect of routine treatment for a number of diseases and illnesses, truly ushering in the era of personalized medicine. However, the reality of implementing genome sequencing as a medical tool depends on the cost of sequencing technology [1]. The price tag on the human genome project was $2.7 billion, requiring the labor of hundreds of scientists, and a decade's worth of time [2]. By contrast, sequencing and analyzing a human genome can now be performed for under $50,000 in about four months' time with the labor of a few individuals [3-5]. This advance was made possible by progressing from traditional Sanger sequencing methods to so-called "next-generation" methods that focused on miniaturization of the sequencing reactions, massive parallelization of data acquisition, and computational analysis. This not only resulted in increased sequencing speeds, but also significantly reduced the cost of genome sequencing [6]. However, in order to expand the use of genomic analysis to the clinic, price, quality, and speed must all be advanced further [7-14].

Sanger sequencing remains the gold standard today for accurate DNA sequencing. Sanger sequencing can reach read lengths of up to roughly 1,000 base pairs, dwarfing most current next-generation methods that average fewer than 100 base pairs [15]. What next-generation methods accomplish is massive parallelization, resulting in throughputs that are orders of magnitude greater than Sanger sequencing. However, the throughput gains come at a cost of a reduced read length [1,16,17]. Therefore, Sanger sequencing will remain an essential laboratory tool for years to come; although, for the purposes of large sequencing

projects (i.e. whole genome sequencing, exome sequencing, RNAseq, ChipSeq, etc.), next-generation methods are the new standard [18].

There are multiple sequencing methods that are utilized in next-generation methods. The two most common can be broadly categorized as Sequencing By Synthesis (SBS) [19-21] and Sequencing By Ligation (SBL) [22,23]. SBS is a method of sequencing which utilizes a DNA Polymerase enzyme to incorporate a single fluorescently labeled nucleotide that contains a reversible terminator. This allows a period of data acquisition before removal of the fluorophore, reversal of the terminator, and continuation of sequencing [24]. Additionally, there are single molecule and real-time SBS approaches [25,26], which, as their names imply, are performed without template amplification and sequenced in real-time using some indicator of nucleotide incorporation. In the present work, we have focused on increasing the read length of SBL.

SBL is a straightforward enzymatic method of sequencing DNA. SBL uses known, universal sequences that flank an unknown genomic tag as anchor primer sites [22]. An anchor primer is hybridized to one of these known regions, and a ligatable end (3' or 5' depending on the direction of desired sequencing) is available. An oligo, called a query primer, is then ligated to the end of the anchor primer. The query primer is a mix of oligos that are degenerate for all positions except a single position that is being sequenced, which allows the sequencing of a single position based on the design of the query primer. After sequencing a single position, the query primer and anchor primer are stripped from the DNA template, effectively resetting the sequencing. The process begins again, sequencing a different position by using a different query primer, and repeating until the entire sequence of the tag has been determined [23]. Increased read length can be accomplished

either by increasing the distance SBL can be performed in a single direction, or by incorporating additional universal regions for more anchor primer sites [5,22].

Currently, the number of sequential bases that SBL-based approaches can sequence is limited by loss of specificity of base pair hybridization at any distance away from the site of ligation. Errors in the first six base pairs adjacent to the site of ligation are rare due to the destabilizing effect of mismatches. However, at a distance of about seven base pairs, the specificity of the SBL reaction is reduced (Figure 1). Therefore it is not possible to simply use longer and longer query primers in order to increase SBL read lengths [27].

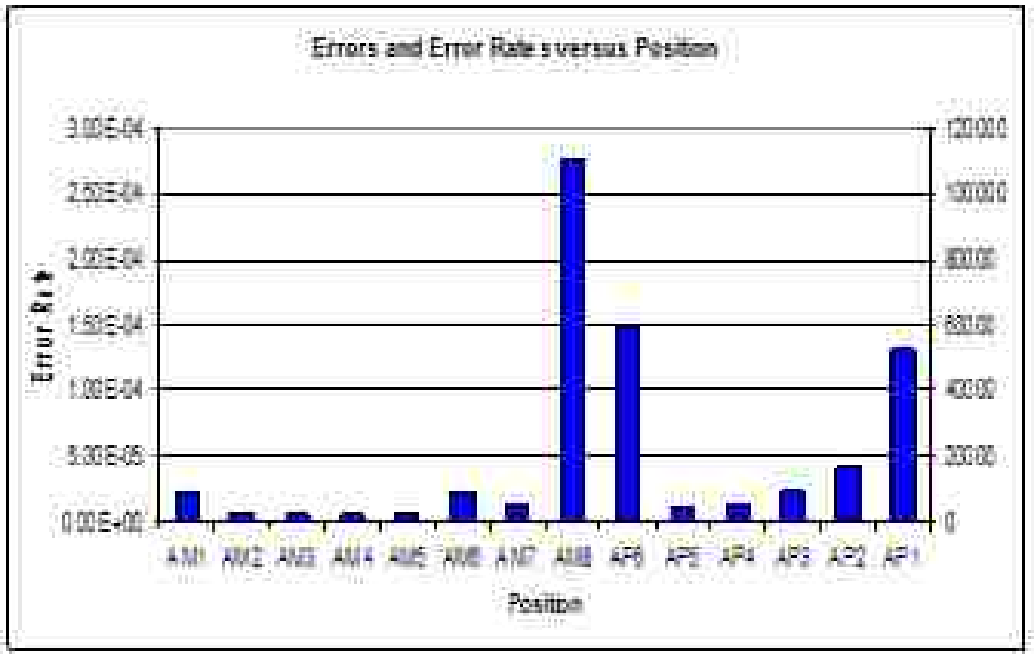


Figure 1. Errors and Error Rate versus Position. Unpublished results in a traditional SBL sequencing run. These reads are separated into two parts, A and B, and are designed either M or P for Minus or Plus, away or towards the site of attachment on the bead. These AM and AP reads are obtained using different hybridized primers. Error and error rates, when not using a cyclic or digestion method, results in loss of specificity the further away a base is from the site of ligation.

In this manuscript, we describe a variation on SBL that utilizes a deoxyinosine in the query primer that can be cleaved by Endonuclease V [28] to increase the read length through successive cycles, which we refer to as cyclic SBL or cSBL. Our approach is conceptually similar to the ABI SOLiD method of SBL, which uses a chemical cleavage of the query primer to get extensions of read lengths. However, in contrast, our method utilizes an enzymatic cleavage using completely off-the-shelf reagents. Deoxyinosine is a universal base [29] that is recognized by Endonuclease V, which cleaves between the 2nd and 3rd phosphodiester bond 3' from the deoxyinosine site [28]. Cyclic SBL is thus identical to standard SBL except that there is a deoxyinosine incorporated in the query primers that is used for cleavage. Therefore, after ligation of a query primer onto an anchor primer, one can use Endonuclease V to cleave off the end of the query primer. This cleavage results in a ligatable end with a portion of the query primer is still ligated to the anchor primer, effectively lengthening the anchor primer for an SBL reaction to increase the SBL read length. The cycles of ligation and Endonuclease V digestion can be repeated to further increase the read length. We have used this approach to extend the read length of SBL to thirteen base pairs in the 3' - 5' direction.

Results

Cyclic SBH

Three cycles of cSBL were performed, giving accurate signal for the first 13 positions of the Test Template. There was a slight increase in non-specific signal with each cycle, but the third cycle still had clearly correct signal with an acceptable signal to noise ratio (Figure 2).

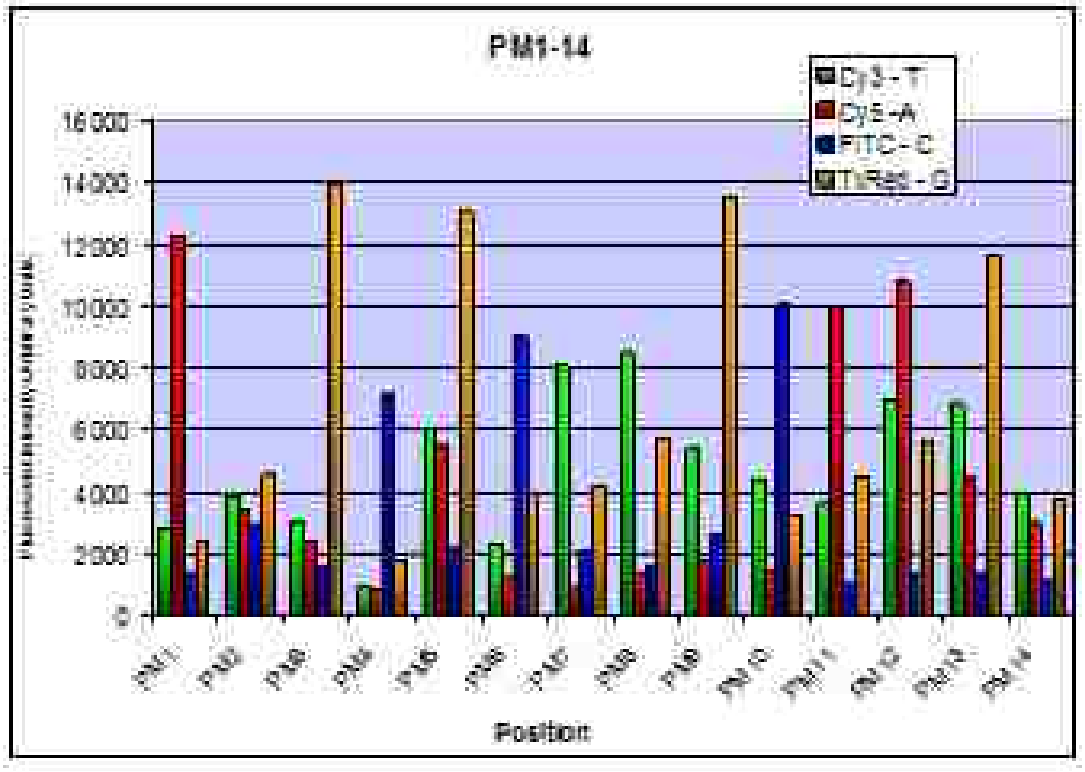


Figure 2. Fluorescent Intensity plotted versus position for each channel, for the Test Template sequence. Sequenced area is underlined. 5' TCT ATG GGC AGT CGG TGA TANGCG CTT GCA AGA GAA TGA GGA AAA CGA AGA 3'.

We were unable to sequence the 14th position and beyond using the cSBL strategy. In order to determine the possible cause of this, we performed a series of tests to explore whether the template DNA had been digested by the Endonuclease V treatment, since this seemed the most likely problem. After the beads had undergone cSBL and stripping of the sequencing strand of DNA, we hybridized a fluorescent probe to the 3' end of the DNA loaded onto the beads and confirmed that the Test Template was still present on the bead.

We also ruled out the issue of secondary structure causing the 3' end of our Test Template to become inaccessible. We performed folding calculations using IDTDNA's Oligo Analyzer software (29) when constructing our Test Template specifically in order to

avoid secondary-structure problems. Calculations for melting temperatures (TM) of secondary structures were performed assuming 50 mM Na+ and 10 mM Mg++. This simulated the highest folding TM at 31.5 degrees, and the fold as modeled by the software was not located near the 14th base pair.

We additionally performed ligation at 50°C using Taq DNA Ligase (NEB), which has a higher optimal temperature, but could not obtain the 14th position or further. We have been unsuccessful in identifying a definitive reason for the observed sequencing limit of 13 continuous bases. However, based on the results from Figure 2, our cSBL strategy does consistently provide at least thirteen base-pair reads in the 3' - 5' direction, and can easily reach twenty-three bases with the addition of a flanking anchor primer site and 5' - 3' sequencing of 10 bases.

Read Length Versus Genome Coverage

To demonstrate the feasibility of a cSBL approach to genome sequencing and calculate gains in using cSBL over traditional SBL methods, we utilized the SawTooth resequencing code developed at the University of New Mexico (M. Murphy et al., to be submitted, 2011). Human genome coverage was simulated using mate-paired data ranging from twenty-six bases to (limit of traditional SBL) to forty bases (theoretical gain from cSBL implementation).

A set of simulated mate-paired tags, each separated by a range of 300-700 bases, was created, ranging in size from 13 paired tags to 20 paired tags. A sufficient number of tags were computationally generated to simulate 10 × coverage. The tags were all generated from chromosome 1, mapped back to the entire genome, and calculations of chromosome

1 coverage were performed. Mapping tags back to the whole genome, instead of just chromosome 1, provided a more realistic comparison to how human genome sequencing is typically performed [30,31]. Tags that mapped to multiple locations, whether in the entire human genome or chromosome 1, were discarded. A tag that maps uniquely or maps back to the reference genome in a single location provides useful data. If a tag maps uniquely to the reference sequence, the loci where it maps are said to be covered by that tag. For a given locus, the number of all such unique mappings when all tags are considered is called the depth of coverage for that locus. SAWTooth uses a general hash index, perhaps the fastest data retrieval structure. Although there are some limitations to general hash indexes, the nature of genomic data and the specialized task of mapping paired end reads to a reference genome, allows the use of hash indexes that circumvent these limitations.

The SawTooth mapping analysis yielded the results summarized in Figures 3, 4, 5. Figure 3 shows raw coverage of chromosome 1 as a function of tag length. Increasing tag lengths from thirteen to twenty, or twenty-six to forty total bases while mate-paired, results in an increased coverage of chromosome 1 from 96% to 97.5%. Gains of coverage are significant when the read lengths are small, but suffer from diminishing returns as read length increases. Also, as expected, depth of coverage increases with tag length (Figure 4).

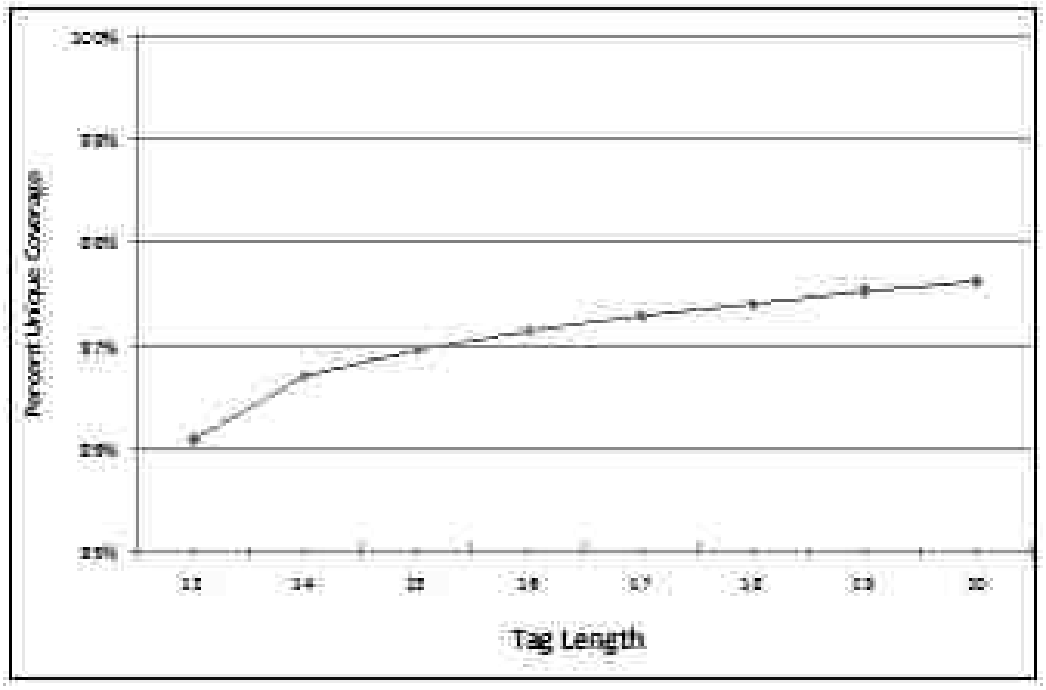


Figure 3. Percent of sequenced regions on chromosome 1 covered by at least one unique mapping, as a function of tag length.

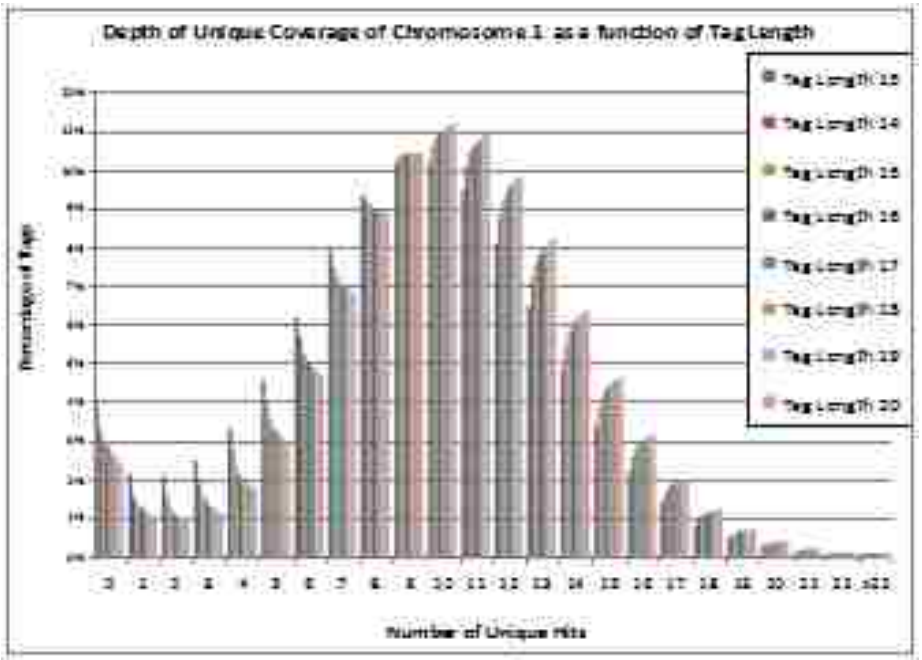


Figure 4. Depth of unique coverage of sequenced regions on chromosome 1 at various tag lengths.

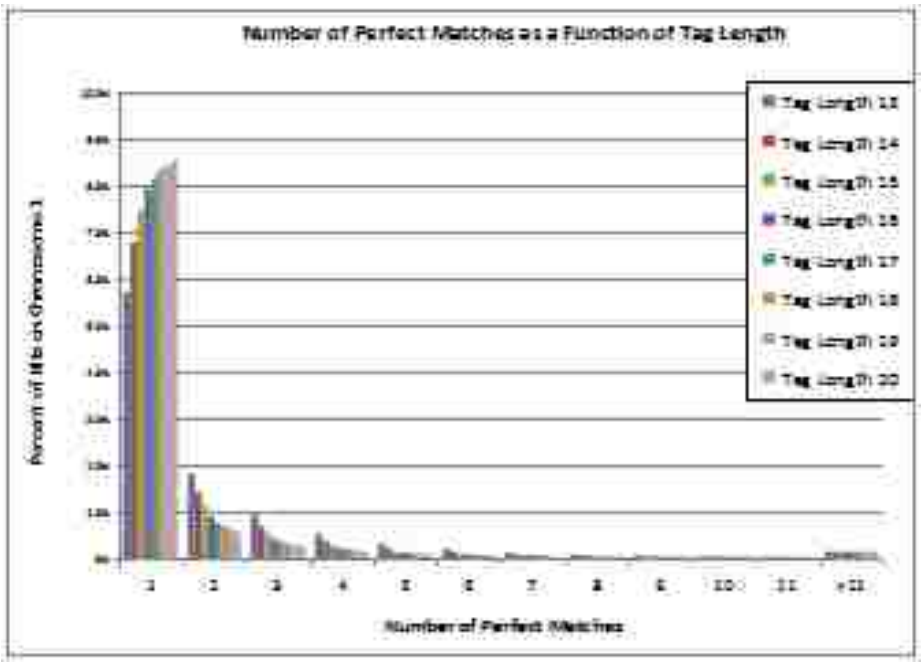


Figure 5. Simulated percentage of reads and the number of perfect matches as a function of tag length.

Next, we performed an analysis of how many times each tag mapped to the genome. One of the more significant benefits gained by increasing tag length from 13 to 20 bases is that far fewer tags must be discarded because they do not map uniquely (see Figure 5). At a tag length of 13 bases, only 57.2% of the tags are used, compared to 85.6% at a tag length of 20, thus effectively increasing throughput.

Discussion

The cSBL protocol described here is a variation on traditional SBL that can increase the read lengths by increasing the number of contiguous bases sequenced. Implementation of the cSBL approach could potentially increase reads to twenty-three base pairs, or forty-six total base pairs with a mate-paired constructed library. In this manuscript, we performed the sequencing on a test DNA template rather than a genome library. however, we expect that any biases or mismatches in our cSBL will be exactly the same as general SBL. These

issues include increased mismatches in specific positions of the query primer [32], or general drops in efficiency when dealing with A or T rich regions of the genome [27]. Additionally, our experiments were performed on beads suspended in solution rather than on beads immobilized on a surface. Therefore, to implement our sequencing strategy in a next generation sequencing platform, the methods would need to be optimized on immobilized beads.

Our cSBL strategy is not truly bi-directional. This is because Endonuclease V cuts in the 3' direction relative to the deoxyinosine position. Therefore, using Endonuclease V for cSBL in the 5' to 3' direction would result in the deoxyinosine remaining in the extended anchor primer. This would limit the number of cSBL cycles in the 5' to 3' direction to two, as attempts to go further will recognize the first incorporated deoxyinosine and limit the extended reads in the 5' to 3' direction.

Conclusions

In summary, we have demonstrated that next-generation sequencing approaches applying the cSBL variation will be able to produce longer read lengths relative to standard SBL. Additionally, cSBL is compatible with and further increases the sequence gains from methods that incorporate additional anchor primer sites. Also, cSBL can complement traditional SBS approaches as cSBL can sequence in the 3' to 5' direction. This variation of traditional SBL approaches has useful applications in many next-generation sequencing methods that are in active use today.

Methods

We have applied cSBL to sequence a known test DNA fragment (Test Template, see Table 1) immobilized on 1.0 um beads (MyOne Beads, Invitrogen) in solution. All

DNA primers used were synthesized by Integrated DNA Technologies. The Test Template was constructed not to have significant secondary structure. The 5' end of the Test Template is modified with a dual biotin on the 5' end to couple to streptavidin-coated beads. The anchor primers (Anchor Primer, see Table 1) were designed to hybridize onto the 5' end of the Test Template, and provide a free 5' phosphate to ligate the query primers (Extension Primers, see Table 1). Multiple anchor primers that were identical except that each progressive primer was shorter by one nucleotide were used. The multiple anchor primers allowed multiple positions to be sequenced with the same set of query. In addition to the query primers, we used a Saturation Primer. The purpose of this was to fully saturate all available ligatable sites, therefore combating drops in signal efficiency and phasing in further cycles. In addition, a standard query primer that did not contain a deoxyinosine was used to sequence the 5th and 10th positions. The 10th position was obtained following a single cycle of cSBL.

Template and Anchor Primers	**DNA Sequence**
Test Template	5′ (Dual Biotin) TCT ATG GGC AGT CGG TGA TAN GCG CTT GCA AGA GAA TGA GGA AAA CGA AGA 3′
Anchor Primer	5′ (Phosphate) A TCA CCG ACT GCC CAT AGA 3′
-1 Anchor Primer	5′ (Phosphate) TCA CCG ACT GCC CAT AGA 3′
-2 Anchor Primer	5′ (Phosphate) CA CCG ACT GCC CAT AGA 3′
-3 Anchor Primer	5′ (Phosphate) A CCG ACT GCC CAT AGA 3′

Extension Sequence Query Primers	**DNA Sequence**
ExSeq4 – A	5′ Cy3 - NNINNANNN 3′
ExSeq4 – T	5′ TYE 665 (Cy5 Analog)- NNINNTNNN 3′
ExSeq4 – C	5′ 6-FAM (FITC Analog)- NNINNCNNN 3′
ExSeq4 – G	5′ TEX 615 (Texas Red Analog)- NNINNGNNN 3′
Saturation Primer	5′ NNINNNNNN 3′

Table 1. Sequences of the Test Template, various Anchor Primers, and Query Primers.

Binding DNA to Beads

The dual-biotin on the test template was bound to the streptavidin-coated beads (MyOne Beads, Invitrogen, Carlsbad, CA). 30 uL of beads were washed three times in Bind and Wash Buffer (10 mM Tris-HCl, 1 mM EDTA, 2.0 M NaCL) and collected using a magnetic particle collector. The beads were then resuspended in 120 uL of BW Buffer and 1.2 uL of 1 mM Test Template sequence (10 uM final concentration) was added incubated at room temperature in a rotisserie for forty-five minutes. Finally, the beads were washed times and resuspended in 60 ul of Wash 1E (10 mM Tris, 50 mM KCl, 2 mM EDTA, and .01% Triton X-100).

Hybridize Anchor Primer onto Template DNA

The beads were washed in Wash 1E (10 mM Tris, 50 mM KCl, 2 mM EDTA, and .01% Triton X-100), then washed once in a 1 × SSPE (150 mM NaCl, 10 mM NaH2PO4, and 1 mM EDTA pH 7.4). The beads were then resuspended in 150 uL 1 × SSPE with 2 uL of 1 mM anchor primer (13 uM final concentration). The solution was incubated at 50°C for 15 minutes and then cooled to room temperature for ten minutes. Lastly, the beads were washed in Wash 1E three times and immediately used in the Query Primer Ligation.

Query Primer Ligation

The beads were collected in resuspended in the ligation buffer (66 mM Tris-HCL, 10 mM MgCl2, 1 mM dithiothreitol, 1 mM ATP, 7.5% Polyethylene glycol [PEG6000]), with a query primer concentration of 3 uM each, and T4 DNA Ligase (2 U/ml, NEB). The ligation reaction was incubated at 30°C for 45 minutes on a rotisserie. Following the reaction the beads were washed three times in Wash 1E and resuspended in Wash 1E. The fluorescent signal was verified using a fluorescent microscope.

Microscope Fluorescent Calibration

The exposure and gain for each fluorescent filter was adjusted with all positions present for each cycle. Camera settings were optimized each cycle of cSBL as signal dropped from one cycle to the next. The individual populations of beads were examined separately with the same settings, and then scored using NIS-Elements Basic Research imaging software (Nikon Instruments Inc, Melville, NY) (Figure 6).

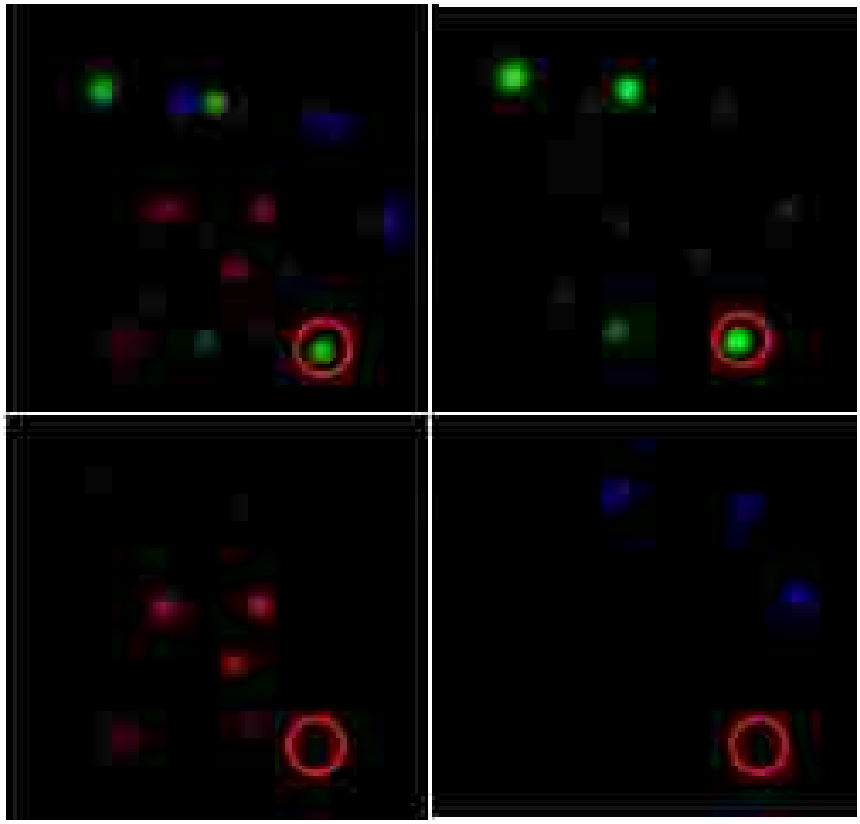


Figure 6. An overlay of three channels of fluorescence. In practice, there are four fluorescent channels, one corresponding to each base. Only three channels and the corresponding overlay are shown here for clarity. NIS-Elements Basic Research 3.0 (Nikon Instruments Inc, Melville, N.Y) software was used to generate this image and analyze the data. Pixel values are taken from beads in each channel to ascertain sequencing accuracy. The pixel values of brightness in each channel are used as a gauge of nucleotide identity.

The pixel values of the brightest channel for a given bead and the values of other channels, provide the signal to noise ratio for comparison.

Pixel Intensity Evaluation as a Measure of Sequencing Accuracy

NIS-Elements Basic Research 3.0 (Nikon Instruments Inc, Melville, NY) was used to determine the pixel intensities in the Cy3, Cy5, FITC, and TxRed channels. Individual channel intensity values ranged from 1-16,383. One-hundred pixels were averaged in each channel and compared. This gave a metric for estimating sequencing accuracy, as the correct signal was known for each position.

Saturation Ligation

A saturation step was performed to fully saturate all Anchor Primers sites not extended during the Query Primer ligation cycle. The ligation was performed in a 1 × T4 DNA Ligase Buffer, with a Saturation Primer concentration of 10 uM and T4 DNA Ligase (2 U/mL), at 30°C for forty-five minutes on a rotisserie.

Endonuclease V Digestion

The beads were washed three times and resuspended in 1 × NEB4 (50 mM Potassium Acetate, 20 mM Tris-Acetate, 10 mM Magnesium Acetate, 1 mM Dithiothreitol) with 100 ug/mL BSA and Endonuclease V at a 2 U/mL concentration. The endonuclease V digestion was incubated at 37 degrees on a rotisserie for ten minutes. Removal of the fluorescence was confirmed visually using a fluorescent microscope. Specific digestion and negligible non-specific Endonuclease V digestion was confirmed by an overnight incubation with Endonuclease V with test-template bound beads. The overnight digestion resulted in no detectable non-specific endonuclease activity when gauged by hybridizing a probe to the distal region of the Test Template.

Endonuclease V Deactivation

Following the Endonuclease V digestion, the beads were extensively washed to remove all Endonuclease V. Enzyme carry forward could cause phasing problems, therefore, a guanidine wash was also performed to inactivate residual enzyme. The bead solution was washed in a 3 M Guanidine solution at room temperature. Following the guanidine wash, the beads were washed three time and resuspended in Wash 1E.

Cyclic Ligation

After Endonuclease V deactivation, the template DNA has been sequenced in one position, but now the anchor primer is effectively lengthened. In traditional SBL, the sequencing strand would be stripped to repeat the sequencing process for a different position. With cSBL, the sequencing of additional bases is dependent upon the preservation of the hybridized sequencing strand of DNA. The process therefore begins again with query primer ligation, and is repeated until the signal to noise ratio is too low to effectively continue sequencing by SBL. At that point, the entire sequencing strand can be stripped and a different length anchor primer can be used to sequence different bases, as in traditional SBL (Figure 7).

Figure 7. Sequencing By Ligation with Endonuclease V Digestion. 1) Sequencing the fourth base in the template tag, by using standard SBL with a Query Oligo that contains a Deoxyinosine (I). 2) Endonuclease V will recognize the Deoxyinosine and cleave the second phosphate bond towards the 3' end. The picture has white light background to make the bead visible as all fluorescence is ablated. 3) Repeat SBL to obtain the next positions.

Authors' contributions

AH, MM, SW, SRA, and JE have all contributed to and participated in drafting this manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the UNM Center for Advanced Research Computing and the UNM Cancer Center Shared Resource for Bioinformatics and Computational Biology for computational resources in support of this work.

This work was supported by the National Institutes of Health [R21 HG004350/564251 and R01HG005852], and the National Science Foundation [DGE-0549500].

REFERENCES

1. McPherson, J.D. (2009) Next-generation gap. *Nat Methods*, **6**, S2-5.
2. Collins, F.S., Morgan, M. and Patrinos, A. (2003) The Human Genome Project: lessons from large-scale biology. *Science*, **300**, 286-290.
3. Lee, C.C., Snyder, T.M. and Quake, S.R. A microfluidic oligonucleotide synthesizer. *Nucleic Acids Res*, **38**, 2514-2521.
4. Pushkarev, D., Neff, N.F. and Quake, S.R. (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol*, **27**, 847-852.
5. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78-81.
6. Yngvadottir, B., Macarthur, D.G., Jin, H. and Tyler-Smith, C. (2009) The promise and reality of personal genomics. *Genome Biol*, **10**, 237.
7. Zhang, W. and Dolan, M.E. Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery. *Pharmacogenomics*, **11**, 249-256.
8. Voelkerding, K.V., Dames, S.A. and Durtschi, J.D. (2009) Next-generation sequencing: from basic research to diagnostics. *Clin Chem*, **55**, 641-658.
9. Nebert, D.W., Zhang, G. and Vesell, E.S. (2008) From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug Metab Rev*, **40**, 187-224.
10. Meyerson, M., Gabriel, S. and Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, **11**, 685-696.
11. Morozova, O. and Marra, M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255-264.
12. Via, M., Gignoux, C. and Burchard, E.G. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med*, **2**, 3.
13. Bell, D.W. Our changing view of the genomic landscape of cancer. *J Pathol*, **220**, 231-243.
14. Tucker, T., Marra, M. and Friedman, J.M. (2009) Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet*, **85**, 142-154.
15. Fredlake, C.P., Hert, D.G., Mardis, E.R. and Barron, A.E. (2006) What is the future of electrophoresis in large-scale genomic sequencing? *Electrophoresis*, **27**, 3689-3702.
16. Bennett, S.T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics*, **6**, 373-382.
17. Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C. *et al.* (2009) The challenges of sequencing by synthesis. *Nat Biotechnol*, **27**, 1013-1023.
18. Pihlak, A., Bauren, G., Hersoug, E., Lonnerberg, P., Metsis, A. and Linnarsson, S. (2008) Rapid genome sequencing with short universal tiling probes. *Nat Biotechnol*, **26**, 676-684.
19. Illumina. (2010).
20. Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S. *et al.* International network of cancer genome projects. *Nature*, **464**, 993-998.

21. Porreca, G.J., Shendure, J. and Church, G.M. (2006) Polony DNA sequencing. *Curr Protoc Mol Biol*, **Chapter 7**, Unit 7 8.
22. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728-1732.
23. Gao, L. and Lu, Z. (2009) The removal of fluorescence in sequencing-by-synthesis. *Biochem Biophys Res Commun*, **387**, 421-424.
24. Biosciences, P. (2010).
25. Xu, M., Fujita, D. and Hanagata, N. (2009) Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small*, **5**, 2638-2649.
26. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.
27. Bloch, K.D. (2001) Digestion of DNA with restriction endonucleases. *Curr Protoc Immunol*, **Chapter 10**, Unit 10 18.
28. Case-Green, S.C. and Southern, E.M. (1994) Studies on the base pairing properties of deoxyinosine by solid phase hybridisation to oligonucleotides. *Nucleic Acids Res*, **22**, 131-136.
29. Allawai, H.T. and SantaLucia, Jr. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* **36**(34): 10581-94
30. Trapnell, C. and Salzberg, S.L. (2009) How to map billions of short reads onto genomes. *Nat Biotechnol*, **27**, 455-457.
31. Medvedev, P., Stanciu, M. and Brudno, M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, **6**, S13-20.

Chapter 4

The Sequence Analysis Workbench: A Framework for Fast, Parallel Genomic Sequence Mapping

by

Maurice H. Murphy[1,2], Susan M. Wilson[2], Antoine Ho[3],

Jeremy S. Edwards[1,3,4], Susan R. Atlas[1,2,5]

[1]UNM Cancer Center, University of New Mexico,

[2]Center for Advanced Research Computing, University of New Mexico,

[3]UNM Department of Molecular Genetics and Microbiology, University of New Mexico,

[4]UNM Department of Chemical and Nuclear Engineering, University of New Mexico,

[5]UNM Department of Physics and Astronomy, University of New Mexico

ABSTRACT

Current sequencing technologies produce RNA and DNA sequence data at very high rates. Generally, downstream analysis requires the intermediate step of mapping of reads to a reference genome. Virtually all next-generation (post-Sanger) sequencing platforms generate giga-base-pairs (Gbp) of data per run, often in the form of mate-paired short–reads (1). We anticipate the daily need to sequence, and subsequently align (map) to a reference genome, several billion mate-pair reads, or single sequence reads, in whole-genome sequencing of human samples. These reads may need to be aligned to a large reference genome, itself comprising several Gbp, e.g. human (~3.0 Gbp), mouse (~ 2.5 Gbp), frog (~ 1.5 Gbp) or zebrafish (~ 1.5 Gbp). An efficient algorithm to perform this mapping is essential given these large dataset sizes. Here we present the SawTooth suite of software applications whose core functionality is the efficient mapping of short-read sequencing data to a reference genome. SawTooth also implements several ancillary applications for validation and statistical analysis of mapping results.

INTRODUCTION

Our initial motivation for developing the SawTooth mapping algorithm was to map data generated on the Polonator polony sequencing platform (http://www.polonator.org). This platform produces data for short-read (typically 13 or 14 bp per tag) paired-end tags (PETs) whose approximate tag separation is known *a priori*, but only to within a broad range of values. Plausible separation intervals may be several hundred or several thousand bp. For example, a Polonator run may produce data where a single read consists of two 13bp tags with a plausible separation range between 700bp and 1200bp. Our target computational platform was a Linux supercomputer consisting of Infiniband-coupled 8-core nodes with 16GB RAM/node. (See Materials and Methods for greater detail.) The resulting algorithm and its implementation were optimized for this core problem, mapping 13mer and 14mer PETs on the targeted computational clusters whose multi-processor nodes provide a minimum of 16GB RAM.

For the core problem, we were able to realize speed-ups on the order of 400x over other currently available and supported mapping software, in particular Bowtie (2), Novoalign (http://www.novocraft.com), and SOAP2(3) . All fast contemporary mapping algorithms, including those just mentioned, rely on indexes. These auxiliary data structures facilitate the alignment of sample sequences to a reference genome, which we refer to as **RefG** in much of the discussion below. These indexes fall generally into two broad categories, suffix-trees and hash indexes (4). Historically, hash-based approaches were the first class of methods to be implemented and are still being used and developed. The well-known BLAST (5) algorithm, developed in 1990, is one example, as are SOAP (6), the more recent Novoalign, and SawTooth, the algorithm that is the subject of this paper. We

first discuss suffix-trees, and their performance, to motivate our reasons for using a hash-based approach.

Traditionally, the construction and use of suffix-trees imposed prohibitive memory requirements, even when represented as suffix-arrays which contain the same information in a more memory-efficient form (7). In recent years, however, innovations in the field of compressed text indexes and associated search methods have rendered suffix-based methods feasible for whole-genome indexing. (Sequence data is a special case of text and is therefore amenable to text-based methods.) Notable examples of compressed-text indexes are based on the Burrows-Wheeler Transform (BWT) (8) of the original text, with subsequent compression and construction of an FM index (9) to enable in-place searching of the compressed text. These form the basis for more recent mapping software such as BWA (10), SOAP2 (3), and Bowtie (2).

A hash index is a well-known referencing data structure which allows key-based data retrieval in constant, O(1), time, making it perhaps the fastest of all data retrieval structures. It comprises three elements, a hash function, a hash table, and a data-store. The hash table is an array of pointers to data elements held in the data-store. This array is accessed directly via an integer offset into the array. As input the hash function takes a key, such as a name or social security number in traditional databases, and returns the integer offset into the hash array which contains a pointer the corresponding data. This integer offset is often called simply the "hash" of the key. Thus, the process of locating the data may be as fast as two memory access operations.

There are some limitations of general hash indexes that may limit their performance or impair their usefulness (11). In the general case, keys are not ordered so sorted lists and

range searches are not intrinsic operations on the data structures. Also, a hash function may generate the same hash for multiple keys. These cases, known as collisions, must somehow be resolved, requiring extra processing and access to the original keys within the index. However, the special nature of genomic data, and our limited purpose of mapping PETs to a reference genome, allow us to create hash indexes that are free from these limitations. In SawTooth, the key is the sequence comprising a tag and the data to be retrieved is an exhaustive list of loci where the tag maps in the reference genome. Importantly, this list is ordered by locus. The hash index and retrieval process is described in greater detail below.

The SawTooth algorithm for mapping a single mate-pair to the reference genome proceeds in two steps:

1. Retrieve two exhaustive lists of loci in the reference genome where each of the two paired-end tags, TAG1 and TAG2, map to the reference genome.

2. Examine all possible combinations of *m* TAG1 candidates and *n* TAG2 candidates. Retain only those that fall within the plausible separation interval.

SawTooth can perform Step 1 in $O(1)$ time and Step 2 in $O(m+n)$ average time, and in $O(m*n)$ worst-case time. The justification for these computational complexity estimates is provided below.

Step 1 – Retrieving lists of loci

SawTooth accomplishes Step 1 in O(1) time by means of a pre-compiled hash index of RefG.

This hash index consists simply of two (4-byte unsigned) integer arrays, one being the hash array (which we refer to as OFFSET_ARRAY in the discussion below), and the other containing the data, i.e. the ordered, exhaustive list of loci (which we refer to as LOCI_ARRAY in the discussion below). It is built for the reference genome (as opposed to the target data) and has these and characteristics and components.

The hash index is based on a pre-defined tag length, 12, 13, or 14 bp. The hash function is defined as simply the binary encoding of the tag sequence. We arbitrarily code the individual nucleotides as follows: A=00, C=01, G=10, T=11. We then create a coding for oligonucleotides by appending the individual codes, e.g. GCAT is coded as 10010011. As well as being an unambiguous binary coding for an oligo, this representation can be directly treated as an integer and consequently as a direct-reference, i.e. as an offset into an array. Tags up to length 16 can be coded into a 4-byte integer hash. Current single – node memory constraints limit the size of the tags used to construct the hash table, i.e. the seed-tags, to a maximum length of 14. This is a "perfect hash function" i.e. there are no collisions and there is no need in the algorithm for collision resolution. It is invertible, i.e., given a hash, the original tag can be uniquely reconstructed.

The data array, LOCI_ARRAY, contains of 4-byte unsigned integers which are the loci where every occurrence of every tag in the reference genome is listed. The hash array OFFSET_ARRAY of 4-byte unsigned integers are the offsets into LOCI_ARRAY where

the first locus in RefG of every possible tag occurs. If the tag occurs nowhere in RefG, then the offset is set to 0xFFFFFFFF, which we use as a "not found marker", NFM.

The complete list of loci where a particular tag occurs is found as follows. The tag is translated into an integer, TAG_HASH, via the hash function, and the element OFFSET_ARRAY[TAG_HASH] is obtained. This is the offset into LOCI_ARRAY where the list of loci for this tag begins. The next tag that occurs in RefG is obtained as TAG_HASH _NEXT. This is usually simply OFFSET_ARRAY[TAG_HASH +1] but is inspected and incremented if it contains a "not found marker". Thus OFFSET_ARRAY[TAG_HASH _NEXT] - OFFSET_ARRAY[TAG_HASH]. This is the number of loci in the reference genome where the tag occurs. The requirements that we have the location in memory where the list of the loci begins, and how many loci there are for this tag, are now completely satisfied.

An example of these data structures for 3-mers appears in Table 1.

Table 1a. Loci Array

Table 1b. Offset Array

Tag Sequence	Loci Array Index	Hash Code	Loci Array
AAA	**0**	**0**	**140**
AAA	1	0	150
AAA	2	0	162
AAA	3	0	143
.	.	.	.
.	.	.	.
.	.	.	.
AAA	74	0	205
AAA	75	0	206
AAA	80	0	5307
AAC	**81**	**1**	**1208**
AAC	82	1	1509
AAC	83	1	2210
.	.		
.	.		
.	.		
AAC	102	1	2884
AAC	103	1	3085
AAC	104	1	3803
AAT	**105**	**2**	**87**
AAT	106	2	981
.	.	.	.
.	.	.	.

Hash Code	Tag Sequence	Offset Array
0	AAA	0
1	AAC	81
3	AAG	105
4	AAT	258
5	ACA	266
6	ACC	283
7	ACG	303
8	ACT	499
9	AGA	513
10	AGC	560
11	AGG	585
12	AGT	645
13	ATA	693
14	ATC	838
15	ATG	900
16	ATT	953
17	CAA	969
18	CAC	1016
19	CAG	1049
.	.	.
.	.	.

Because the hash code is translated directly to the index of a tag in OFFSET_ARRAY there is no need to store either the hash codes or the original tags; only the list of offsets is required. Thus memory usage is calculated as follows: Memory to store LOCI_ARRAY ≈ 4*|RefG|

Memory to store OFFSET_ARRAY = 4*(4^SeedSize). The calculation for LOCI_ARRAY storage is approximate because tags can begin on any base of RefG except for those less that the hash seed size distance from either end of a contig. For the Human

Reference Sequence b36.3 with 2,858,018,193 bases this comes to 11,432,072,772 bytes and is independent of hash seed size.

Table 2. Memory Required for Hashing Data Structures for the Human Reference Sequence b36.3

Hash Seed Size	Number of Possible Distinct Tags (4*4^SeedSize)	Average Loci Per Tag	Memory Required By OFFSET_ARRAY	Total Memory Required
11	**4,194,304**	**715**	**16,777,216**	**11,444,777,216**
12	**16,777,216**	**179**	**67,108,864**	**11,495,108,864**
13	**67,108,864**	**45**	**268,435,456**	**11,696,435,456**
14	**268,435,456**	**11**	**1,073,741,824**	**12,501,741,824**
15	**1,073,741,824**	**2.794**	**4,294,967,296**	**15,722,967,296**
16	**4,294,967,296**	**0.698**	**17,179,869,184**	**28,607,869,184**
17	**17,179,869,184**	**0.175**	**68,719,476,736**	**80,147,476,736**

Step 2 – Filtering pairs with plausible separation

A naïve approach to filtering tag-pairs with plausible separation would be to actually examine every one of the *m*n* combinations. This, however, is unnecessary. SawTooth accomplishes Step 2 in *O(m+n)* average time by exploiting the fact that the two lists of loci are both ordered by locus. This ordering has two consequences.

1- When looking for the mates for particular TAG1 locus, there is no need to examine any TAG2 loci after encountering the first TAG2 locus that exceeds the maximum separation. Because they are ordered, all subsequent TAG2 loci in the TAG2 locus list are greater still, and must also exceed the maximum separation.

2- Once the plausible TAG2 mates for a particular TAG1, say TAG1 locus(i), have been found, the search for mates to the next TAG1 locus , TAG1 locus(i+1) may begin with the first TAG2 locus hat was a mate for TAG1 locus(i). In the case where TAG1 locus(i) had no mates, the search can begin with the last TAG2 locus examined which was the first which exceeded the maximum separation from TAG1 locus(i) .

These two constraints allow, in the case of typical tags, for a single traversal of each the two loci lists, with occasional re-examination of TAG2 loci, and requiring *O(m+n)* comparisons. The worst case *O(m*n)* occurs when every TAG2 is a plausible mate for every TAG1 as would happen in a long sequence of replicated monomers. This situation is

easily handled by terminating the search when a maximum number of matches have been found.

Extension of Paired-End Mapping to the Mapping of Longer Tags

The algorithm described above has been extended in SawTooth to handle tags of arbitrary length, whether as single tags or paired-end tags of length greater than 14. The extension based on the fact that the result of pair separation filtering is an ordered list of loci, just as is the result of a search in the hash-index. With that in mind, the extension to longer tags is obvious and direct. Longer single tags are simply split into sub-tags with the same length as the seed, and treated as mate-pairs with a separation of 0 for contiguous sub-tags, or negative separation for overlapping sub-tags. For example, using an index seed length of 14, a longer tag of length 35 is split into 3 sub-tags beginning at positions in the original tag of 1, 15, and 21 as in the representation below.

```
Tag: ACGTACGTACGTACGTACGTACGTACGTACGTAC
T01: ACGTACGTACGTAC
T15:                GTACGTACGTACGT
T21:                       CGTACGTACGTAC
```

The lists of loci for T15 and T21 are filtered for a separation of exactly -7. The resulting list of loci is filtered with the T01 list for a separation of exactly 0. For longer tags yet, the process is repeatedly applied.

The list of loci for single tags produced this way are de-referenced and reported as mapping results. For paired-end tags, the lists of loci are then submitted to the pair separation filtering algorithm.

Other Tools in the SawTooth Toolkit

In addition to the core functionality described above, SawTooth implements several other tools for the processing, validation and analysis of mapping results. These include:

- Building, saving and loading hash indexes based on multiple FASTA files and multiple-section FASTA files whether genomic or transcriptomic.
- Generating sample tags and tag-pairs from a reference sequence
- Validating results for correctness against the reference genome
- Comparison of results generated by different mapping software
- Analysis of mapping results to determine depth of coverage
- Translation between various formats of input and results files

MATERIALS AND METHODS

ANSI C/C++ was chosen as the core development computer language for SawTooth for several reasons. It is available in the Microsoft Visual Studio environment, which provides a convenient development platform and is portable to our target Linux production architecture. Second, the language modularity and the availability of stubs for the Message Passing Interface (MPI) specification (12) led to code that is easily extensible with respect to its algorithmic and parallel (multi-node) capabilities. Third, C++ provides intrinsic high level data-structures such as vectors, lists, maps, and strings, along with fast and reliable algorithms that operate on them, through the Standard Template Library (STL). Finally, strict adherence to the ANSI language standard ensures that that SawTooth

will be portable across a great variety of platforms. SawTooth is currently implemented on both Intel and PowerPC hardware architectures. It runs under the Linux, Windows, and AIX operating systems; and has been successfully compiled with Microsoft Visual Studio, the GNU G++ compiler, the Cygwin G++ compiler, the Intel ICC compiler, and the IBM XLC compiler.

On-node multi-processor performance was implemented through the OpenMP multi-threading library. Multi-node, distributed computing enhancements, primarily for segmented hash tables, are under development using the MPI.

Performance benchmarking was performed on a 1792 node/14,336-processor-core, 172 TFlop SGI/Intel Altix ICE 8200 supercomputer (Intel Xeon 3.0 GHz, 64-bit architecture; 8 cores/node, 16 GB RAM/node; high-speed InfiniBand and IPC networks). Code validation was performed on a smaller version of this architecture (22 nodes). Timings for all benchmarks reported below are based on elapsed wall-time, expressed in seconds. All tests were performed via batch submission on dedicated nodes, with no processes running other than the mapping software and system processes.

RESULTS

Benchmarks for the purpose comparison of search times among SawTooth, Bowtie, and Novoalign were performed in identical execution environments. Except for the benchmarking with respect to number of processors each application was run simultaneously on 8 nodes, utilizing all 8 processors (cores) on each node.

Validation Test 1: checked 2 million mappings, exact correspondence with Novo some differences with Bowtie.

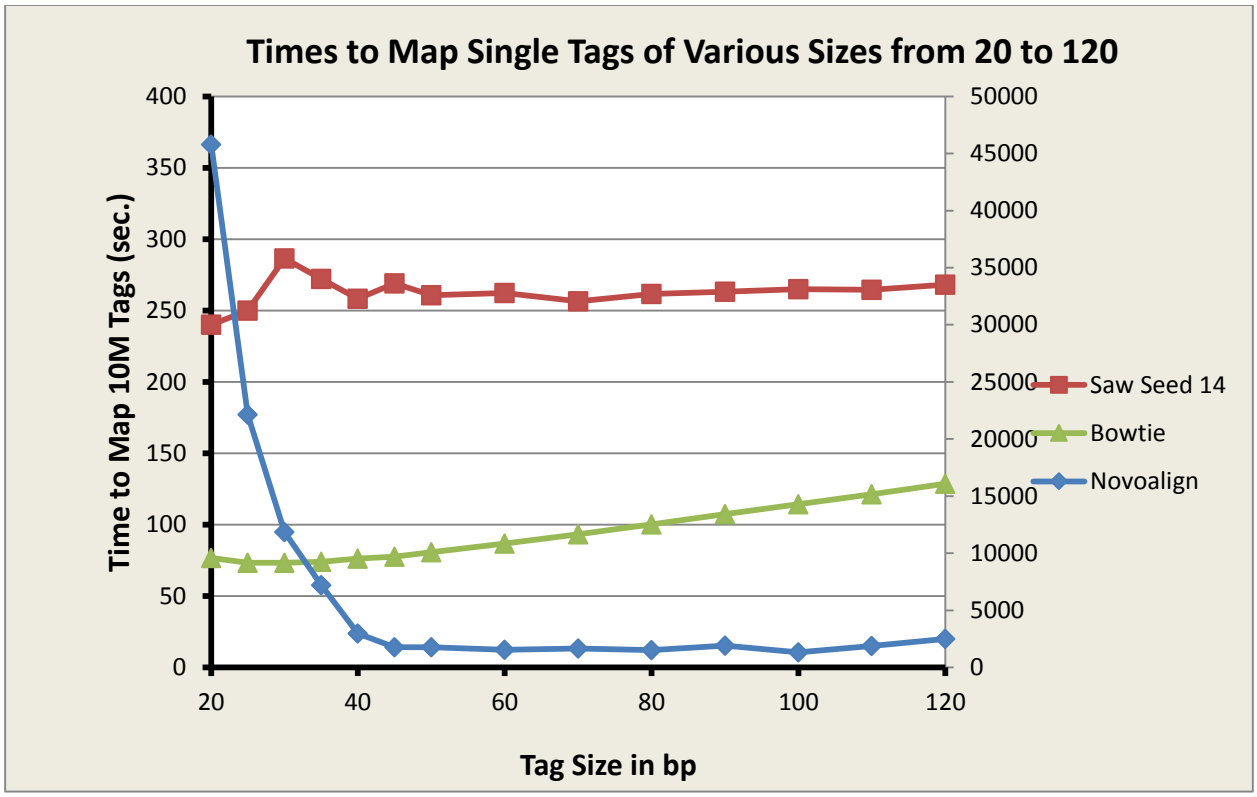
Benchmark Test 1: search times to map 10M paired-end tags of various sizes 13-20

Tag Size (bp)	SawTooth (sec.)	Novoalign (sec.)	Bowtie (Seconds)	Speedup Relative to Novoalign	Speedup Relative to Bowtie
14	463		152588	0	329
15	558		135556	0	243
16	519	184003	100919	355	194
17	479	153142	70849	319	148
18	448	126530	51069	282	114
19	428	103814	38754	242	90
20	409	86302	31236	211	76

Table 1. Bowtie and SawTooth search times (sec.) to map 10M paired-end tags of various tag sizes. The SawTooth index was built with seed size 14.

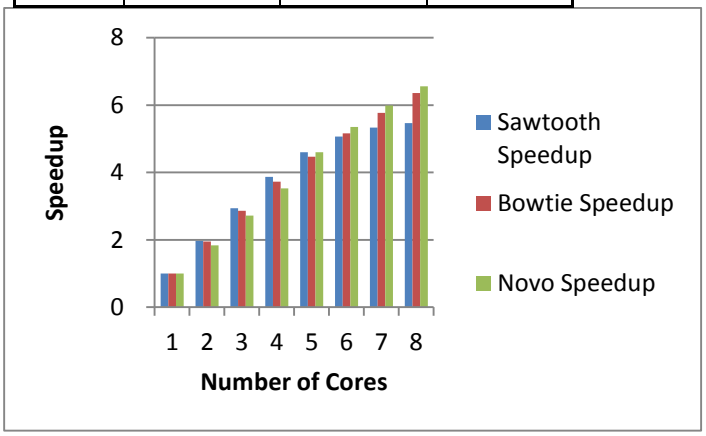
Benchmark Test 2: Search times to map 10M single tags various sizes 20-120

Single Tag Size (bp)	SawTooth (sec.)	Novoalign (sec.)	Bowtie (sec.)	Speedup Relative to Novoalign	Speedup Relative to Bowtie
20	240	45792	77	190.8	0.32
25	250	22141	73	88.6	0.29
30	286	11861	73	41.4	0.26
35	272	7195	74	26.4	0.27
40	258	2979	76	11.5	0.30
45	269	1763	78	6.6	0.29
50	261	1763	81	6.8	0.31
60	262	1556	87	5.9	0.33
70	256	1644	93	6.4	0.36
80	262	1519	100	5.8	0.38
90	263	1907	107	7.2	0.41
100	265	1330	114	5.0	0.43
110	264	1866	121	7.1	0.46
120	268	2501	129	9.3	0.48



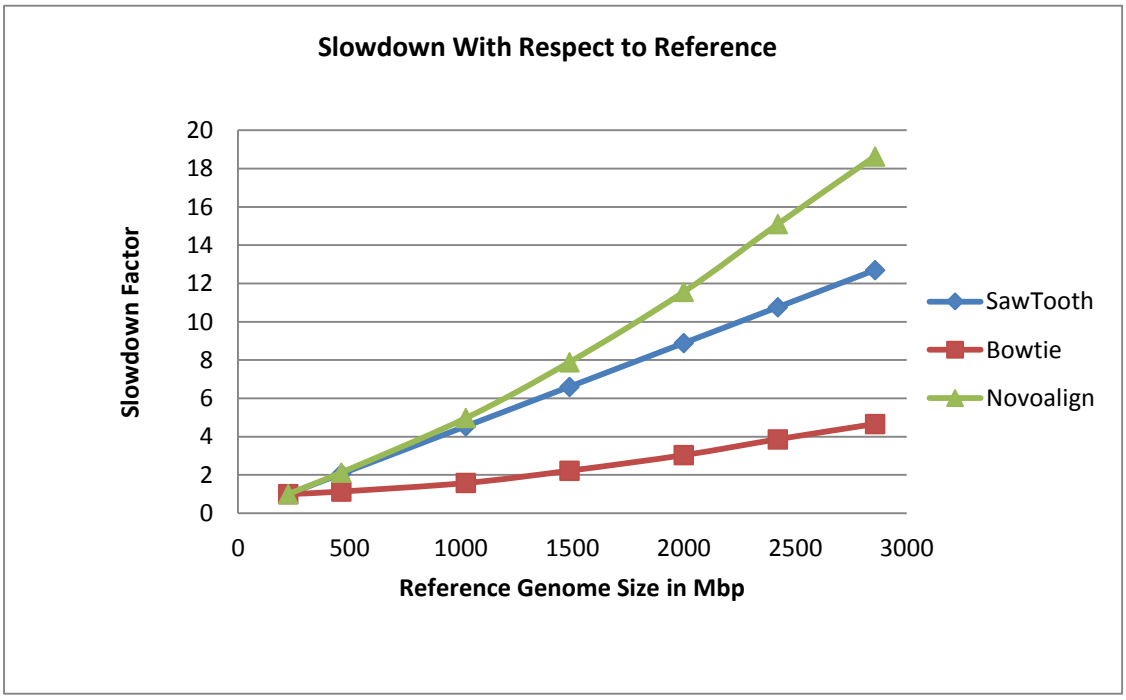
Benchmark Test 3: Search times to map 1M paired-end tags size 16 using 1-8 cores

Cores	Sawtooth Speedup	Bowtie Speedup	Novo Speedup
1	1.0	1.0	1.0
2	2.0	1.9	1.8
3	2.9	2.9	2.7
4	3.9	3.7	3.5
5	4.6	4.5	4.6
6	5.1	5.2	5.4
7	5.3	5.8	6.0
8	5.5	6.4	6.6



Benchmark Test 4: search times to map 10M paired-end tags size 16 to various sized genomes, where maps are known to exist a priori (adding successive subsets of chromosomes that add up 300Mbp increments to chr1)

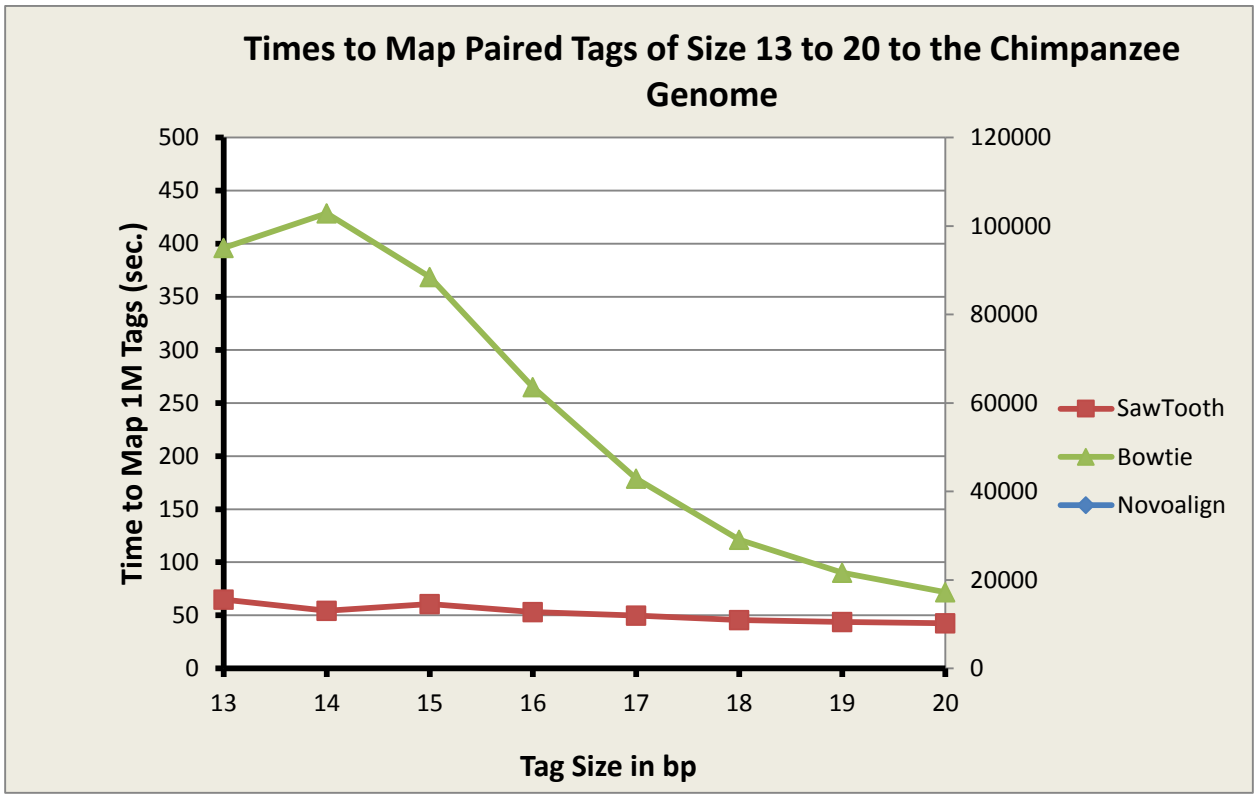
Chromosomes in Reference Sequence	Reference Genome Size in Mbp	SawTooth (sec.)	Bowtie (sec.)	Novoalign (sec.)	SawTooth Slowdown Factor	Bowtie Slowdown Factor	Novoalign Slowdown Factor
1 to 1	225	86	5418	14731	1.0	1.0	1.0
1 to 2	463	97	11526	26861	2.1	1.1	2.1
1 to 5	1022	135	26953	55601	4.5	1.6	5.0
1 to 8	1487	190	42766	83436	6.6	2.2	7.9
1 to 12	2000	260	62652	117471	8.9	3.0	11.6
1 to 17	2422	331	81854	151096	10.8	3.9	15.1
1 to 24	2858	399	101009	184003	12.7	4.7	18.6



Benchmark Test 5: search times to map 1M paired-end tags sizes 13 to 20 to the chimpanzee genome, ~3Gbp also.

Tag Size (bp)	SawTooth (sec.)	Novoalign (sec.)	Bowtie (sec.)	Speedup Relative to Novoalign	Speedup Relative to Bowtie
13	64.8		95084.0		1466.6
14	54.2		102877.7		1896.9
15	60.6		88469.3		1459.1
16	53.2		63629.7		1195.3
17	49.7		42897.0		862.5
18	45.6		29075.7		638.1
19	43.7		21675.0		495.6

20	42.6		17267.3		405.3



DISCUSSION

While much recent work is devoted to developing suffix-tree based mappers, hash-based mappers still have the potential for superior performance. SawTooth is one such application that provides a significant speed up over current suffix-tree based mappers and previous hash-based mappers.

In addressing the core problem for which SawTooth was developed, i.e. the exact mapping of short (13mer to 20mer) paired end tags to a large (~3Gbp) reference sequence, it provides speedups of 76 to several hundred over other popular mapping software packages (Bowtie and Novoalign)

Spaced-seed indexing (14) can greatly increase the specificity of mapping longer tags and improve performance in that area where SawTooth lags. (15)(16)(17). The seed sequences on which the hash index is based need not be contiguous. In the simplest implementation of a spaced-seed index the seeds are composed of every n^{th} base in the reference sequence, perhaps every third or fourth. A spaced-seed index based on every fourth base and of net seed length 14 would span a region of 43 (4*13 +1) bases in the reference sequence.

The next phase of development will be to implement SNP/SNV discovery. This may be accomplished by one of several different methods. SawTooth is still in an early stage of development and these directions for future development seem very promising. We believe that the SawTooth algorithm may be improved. We further believe that the methodology is so powerful that it may serve as the kernel of hybrid algorithms incorporating various other methods and concepts. Possible areas of future efforts are as follows.

SNP discovery can be accomplished immediately by simply permuting each base in the target tags and repeatedly applying the existing algorithm. This brute force approach is computationally intensive; permuting each base through three alternates in paired-end 14mers requires 84 (3x28) iterations of the algorithm. However the speed of SawTooth makes this approach feasible.

In a more promising approach, the very fast exact mapping of 14-mer pairs could be used to provide small sets of candidate sequence alignments which may then be analyzed by more comprehensive yet much slower homology scoring algorithms such as Smith–Waterman which requires *O(mn)* time and space. This is an approach similar to that used by the BLAST algorithm for generalized sequence alignments. This approach would lead directly to the detection of indels, (insertions and deletions (18) as well as polymorphisms.

FUNDING

This work was supported by the National Institutes of Health NCE [# R21 G004350/564251], and the National Science Foundation [DGE-0549500]."

REFERENCES

1. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46, 10.1038/nrg2626.
2. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25, 10.1186/gb-2009-10-3-r25.
3. Li,R., Yu,C., Li,Y., Lam,T.-W., Yiu,S.-M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966 -1967, 10.1093/bioinformatics/btp336.
4. Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, **11**, 473 -483, 10.1093/bib/bbq015.
5. Altschul, S, Gish, W, Miller, W, Myers, E and Lipman, D (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403–410, 10.1006/jmbi.1990.9999.
6. Li,R., Li,Y., Kristiansen,K. and Wang,J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713 -714, 10.1093/bioinformatics/btn025.
7. Navarro,G. and Mäkinen,V. (2006) Compressed full-text indexes. *ACM COMPUTING SURVEYS*, **39**, 2007.
8. D,B.M. and W. and Burrows M and Wheeler D (1994) A block sorting lossless data compression algorithm Technical Report 124, Digital Equipment Corporation Available at: http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.html.
9. Ferragina,P. and Manzini,G. (2000) Opportunistic data structures with applications. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, Washington, DC, USA, p. 390–. Available at: http://portal.acm.org/citation.cfm?id=795666.796543 [Accessed May 25, 2011].
10. Li H,D.R. and Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler Transform. *Bioinformatics*, **25**, 1754–1760, 10.1093/bioinformatics/btp324.
11. Cormen,T.H., Leiserson,C.E., Rivest,R.L. and Stein,C. (2009) Introduction to Algorithms third edition. The MIT Press.
12. Walker,D.W., Walker,D.W., Dongarra,J.J. and Dongarra,J.J. (1996) MPI: A Standard Message Passing Interface. *Supercomputer*, **12**, 56–68.
13. Navarro,G. and Raffinot,M. (2007) Flexible Pattern Matching in Strings: Practical On-Line Search Algorithms for Texts and Biological Sequences 1st ed. Cambridge University Press.
14. Trapnell,C. and Salzberg,S.L. (2009) How to map billions of short reads onto genomes. *Nat Biotechnol*, **27**, 455-457, 10.1038/nbt0509-455.
15. Brejova,B., Brown,D.G. and Vinar,T. (2003) Vector Seeds: An Extension to Spaced Seeds Allows Substantial Improvements in Sensitivity and Specificity. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.3097.
16. Xu,J., Brown,D., Li,M. and Ma,B. (2004) Optimizing Multiple Spaced Seeds for Homology Search. *IN: PROCEEDINGS OF THE 15TH SYMPOSIUM ON*

COMBINATORIAL PATTERN MATCHING. VOLUME 3109 OF LECTURE NOTES IN COMPUTER SCIENCE, **2004**, 47--58.
17. Choi,K.P., Zeng,F. and Zhang,L. (2004) Good spaced seeds for homology search. *Bioinformatics*, **20**, 1053-1059, 10.1093/bioinformatics/bth037.
18. Kondrashov AS,R.I. and Kondrashov AS, Rogozin IB (2004) Context of deletions and insertions in human coding sequences. *Hum. Mutat.*, **23**, 177–85, 10.1002/humu.10312.
19. Cichelli,R.J. (1980) Minimal perfect hash functions made simple. *Commun. ACM*, **23**, 17–19, 10.1145/358808.358813.

Chapter 5

Paired-End Library Construction

Introduction

The initial goal of the project described in this chapter, focused on sequencing the bacteria, *Actinomyces kibdesporangium*, culminating in a finished *de novo* draft assembly. Sequencing was performed using an IonTorrent Personal Genome Machine (PGM). The project subsequently shifted focus to the library construction and optimizing a library construction protocol that would allow first paired-end and then mate-paired reads to be obtained with the IonTorrent PGM, making *de novo* draft assembly more accurate and complete (1-7).

Actinomyces kibdesporangium does not have a complete reference genome available yet, with only a single gene cluster having been sequenced thus far. The laboratory was approached by a chemistry lab, interested in biochemical pathways the bacteria is capable of.

Library construction is an important precursor to sequencing, not only to produce the DNA templates to be sequenced. Additionally, this process can dictate what sequencing data will be gained. Sequencing By Synthesis (SBS) must proceed in the 5' to 3' direction due to the usage of the enzyme DNA polymerase (8-10). Therefore, even with the most straightforward sequence library construction methods, one can only obtain sequence data at one end of a DNA tag (11-13). The limits of most next-generation sequencing methods that utilize SBS reach around 200 hundred bases due to the multi-parallelization that occurs (9,11,14), resulting in loss of chemical efficiency for each

individual step. Traditional Sanger sequencing, considered the "original" SBS method, has much lower throughput and can produce reads on the order of roughly one kilobase (9,15-17). Miniaturization reduces the efficiency of the biochemistry, resulting in a significant loss of read length, however, there is an exponential increase in the number of reads that can be sequenced, vastly increasing overall throughput (17,18). Different library construction methods can consider these limitations and be tailored to obtain the most useful data possible.

Sequencing data can be either non-paired or mate-paired. A non-paired library provides only a single read from each tag, while a mate-paired library produces tags that are paired, but separated in the genome by a known range of distances (9,14,17). This distance makes it possible, using computational methods, to assemble large genomes that may have redundant regions, and offers a greater potential for resolving ambiguous matches (19). Though mate-paired reads are always more useful in the assembly of large genomes, there are still reasons not to use mate pairs. Most re-sequencing of human genomes can still provide very valuable data without mate-pairs because a reference genome can be used to determine SNPs or indels (20). Even for *de novo* sequencing projects, the genome may be small enough to be assembled without using mate-pairs (11,20,21). The amount of labor and optimization that goes into mate-paired library construction can be deterring factors for making a mate-paired library, especially if one is not necessary.

This project sought to optimize a library construction method that would make it possible to obtain paired-end reads on the IonTorrent, a next-generation sequencing machine that utilizes SBS. This would potentially allow optimization and application of a

mate-paired sequencing approach in the future. In this chapter, we describe the procedures used and preliminary results from this work. However, preliminary paired Ion Torrent data suggested our current protocol is not usable for generating paired-end data, as only a percentage of the supposedly paired reads were actually paired, making utilization of the data problematic.

Paired-Library Assessment

The paired-end library was constructed in a small number of runs (in Methodology below). A paired-end run actually consists of two runs, a forward run followed by a reverse run. In all cases, the reverse run contained fewer beads, or reads, than the forward run. This was to be expected due to lack of efficiency in various enzymatic steps to prepare the template DNA for the run. The IonTorrent PGM has millions of wells that the machine can track between runs. Therefore, it is simply a matter of connecting reads obtained in the first forward run, and pair each with the read from the corresponding well in the reverse sequencing run. The main issue was was whether every pair of reads from a single well would actually be paired, and therefore verification of the data needed to be performed.

Initial verification of successful paired-end sequencing was performed by using a simple perl script (M. Murphy) that checked for overlap in the two supposedly paired sequences. A pair of runs, a forward and a reverse sequencing run, was performed on a bacterial genome, *Actinomyces kibdesporangium.* Template DNA had been size-selected via gel electrophoresis prior to ligation of adaptors, therefore confidence that the fragment size was in the 200-250 base range was high. For our sequencing run, we used the Ion PGM™ 200 Xpress™ Template Kit (Cat 4474280), which creates reads roughly 200 bases

long. Therefore, we expected to obtain paired-end reads that would be overlapping. The data is output into three files; one forward, one reverse, and one unpaired file. We only analyzed the first two, which contained roughly 100,000 reads. We performed a simple check for paired-end reads by taking the last twelve bases of the forward read, and then looking for a reverse-complement match in the supposedly paired-read in the second file. We obtained a match roughly a quarter of the time (Table 1).

Total Pairs of Reads	**96,102**
Pairs with Overlapping Sequence (Paired)	21,039 (22.02%)
Pairs with No Overlapping Sequence (Unpaired)	74,521 (77.54%)

Table 1 – Reads evaluated with perl script for overlapping sequence as an indicator of paired reads.

We then decided to follow up and check to see if there were any paired-end reads we did not catch due to the fact that the paired-end reads were too short to overlap, although we predicted that this would be a fairly low percentage of our reads. We used Bowtie 2 (22) and a very rough draft genome of *Actinomyces kibdesporangium*, that had been assembled by collaborators with over 2 gigabases worth of IonTorrent sequencing information from another assembler, CLC Bio. Bowtie 2 is an assembler that can match paired-end reads to a reference and then output statistics on what percentage of the reads were paired-end. Bowtie 2 also distinguishes between cases where the paired reads were concordant or discordant, meaning that they were separated or overlapped, respectively. Using the same dataset as our initial perl script check, we were able to obtain the following results, which somewhat affirms our quick perl script results, that only a percentage of the paired reads were truly providing paired data (Table 2).

Unpaired Reads (From Unpaired Pairs above)	149042
Did Not Align to Reference	90,266 (60.56%)
Aligned Once	57,389 (38.51%)
Aligned More than Once	1387 (0.93%)

Table 2 – Bowtie 2 results, using rough draft as reference and mapping paired-end reads to verify paired-end status.

However, in this particular dataset, we also discovered that the reads themselves were not aligning properly to our reference genome (Did Not Align to Reference). Ideally, the reads would map only once (Aligned Once), making evaluation possible as reads that mapped ambiguously in multiple areas (Aligned More Than Once) cannot be used regardless. Given the read lengths, we expected a majority of the reads to map once, given a reliable reference genome. If anything, these results gives more insight into the current draft and sequencing methods than it does the paired-end library. However, as seen from Table 2, 61% of the dataset was unusable because it did not even match in the reference, meaning Bowtie 2 could not be used to actually determine the rate of paired-end reads. Therefore, it was determined that only reads that aligned to the genome once would be used. The percentage of paired reads to *eligible* reads that actually matched to the rough draft genome was roughly 41.8% of the reads paired (Table 3).

Total Reads that Aligned Once	101,938
Paired Pairs	42,078 (41.8%)
Unpaired Pairs	58,776 (57.6%)

Table 3 – A subset of only reads that aligned exactly once according to Bowtie 2, evaluated for overlap, indicating paired-end reads.

The initial expectation for this paired-end library was that most of the reads present in the reverse run would be truly paired, instead, less than half of the reads ended up being paired. This shortfall (Table 3) is problematic because in any *de novo* sequencing effort, there is no way to ascertain which reads in the reverse run are truly paired . Even with resequencing efforts, where a reference genome would be available, information involving structural variations or rearrangements would be lost because none of the supposedly paired-end data is reliable.

Future work on this project will require further optimization of the paired-end library construction method, as it will be vital in enhancing a mate-paired library construction method.

Methodology

Paired-End Library Construction

All protocols were adapted from IonTorrent's Demonstrated Protocol: Paired-End Sequencing on the PGM™ System (23).

The IonTorrent is a next-generation sequencing machine from Life Technologies (24). It utilizes an SBS method that obtains data in real-time by using pH sensors to detect the incorporation of nucleotides. General library construction involves shearing double-stranded DNA, and then ligating adaptors onto both ends of the DNA. The DNA template is then single-stranded and put through an emulsion PCR (ePCR) method, that will clonally cover polystyrene beads with the DNA template. These beads are then put into wells in a specialized chip that is covered in pH sensors. After this, SBS will proceed from the 5' to

3’ end, obtaining a single read from each well – resulting in millions of beads, each providing 150-250 bases of data.

One of the biggest changes in protocol from single-end reads to paired-end reads involves adding an Nt.BbvCI site onto the Sequencing Primer. After sequencing the forward reads routinely, the Nt.BbvCI enzyme is added along with T7 Endonuclease. Nt.BbvCI will nick the recognition site, while T7 Endonuclease digests all DNA that is 3’ of the nick. This results in the original template strand being digested away while the newly synthesized sequencing strand remains. The nick is far enough 3’ to spare the 5’ side DNA, which can then prime a follow-up SBS run, but now running in the opposite direction, creating a paired-end read.

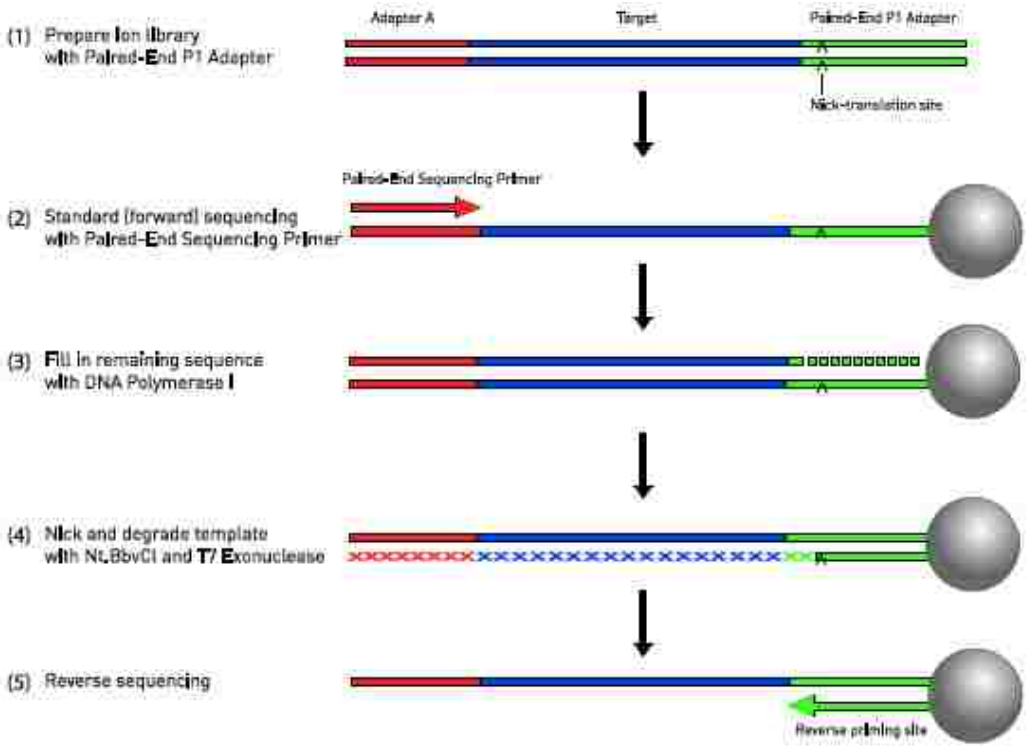


Figure 1 – A general summary of the paired-end sequencing strategy from (23).

Primer Construction

The primers themselves are essentially identical the default primers used however, there is a Nt.BbvCI site incorporated in the Adaptors and the Sequencing Primer was altered to compensate for this. Phosphorothioate bonds are used to prevent digestion. P1 adaptors will be adhered to the bead, whereas the P2 adaptors will be free end of the DNA, that will later prime with the Paired-End Sequencing Primer.

Oligo Name	Sequence
Paired-End Sequencing Primer	5′ – C*C*A*T* CTC ATC CCT GCG TGT CTC CGA C – 3′ * = Phosphorothioate bond.
P1 Adaptor 1	5′—CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATCCTCAGC —3′
P1 Adaptor 2	5′— GCTGAGGATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTGG*T*T —3'
P2 Adaptor 1	5′— CCATCTCATCCCTGCGTGTCTCCGACTCAG —3'
P2 Adaptor 2	5′— CTGAGTCGGAGACACGCAGGGATGAGATGG*T*T —3'

Table 1- Primer construction.

DNA Template Preparation

The genomic DNA was first fragmented, and adaptor ligated. A 200 base-read library was prepared using Ion Xpress™ Plus Fragment Library Kit (Life Technologies. Cat - 4471269). The only alteration was ligating the above P1 and P2 adaptors instead of the default adaptors.

Next, this template was clonally amplified to cover polystyrene beads via an emulsion PCR method. This was performed with a Ion Xpress™ Template 200 Kit (Cat - 4471253). This was completed with no alterations to the written protocol.

Forward Sequencing

Reagents and the protocol used for forward sequencing were performed as in the Ion Sequencing 200 Kit (Cat - 4471258). The only change to the protocol was substituting in the Paired-End Sequencing primer, as dicussed above. Forward sequencing was performed as usual.

Reverse Sequencing

Reagents and the protocol used for reverse sequencing were as in the Demonstrated Protocol: Paired-End Sequencing on the PGM™ System (Part Number – MAN0006191).

References

1. Ho, A., Murphy, M., Wilson, S., Atlas, S.R. and Edwards, J.S. (2011) Sequencing by ligation variation with endonuclease V digestion and deoxyinosine-containing query oligonucleotides. *BMC genomics*, **12**, 598.
2. McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, **19**, 1527-1541.
3. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, **309**, 1728-1732.
4. Young, A.L., Abaan, H.O., Zerbino, D., Mullikin, J.C., Birney, E. and Margulies, E.H. (2010) A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome research*, **20**, 249-256.
5. Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome research*, **20**, 1165-1173.
6. Li, Y., Hu, Y., Bolund, L. and Wang, J. (2010) State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human genomics*, **4**, 271-277.
7. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, **20**, 265-272.
8. Gao, L. and Lu, Z. (2009) The removal of fluorescence in sequencing-by-synthesis. *Biochemical and biophysical research communications*, **387**, 421-424.
9. Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C. *et al.* (2009) The challenges of sequencing by synthesis. *Nature biotechnology*, **27**, 1013-1023.
10. Hamilton, S.C., Farchaus, J.W. and Davis, M.C. (2001) DNA polymerases as engines for biotechnology. *BioTechniques*, **31**, 370-376, 378-380, 382-373.
11. Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics*, **11**, 31-46.
12. Bashir, A., Volik, S., Collins, C., Bafna, V. and Raphael, B.J. (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS computational biology*, **4**, e1000051.
13. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, **318**, 420-426.
14. McPherson, J.D. (2009) Next-generation gap. *Nature methods*, **6**, S2-5.
15. Yngvadottir, B., Macarthur, D.G., Jin, H. and Tyler-Smith, C. (2009) The promise and reality of personal genomics. *Genome biology*, **10**, 237.
16. Kato, K. (2009) Impact of the next generation DNA sequencers. *International journal of clinical and experimental medicine*, **2**, 193-202.

17. Bennett, S.T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics*, **6**, 373-382.
18. Morozova, O. and Marra, M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255-264.
19. Trapnell, C. and Salzberg, S.L. (2009) How to map billions of short reads onto genomes. *Nature biotechnology*, **27**, 455-457.
20. Paszkiewicz, K. and Studholme, D.J. (2010) De novo assembly of short sequence reads. *Briefings in bioinformatics*, **11**, 457-472.
21. Henson, J., Tischler, G. and Ning, Z. (2012) Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, **13**, 901-915.
22. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**, R25.
23. IonTorrent. (2012) Paired-End Sequencing on the PGM System.
24. Zhao, J. and Grant, S.F. (2011) Advances in whole genome sequencing technology. *Current pharmaceutical biotechnology*, **12**, 293-305.

Chapter 6

De Novo Assembler Comparison

Introduction

The process of *de novo* genome assembly is a complicated affair. There is a lack of true verification of published draft genomes that will inevitably see many revisions (1-3). There are many programs available for assembly, all of which vary in their algorithms, efficiencies, and effectiveness for various next-generation sequencing technologies, and the list of assemblers is still growing every day (3). The process of choosing an assembler and optimizing is a challenge that has growing complications by the day.

De novo assembly and mapping assembly are two distinct tasks. *De novo*, as the term implies, is from nothing, and refers to the assembly of a genome utterly from scratch (1,3-6). On the other hand, mapping assembly takes next-generation sequencing data and compares it to an available reference genome (1,7,8). The information gained from each endeavor varies as well. *De novo* assembly culminates in the completion of a draft genome, providing basic genome structure, gene synteny, and protein information (2,5,9). A mapping assembly requires a reference genome, but can provide insight into differences between sample genomes and the reference genome. These differences can be structural variations, Single Nucleotide Variations and Indels, expression differences, epigenetic factors, or copy number variations (10-14). These variations, in turn, can be potential contributing factors of phenotypes or disease states of interest.

Preliminary data regarding the *de novo* assembly of *Actinomyces kibdesporangium*, as described in chapter 5, prompted a need to evaluate assemblers. An issue complicating

the generation of a *de novo* draft quality genome is the lack of verification. One can provide evidence of a strong dataset, provide quality control data regarding the acquired sequence data, but there is no method to directly verify the accuracy of a draft genome (15,16). Often, published draft genomes undergo multiple revisions as the genome is re-sequenced and reassembled, gaining more detailed annotations as further painstaking work is performed to verify the genome (17,18). Therefore, one cannot simply generate a draft genome using multiple assemblers and determine from those results which draft genome is the most accurate. Granted, many programs include internal checks that can provide various relevant metrics such as contig sizes and coverage, but there is no way to measure mistakes in alignment or base calling.

This work attempts to evaluate the capabilities of three popular *de novo* assemblers by assembling a well-characterized reference genome: *Escherichia coli* K12 MG1655. The assemblers tested are MIRA (Mimicking Intelligent Read Assembly), CLC Genomics Workbench, and Velvet Assembler (19,20). In addition to testing the assemblers, experimental data from Ion Torrent, Illumina / Solexa, and Roche 454 were obtained for E. coli K12 MG1655. These sequencing data were put through the various packages to create multiple *de novo* assemblies. These assemblies were then compared to the reference genome using the Mauve Multiple Genome Alignment program (21), which gave data regarding contig length, percentage of genome covered, as well as misalignments, gaps in alignment, and incorrect base calls (22). Using a reference genome provided an opportunity to evaluate *de novo* assemblies, which would not be possible when comparing true *de novo* sequencing efforts (21,23,24).

Results

As expected, there was no clear, undisputed best performing assembler. Choice of assembler will depend on largely on the data on hand to be assembled. Generally speaking, relying purely on contig length, Roche 454 data is best assembled on CLC Genomic Workbench, Ion Torrent is best assembled on MIRA, and Illumina / Solexa data generates the longest contigs using Velvet Assembler.

MIRA seemed to perform very poorly with Solexa reads, which are mate-paired. However, when given longer reads, as in the case of Roche 454 and Ion Torrent data, Mira did much better, and was best in assembling Ion Torrent data, which provided 300-500 base read lengths. This is in spite of the fact that the data is prone to homopolymer errors, where there are multiples of the same base in a row. In general, yields lower quality reads than Solexa, though there are typically a greater number of reads. This implies that with MIRA, the size of read lengths, assuming adequate data quality, is an important factor for obtaining higher quality assemblies.

Velvet Assembler, in contrast, performed very poorly with longer reads, especially if the accuracy of the reads were lower. Roche 454 data available was of higher quality, but shorter than Ion Torrent data. Velvet assembled 454 data poorly, but when given Ion Torrent data, Velvet Assembler could not even complete an assembly, running out of memory even on a machine with 16 GB of RAM, implying that the algorithms that Velvet Assembler used created too many contigs that were very small in size, and overly taxed the computer system. However, when given very accurate, shorter data from Solexa, Velvet was able to create the longest contigs out of any of the datasets. It seems that the most

important factor for Velvet Assembler is the accuracy of reads, and that the algorithms Velvet Assembler employs for dealing with less accurate reads are ineffective.

CLC Genomics Workbench seems to fall somewhere in between the niche specialties of Velvet Assembler and MIRA. CLC was able to generate the longest contigs with 454 data, which are longer than Solexa reads, but shorter than the Ion Torrent reads. CLC Bio was still able to assemble any of the data types adequately, but was outperformed by Ion Torrent data on MIRA and Solexa data on Velvet Assembler.

Initial Mapping of Sequence Data

Prior to using Mauve to compare the *de novo* assemblies that the various programs would create, a mapping was performed using the respective experimental data. The reason for this was that there was concern that the individual E. *coli* K12 MG1655 strains may have SNPs, indels, or rearrangements that may have inflated the perceived errors rates of the assemblers. The mapping was performed using CLC Bio with the experimental data that was also used in the *de novo* assemblies.

Fortunately, as predicted, the number of SNPs observed from the mapping was very low, numbering about a dozen (Table 1). Considering the number of SNPs found from the assembly numbered in the hundreds, it was concluded that SNPs from the individual strains from the specific sequencing runs did not result in significant differences (Table 1).

Additionally, the variants detected in the CLC Bio mapping are most likely not all real variants. One must consider the frequencies of the SNP detected, the coverage, and the type of error that was detected. For example, both Ion Torrent and Roche 454 sequencing are prone to producing Indel errors. The developers of mappers, such as CLC

Bio, are well aware of this issue, therefore, the software has options to filter out homopolymer errors. Despite this built-in filter effect, there were still Indel errors that were detected by CLC Bio (see Table 1) and especially Ion Torrent (see Supplementary Table 4).

There seemed to be a correlation between the frequencies of detected mapping variants and the presence of that same SNV or Indel in the de novo assembly and genome comparisons performed by Mauve. In general, if a variant was detected with 100% frequency, the SNP would be present in the Mauve assembled genome as well. Any detected variant with a frequency lower than roughly 90%, seemed to be the result of a sequencing error, particularly if it was an Indel variant with Roche 454 or Ion Torrent technology.

Therefore, given that the lists of variants detected was short, and that even in these short lists, that most of the variants detected are probably not real variants, but rather products of sequencing error, it is safe to assume that mutation rates for the separate *E. coli* K12 MG1655 strains in the different labs had very similar genomes.

Reference Position	Consensus Position	Variant type	Length	Reference	Variants	Allele variants	Frequencies	Counts	Coverage
19780	18829	SNV	1	A	2	A/T	59.3/40.7	16/11	27
19796	18845	InDel	0	-	2	-/C	60.7/39.3	17/11	28
257847	254494	InDel	0	-	2	-/G	59.4/40.6	19/13	32
257869	254516	SNV	1	A	1	C	55.6	15	27
257911	254556	InDel	2	GC	2	GC/-	63.6/36.4	21/12	33
3364777	3338479	SNV	1	T	2	T/C	59.3/40.7	16/11	27
3558478	3528742	InDel	1	G	1	-	100	20	20
3957957	3921226	SNV	1	C	1	T	100	15	15

Table 1 – Variants detected from Roche 454 mapping data to reference genome

Bases Covered and Contig n50

One significant finding was that despite the variation in assembly performance, the number of missed bases remained relatively constant. The best performing assemblies were both from MIRA, missed roughly 1.6% (454) and 1.2% (Ion Torrent) of the genome, but the worst assembly (454 on Velvet), that contained the shortest contig sizes, only missed 3.3% of the genome. While this might not seem like a large difference in percentage, but when assembling genomes *de novo*, it gets exponentially more difficult to obtain the last repetitive regions that are impossible to identify without longer reads or extremely long mate pairs (Figure 1).

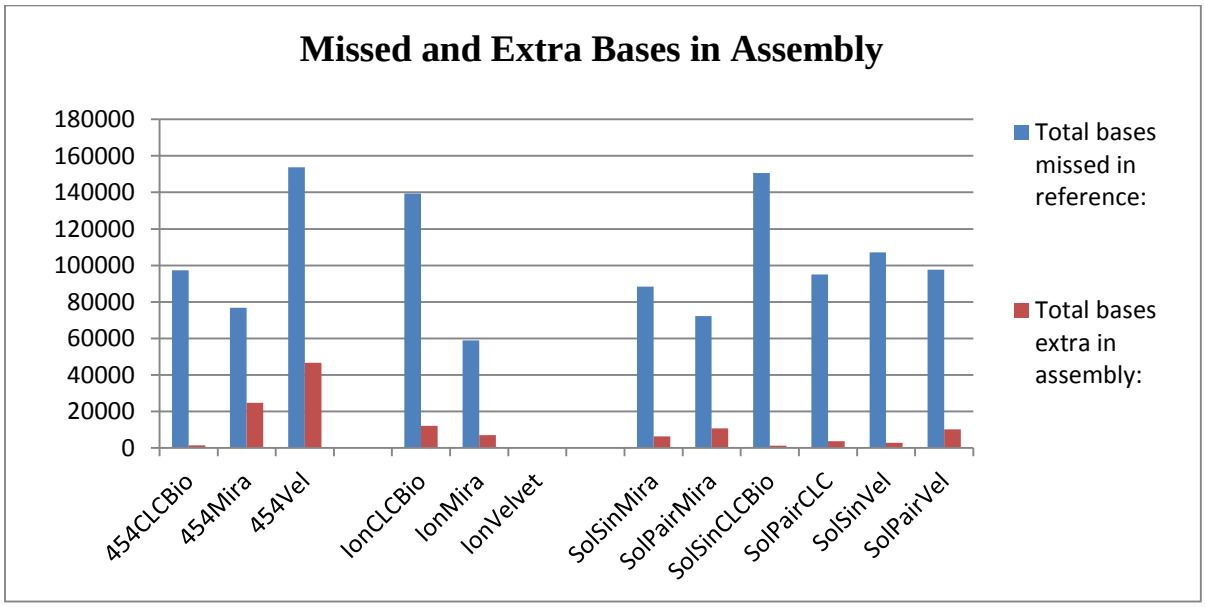


Figure 1 – Bases covered and added when comparing the various assemblies against the reference genome.

The number of extra bases incorrectly added to the assembly, varied across our computed experiments, and roughly corresponded to the aggressiveness of an assembly. For example, when comparing Solexa Single and Solexa Paired, which is the same data, with the paired data either taken into account or not, the base calls and quality scores do not change. What does change is that the assemblers use the paired status and separation

in order to improve the assembly. This improves contig size, but also results in extra bases being added to the assembly.

Contig sizes are one of the primary statistics used as a quick, though overly simplified indicator of an assembly's quality (Figure 2). It is true that assemblies are always looking for longer contigs, with the theoretically final goal of having a contig that encapsulates the entire genome.

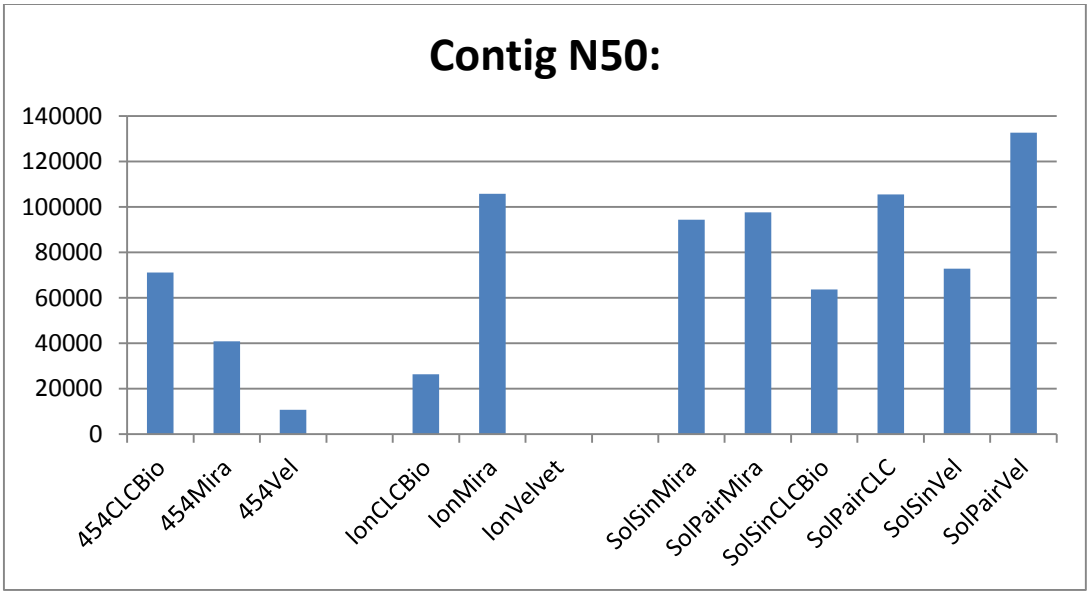


Figure 2 – Contig n50. The n50 is defined as the contig size at which half the bases in all the sequences are shorter than. Longer is indicative of a more complete assembly.

For all computational experiments performed here, Solexa data was used as a control. Realistically, there is no reason to utilize unpaired Solexa data for assembly purposes because the reads are mate-paired, and that inherently provides more information than single reads. For the purposes of our comparison, Solexa Single and Solexa Paired datasets were used separately to check that assembly statistics improved for the paired data. If no improvements are seen when utilizing paired data, then it can be assumed that the assembler was run incorrectly. As noted above, MIRA was unable to make use of either

single or paired Solexa data, though this was most likely due to the fact MIRA was not optimized to handle such reads without access to a supercomputer. CLC Genomics Workbench and Velvet Assembler did much better with the paired data, increasing contig size many times over, though Velvet was more efficient, even if both CLC Bio and Velvet have similar Solexa Single contig sizes.

We further found the contig sizes show that CLC Bio, which seemed to be the most versatile overall, best handled 454 datasets. Ion Torrent data looked excellent in MIRA, whereas Velvet was unable to use it at all, and CLC created a poor assembly.

Mistakes in Assembly

SNPs found with Mauve were not actually SNPs, but reflect missed base calls because all three sequencing platforms should have been sequencing the same genomes (Figure 3). Counter-intuitively, the number of errors increased when trying to utilize Solexa Paired data instead of just the Solexa Single, but this was perhaps due to trying to create a longer assembly. Velvet seems to have very sensitive SNP detection, calling positions of uncertain sequencing SNPs, whereas other software relies on a clear alternate "allele variant".

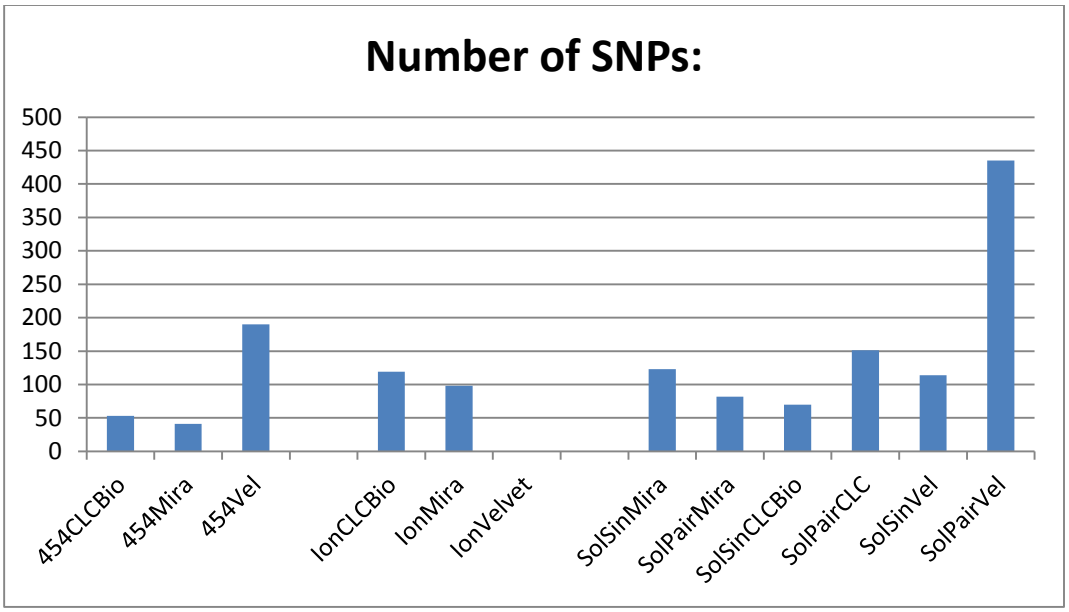


Figure 3 – Number of SNPs in each assembly, an indicator of incorrect base calls. The specific SNPs found are available in the supplementary information.

In addition to SNPs, there may be gaps in the assembly when compared to the reference (Figure 4). The Ion Torrent data seemed to create more gaps, but definitely a much greater amount on CLC Bio than on Mira.

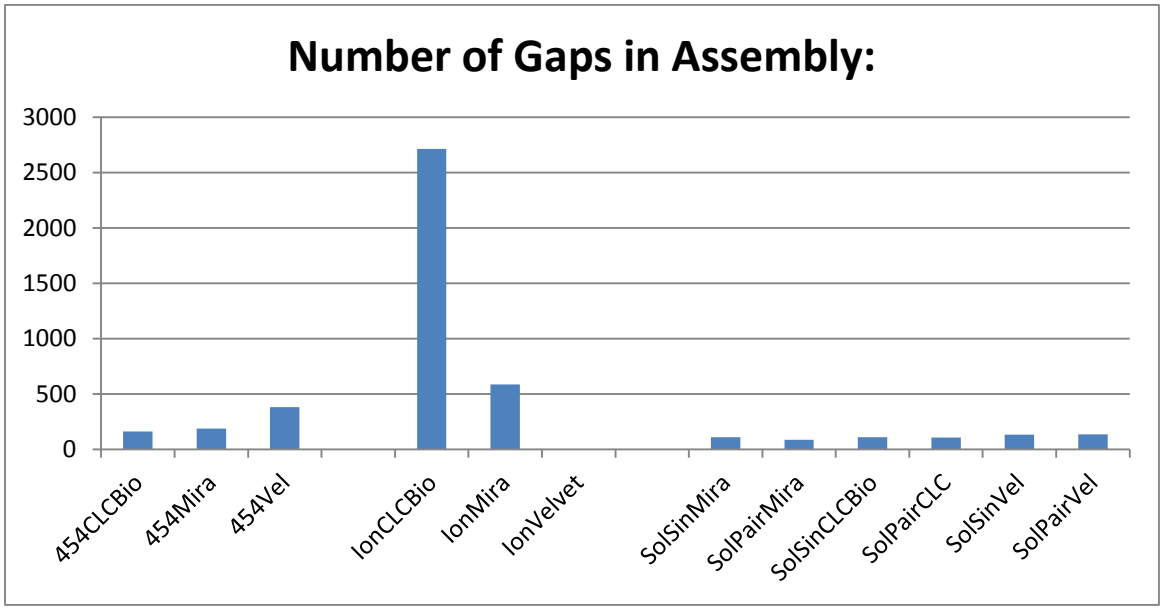


Figure 4 – Gaps in assembly.

Finally, we examined the number of Inter-LCB Boundaries (Locally Collinear Blocks) (Figure 5). An LCB is a region that does not contain any rearrangements. For the purposes of our genome comparison, which is the same genome in all cases, an inter-LCB Boundary can be thought of as a misalignment. Mauve is normally used to compare different genomes, and in those cases, regions of the genomes may be moved around due to recombination.

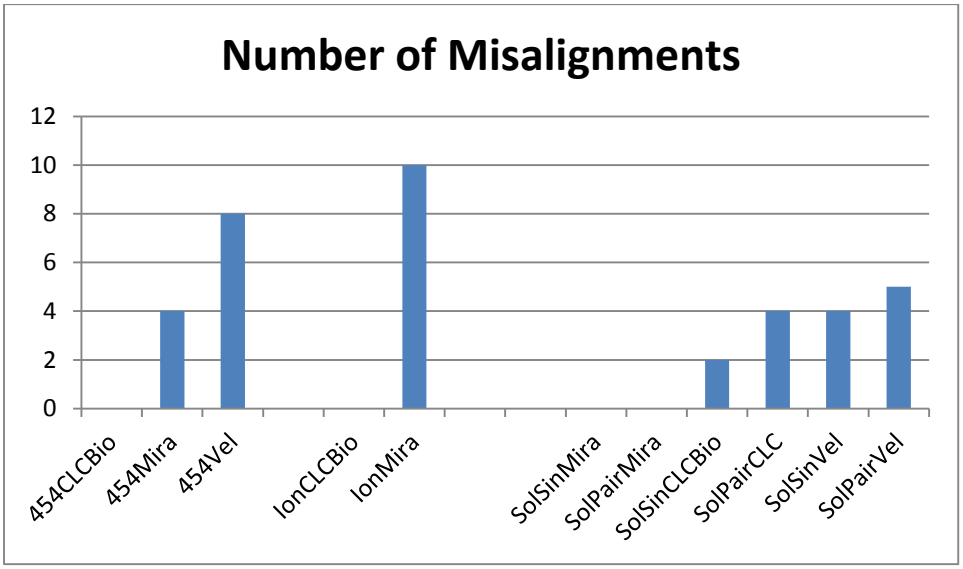


Figure 5 – Number of Inter-Linear Colocalized Blocks (LCBs), or misalignments. A LCB is a block of correctly assembled sequence, however, to align a draft genome to a reference, it may become necessary to break these blocks of correctly assembled sequence into smaller blocks because the assembler has placed it in a different location than the proper spot in the reference genome.

Discussion

Assembler Comparison

Choosing the proper assembler for the data on hand is very important, and as the CLC / Velvet comparison demonstrated, contig size cannot be the only factor that is considered. There is no denying, however, that longer contig size is a desirable metric in

the *de novo* assembly of a genome. It is interesting to observe how the different assemblers dealt with, or failed to deal with the variantions in the data. The various next-generation sequencing technologies each have certain quirks, which are either less or more compatible with the different assemblers. Homopolymer mistakes by the Ion Torrent or 454, for example, caused computational problems for Velvet Assembler, although it was able to take advantage of Solexa's data, despite the shorter reads.

When comparing *de novo* assemblers, one cannot just consider n50 contig size alone though it is a logical statistic to begin with, since it still remains a useful metric to gauge an assembly's effectiveness. The mistakes that the assemblers make, are another consideration. CLC Bio, although creating shorter contigs than Mira with Ion Torrent data, or Velvet with Solexa data made fewer misalignments, even making none with 454 and Ion Torrent data. Seemingly, CLC Bio is a more conservative assembler that produces less mistakes at the cost of shorter n50 contig lengths.

As new sequencing technologies emerge, correspondingly, new software will be developed to utilize the data. Any paradigm shifts in next-generation data will provoke a reactionary creation of software leading to entirely different computational issues that arise with the shape of the new data. Conversely, new sequencing technology will render irrelevant many of the issues that current software has taken pains to address and adapt to obsolete. This is not to say that new sequencing technologies will necessarily render every current technology moot. Sanger Sequencing is still capable of generating longer reads, though the method has largely fallen out of use due to low throughput, and still has potential applications. Benchmarking and comparison of software will continue to be a necessary

part of *de novo* or assembly mapping, because even if the shape of data and the algorithms of software changes, finding compatible pairings will remain a relevant endeavor.

SNV Locations Within Assemblies

As a general trend, it seems that the SNVs, often falsely, detected by the various assemblers clustered around certain positions, roughly finding a dozen mistakes in a range of a hundred bases in many cases for example. There did not seem to be regions in common between the three different datasets, though there was often overlap in regions that gave the individual assemblers the most problem. This implies that false SNV detection arose due to using specific genomic samples prepared by the individual laboratories.

Comparison of SNVS from 454 Mapping and Assembled Genomes

Within the 454 results, only a single SNP (Position 3957957) from the mapping was found in the Mira and CLCBio assemblies, implying that this is a real SNP. Velvet failed to detect the same SNP.

Comparison of SNVs from Ion Torrent Mapping and Assembled Genomes

There were many SNVs detected, though many were falsely detected due to the technology used. Therefore, only SNPs were examined in-depth. SNPs found in the mapping with 100% presence and consequently detected in at least the MIRA assembly were found at (57694, 2143337, and 3957957). CLCBio failed to detect any of these SNPs that were observed in the mapping and was corroborated by the MIRA assembly.

Comparison of SNVs from Solexa Mapping and Assembled Genomes

There were two sets of Solexa data, due to the fact that the data was analyzed both correctly assuming and utilizing the paired information, and incorrectly assuming and ignoring paired data in a 'Single' run. There were thirteen SNPs found in the Single mapping (257911, 547694, 547836, 1773495, 2171387, 3364777, 3421312, 3558478, 3957957, 4038792, 4169912, and 4294405). Of these, only six were found in the Paired run (547694, 1773495, 3421312, 3957957, 4038792, and 4169912). This seems intuitive that within a mapping context, the paired data would lead to more accurate SNV calling, as opposed to assemblies where using paired data gave longer contigs, but resulted in more false SNVs being called.

SNP Position	**CLCBio**	**MIRA**	**Velvet**
547694	Yes	No	Yes
1773495	Yes	No	No
3421312	No	No	No
3957957	Yes	No	Yes
4038792	No	No	No
4169912	No	No	Yes

SNP at Position 547694

Interestingly, there was a SNP found in the Ion Torrent, Single, and Paired Solexa Mappings, though not the 454 Mapping. However, using the same 454 data, all three assemblers, CLCBio, MIRA, and Velvet, detected the same SNV at that exact location. This implies that this detected SNP, across three different laboratory specimens, may be an

actual SNV of the strain. The three sequencing runs were performed from the same E. coli K12, MG1655 strain, but they were performed by different laboratories. Additionally, no other SNP is detected in common besides this particular one.

Methodology

Data Sets

All sequencing information was obtained from sequencing runs of *Escherichia coli* K12 (NC_010473), strain MG1655. The Roche 454 dataset was downloaded from CLC Bio's publically available dataset (http://www.clcbio.com/support/downloads/). The Illumina Solexa data was obtained from the company website (http://www.illumina.com/systems/miseq/scientific_data.ilmn). The Ion Torrent dataset was obtained from the company's main page (http://ioncommunity.lifetechnologies.com/docs/DOC-2265) (See Table 2).

Sequencing Platform	Coverage	Read Structure
Roche 454	21.7	Single ~ 225 bases
Solexa	37.9	Paired 35 bases Separated by 150-300
Ion Torrent	47.1	Single ~ 430 bases

Table 2 – Sequencing platforms and dataset coverage.

CLC Genomics Workbench

CLC Genomics Workbench is a commercial, licensed product, and as such, required the least amount of optimization and effort to run. All sequencing data was

imported with default options, required no additional pre- or post-processing, and were all assembled with specialized settings that were developer-recommended. The only option that had to be selected was the homopolymer filter when dealing with Roche 454 and Ion Torrent data, to prevent the homopolymer miscalls from being called Indels.

Velvet Assembler

Roche 454

Data came as two files, a fasta file and a .qual file. Pre-processing was performed with Galaxy 101 (https://main.g2.bx.psu.edu/root), to combine both files into a single fastq file.

The command lines used to run Velvet were:

```
./velveth 454Galaxy53 53 -fastq -short data/454EColiGalaxy.fastq

./velvetg 454Galaxy53/ -cov_cutoff auto -exp_cov auto
-min_contig_lgth 300
```

The first command line created the directory 454Galaxy53 with a 53 k-mer size, and designated the merged fastq file as input. 53 k-mer size was chosen based on running all possible k-mer sizes from 15 to 61, and choosing the run that yielded the largest n50.

The second command performed the actual assembly, with automatically-chosen coverage cutoffs and expected coverage. A minimum contig length of 300 was chosen to filter out contigs too short to be useful.

Ion Torrent

None of the Ion Torrent runs finished on Velvet due to memory limitations. No pre-processing was necessary.

The command lines were as follows:

```
./velveth IonTorEColi 21 21 -short -fastq data/EColi_in.iontor.fastq

./velvetg IonTorEColi21 -cov_cutoff auto -exp_cov auto
-min_contig_lgth 200
```

Solexa Single

The command lines were as follows:

```
./velveth SolexaEColiSingle21 21 -short -fastq
data/s_1_1sequence.fastq data/s_1_2sequence.fastq


./velvetg SolexaEColiSingle21 -cut_off auto -exp_cov auto
-min_contig_lgth 100
```

Solexa Paired

Pre-processing was necessary to reformat the two fastq files, which contained each half of the mate-paired reads, however Velvet Assembler assumes that the paired reads are next to their mate pair. A perl script, bundled in with the software package was able to shuffle the sequences (shuffleSequences_fasta.pl), but otherwise, quality scores and bases were unaltered.

Command lines:

```
./velveth SolexaPaired25 25 -shortPaired -fastq data/paired_1.fastq

./velvetg SolexaPaired25 -cov_cutoff auto -exp_cov 35
```

Note the –shortPaired flag in the first command line, and the use of single new fastq file generated by the pre-processing perl script.

MIRA

Roche 454

As in the case of Velvet Assembler, the 454 data was separated into two files: an .fna sequence file and an accompanying .qual quality file. Galaxy 101 was used to generate a single .fastq file for MIRA.

The command line was:

```
mira --project=454EColiGalaxy --job=denovo,genome,accurate,454 -notraceinfo
```

The job flags indicate that the task was a *de novo* assembly of a genome. The accurate flag was used instead of draft, and the 454 flag indicates the sequence data type. The –notraceinfo flag indicates that the original .sff file, recommended by MIRA developers to be included in Roche 454 assembly, was not available, and therefore would not be used to assist in assembly.

Ion Torrent

Ion Torrent data was extracted from the original .sff data files using a third-party script entitled sff_extract (http://bioinf.comav.upv.es/sff_extract/index.html).

The command line was:

```
mira --project=EColi --job=denovo,genome,accurate,iontor
```

Solexa Single

Solexa files required pre-processing to be run on MIRA due to the fact that the files were from an earlier version of the Illumina / Solexa pipeline that is no longer used. Galaxy 101 was used to convert the discontinued Solexa format into a standard Sanger .fastq format.

The command line was:

```
mira --project=EColiGalaxy2 --job=denovo,genome,accurate,solexa
```

Solexa Paired

Running Solexa Paired data was very similar to running the single version.

The command line was:

```
mira --project=EColiGalaxy2 --job=denovo,genome,accurate,solexa
SOLEXA_SETTINGS -GE:tismin=150:tismax=300
```

The above stipulates that there are paired reads, and that the minimum and maximum separation is 150 and 300 bases, respectively.

Mauve Genome Alignment Software

All assemblies created a .fasta file, containing a list of contigs. These contigs were then analyzed using the Move Contigs option of Mauve. The reference file chosen was the *E. coli* K12 genome, and the experimental chosen was the appropriate fasta file from each assembly.

Mauve, broadly speaking, aligns the draft genomes created from the various assemblers and compares it to the designated reference genome. From this, Mauve is

capable of outputting metrics such as SNPs, rearrangements, extra or missing bases, as indicated in previous figures in this chapter.

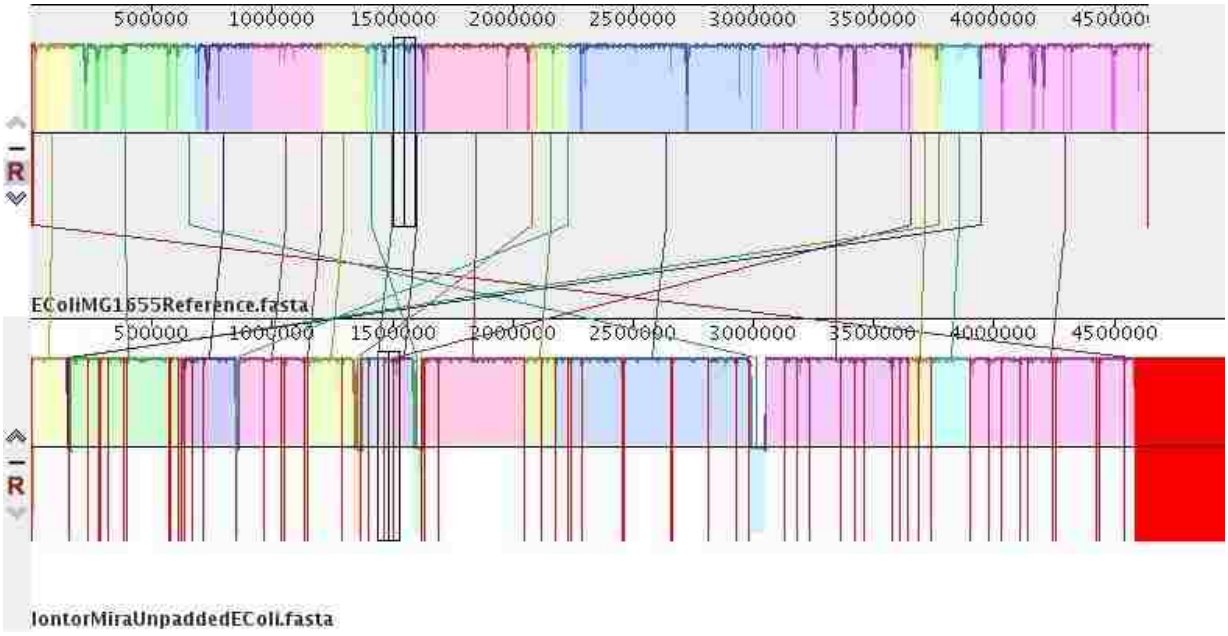


Figure 6 – Mauve comparison of E. coli K12 MG1655 reference genome and Ion Torrent assembly as performed by Mira.

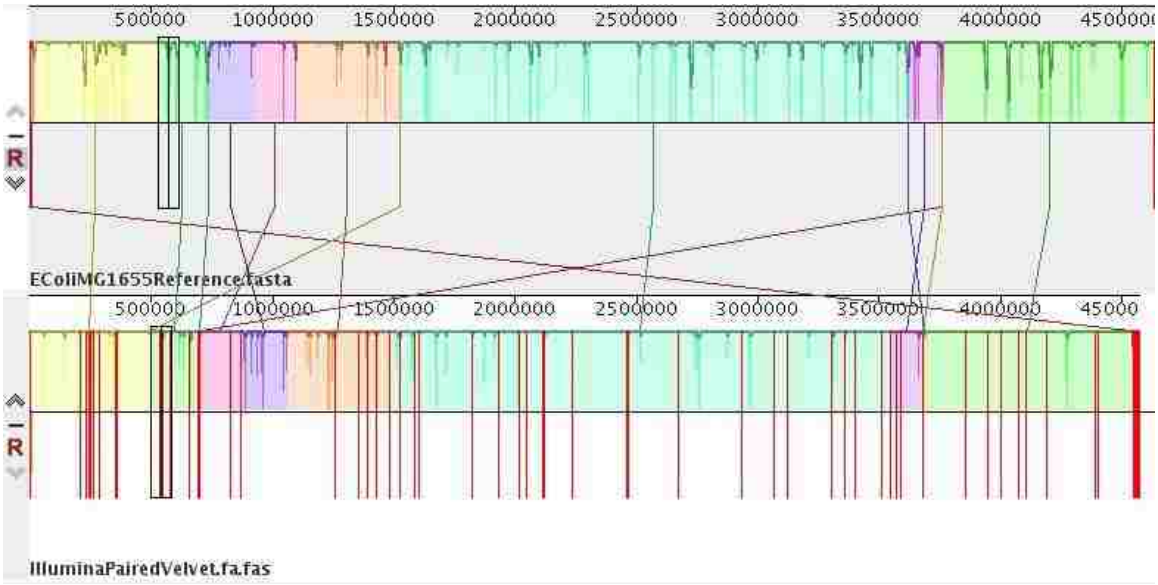


Figure 7 – Mauve comparison of E. coli K12 MG1655 reference genome and Solexa assembly as performed by Velvet.

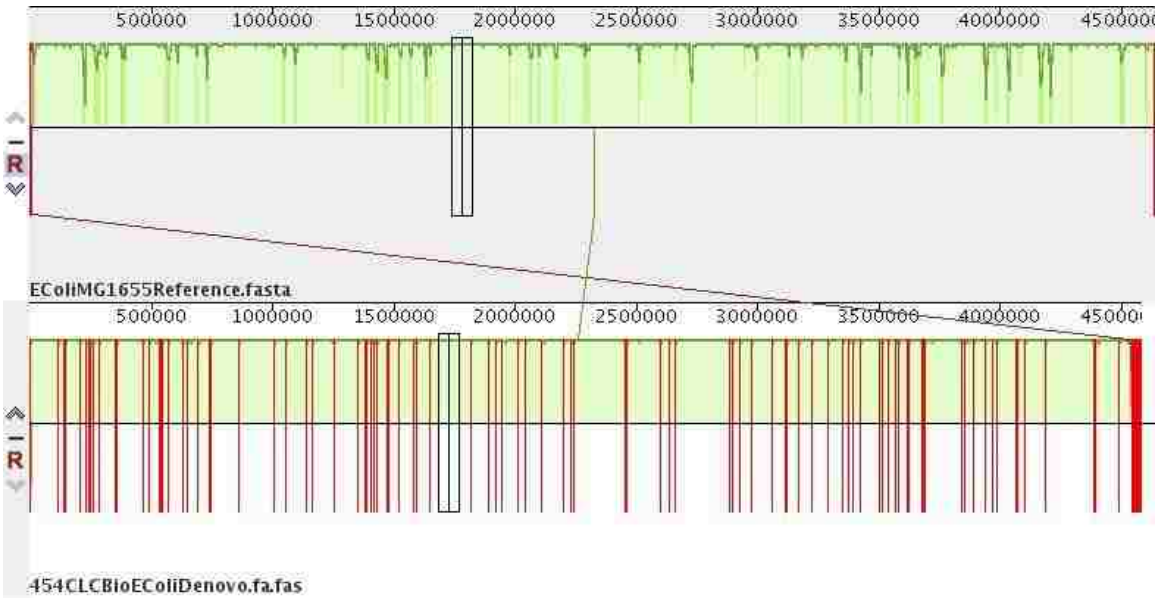


Figure 8 – Mauve comparison of E. coli K12 MG1655 reference genome and 454 assembly as performed by CLC Bio.

Supplementary Information

Included below are the various data tables generated by Mauve following reference and experimental genome comparisons.

	454CLCBio	454Mira	454Vel
Number of Contigs:	204	257	765
Number reference replicons:	1	1	1
Number of assembly bases:	4582488	4640522	4532481
Number of reference bases:	4639675	4639675	4639675
Number of LCBs:	2	10	16
Number of Blocks:	106	180	734
Breakpoint Distance:	106	180	734
DCJ Distance:	106	180	734
SCJ Distance:	212	360	1468
Number of Complete Coding Sequences:	0	0	0
Number of Broken Coding Sequences:	0	0	0
Number of SNPs:	53	41	190
Number of Gaps in Reference:	112	204	766
Number of Gaps in Assembly:	161	186	379
Total bases missed in reference:	97243	76813	153855
Percent bases missed:	2.0959 %	1.6556 %	3.3161 %
Total bases extra in assembly:	1308	24692	46654
Percent bases extra:	0.0285 %	0.5321 %	1.0293 %
Number of missing chromosomes:	0	0	0
Number of extra contigs:	99	82	37
Number of Shared Boundaries:	1	1	1
Number of Inter-LCB Boundaries:	0	4	8
Contig N50:	71127	40837	10773
Contig N90:	23313	13039	3140
Min contig length:	124	193	203
Max contig length:	222256	110514	41423

	IonCLCBio	IonMira	IonVelvet
Number of Contigs:	2439	906	NA
Number reference replicons:	1	1	
Number of assembly bases:	5608622	4963063	
Number of reference bases:	4639675	4639675	
Number of LCBs:	2	23	
Number of Blocks:	287	81	
Breakpoint Distance:	287	81	
DCJ Distance:	287	81	
SCJ Distance:	574	162	
Number of Complete Coding Sequences:	0	0	
Number of Broken Coding Sequences:	0	0	
Number of SNPs:	119	98	
Number of Gaps in Reference:	284	193	
Number of Gaps in Assembly:	2714	586	
Total bases missed in reference:	139421	58793	
Percent bases missed:	3.005 %	1.2672 %	
Total bases extra in assembly:	12101	6884	
Percent bases extra:	0.2158 %	0.1387 %	
Number of missing chromosomes:	0	0	
Number of extra contigs:	2153	837	
Number of Shared Boundaries:	2	1	
Number of Inter-LCB Boundaries:	0	10	
Contig N50:	26417	105851	
Contig N90:	508	4473	
Min contig length:	51	66	
Max contig length:	100830	357573	

	SolSinMira	SolPairMira
Number of Contigs:	168	115
Number reference replicons:	1	1
Number of assembly bases:	4605123	4597867
Number of reference bases:	4639675	4639675
Number of LCBs:	8	32
Number of Blocks:	105	85
Breakpoint Distance:	105	85
DCJ Distance:	105	84
SCJ Distance:	210	170
Number of Complete Coding Sequences:	0	0
Number of Broken Coding Sequences:	0	0
Number of SNPs:	123	82
Number of Gaps in Reference:	108	81
Number of Gaps in Assembly:	109	86
Total bases missed in reference:	88282	72138
Percent bases missed:	1.90%	1.55%
Total bases extra in assembly:	6172	10726
Percent bases extra:	0.13%	0.23%
Number of missing chromosomes:	0	0
Number of extra contigs:	67	44
Number of Shared Boundaries:	1	1
Number of Inter-LCB Boundaries:	3	17
Contig N50:	94363	97697
Contig N90:	23671	31742
Min contig length:	161	185
Max contig length:	174248	233168

	SolSinCLCBio	SolPairCLC
Number of Contigs:	243	756
Number reference replicons:	1	1
Number of assembly bases:	4591999	4747382
Number of reference bases:	4639675	4639675
Number of LCBs:	4	11
Number of Blocks:	120	95
Breakpoint Distance:	120	95
DCJ Distance:	120	94
SCJ Distance:	240	190
Number of Complete Coding Sequences:	0	0
Number of Broken Coding Sequences:	0	0
Number of SNPs:	70	151
Number of Gaps in Reference:	115	107
Number of Gaps in Assembly:	108	105
Total bases missed in reference:	150526	95024
Percent bases missed:	3.2443 %	2.0481 %
Total bases extra in assembly:	1162	3563
Percent bases extra:	0.0253 %	0.0751 %
Number of missing chromosomes:	0	0
Number of extra contigs:	124	667
Number of Shared Boundaries:	1	1
Number of Inter-LCB Boundaries:	2	4
Contig N50:	63634	105511
Contig N90:	17186	14826
Min contig length:	153	107
Max contig length:	183879	204893

	SolSinVel	SolPairVel
Number of Contigs:	231	138
Number reference replicons:	1	1
Number of assembly bases:	4571703	4578906
Number of reference bases:	4639675	4639675
Number of LCBs:	8	14
Number of Blocks:	127	73
Breakpoint Distance:	127	73
DCJ Distance:	127	73
SCJ Distance:	254	146
Number of Complete Coding Sequences:	0	0
Number of Broken Coding Sequences:	0	0
Number of SNPs:	114	435
Number of Gaps in Reference:	135	135
Number of Gaps in Assembly:	131	133
Total bases missed in reference:	107034	97615
Percent bases missed:	2.3069 %	2.1039 %
Total bases extra in assembly:	2833	10125
Percent bases extra:	0.062 %	0.2211 %
Number of missing chromosomes:	0	0
Number of extra contigs:	107	73
Number of Shared Boundaries:	1	1
Number of Inter-LCB Boundaries:	4	5
Contig N50:	72848	132727
Contig N90:	19239	33502
Min contig length:	121	121
Max contig length:	174094	390878

Single Nucleotide Variation Mapping Information

Reference Position	Consensus Position	Variant type	Length	Reference	Variants	Allele variants	Frequencies	Counts
547694	542071	SNV	1	A	1	G	100	207
1773495	1762314	SNV	1	T	2	C/T	56.8/40.5	21/15
3421312	3401828	SNV	1	A	2	C/A	51.5/45.5	17/15
3957957	3930210	SNV	1	C	1	T	100	151
4038792	4009478	SNV	1	T	2	T/G	58.1/41.9	25/18
4169912	4137318	SNV	1	T	2	T/G	62.8/37.2	27/16

Supplementary Table 1 – Illumina Paired SNVs

Reference Position	Consensus Position	Variant type	Length	Reference	Variants	Allele variants	Frequencies	Counts	Coverage
257911	253956	InDel	2	GC	2	GC/-	63.0/37.0	51/30	81
547694	537927	SNV	1	A	1	G	100	100	100
547836	538069	InDel	0	-	1	G	93.2	82	88
1773495	1754290	SNV	1	T	2	C/T	53.3/46.7	8/7	15
2171387	2148528	InDel	0	-	1	CC	86.4	51	59
3364777	3332422	SNV	1	T	2	T/C	63.7/36.3	86/49	135
3421312	3388957	SNV	1	A	1	C	50	7	14
3558478	3522267	InDel	1	G	1	-	98.9	88	89
3957957	3912982	SNV	1	C	1	T	100	70	70
4038792	3990504	SNV	1	T	2	T/G	57.7/42.3	15/11	26
4169912	4117421	SNV	1	T	2	T/G	64.7/35.3	11/6	17
4294405	4237561	InDel	0	-	1	GC	90.9	40	44

Supplementary Table 2 – Illumina Single SNVs

Reference Position	Consensus Position	Variant type	Length	Reference	Variants	Allele variants	Frequencies	Counts	Coverage
19780	18829	SNV	1	A	2	A/T	59.3/40.7	16/11	27
19796	18845	InDel	0	-	2	-/C	60.7/39.3	17/11	28
257847	254494	InDel	0	-	2	-/G	59.4/40.6	19/13	32
257869	254516	SNV	1	A	1	C	55.6	15	27
257911	254556	InDel	2	GC	2	GC/-	63.6/36.4	21/12	33
3364777	3338479	SNV	1	T	2	T/C	59.3/40.7	16/11	27
3558478	3528742	InDel	1	G	1	-	100	20	20

3957957	3921226	SNV	1	C	1	T	100	15	15

Supplementary Table 3 – Roche 454 SNVs

Reference Position	Consensus Position	Variant type	Length	Reference	Variants	Allele variants	Frequencies	Counts	Coverage
183790	183379	InDel	1	G	2	G/-	58.6/37.9	17/11	29
475972	473903	InDel	1	C	2	C/-	64.7/35.3	22/12	34
478029	475960	InDel	1	G	2	-/G	52.8/47.2	19/17	36
543644	541571	InDel	1	G	1	-	53.7	29	54
547694	545621	SNV	1	A	1	G	93.5	43	46
579799	577214	InDel	1	C	2	C/-	55.6/44.4	15/12	27
664654	661731	InDel	1	C	1	-	37.2	16	43
668151	665227	InDel	1	T	2	-/T	59.1/40.9	13/9	22
668152	665228	InDel	1	G	2	G/-	58.3/41.7	7/5	12
730122	726896	InDel	1	C	2	C/-	50.0/50.0	9/9	18
754652	751424	InDel	1	G	1	-	43.6	17	39
780499	777269	InDel	1	C	1	-	35.5	11	31
780720	777490	InDel	1	C	1	-	60	12	20
855494	852249	InDel	1	T	1	-	35.9	14	39
926825	923573	InDel	1	G	2	G/-	62.9/37.1	39/23	62
1052542	1049268	InDel	1	G	2	G/-	62.3/37.7	38/23	61
1086867	1083591	InDel	1	G	2	G/-	61.8/35.3	21/12	34
1286188	1282713	InDel	1	A	2	A/-	50.0/46.2	13/12	26
1286636	1283161	InDel	1	G	2	G/-	52.4/42.9	11/9	21
1323138	1319653	InDel	1	G	1	-	81.8	18	22
1350393	1346905	InDel	1	C	2	-/C	51.5/48.5	34/32	66
1405778	1402047	InDel	1	A	2	A/-	65.0/35.0	26/14	40
1538768	1534559	InDel	1	C	2	C/-	58.1/41.9	25/18	43
1588907	1584690	InDel	1	G	2	G/-	63.6/36.4	28/16	44
1770219	1765973	InDel	1	C	2	-/C	56.3/43.8	18/14	32
1787705	1783453	InDel	1	G	1	-	37.9	22	58
1974813	1970535	InDel	1	G	1	-	41.7	15	36
1976527	1972249	SNV	1	G	1	T	93.3	14	15
1976560	1972279	InDel	1	G	1	-	90	9	10
2143337	2138718	SNV	1	C	1	A	100	71	71
2210248	2205391	InDel	1	G	2	G/-	54.8/38.7	17/12	31
2378039	2373054	InDel	1	C	2	C/-	59.6/40.4	28/19	47
2469665	2464669	InDel	1	G	2	G/-	48.5/48.5	16/16	33
2475051	2470055	InDel	1	G	1	-	40.9	9	22
2622901	2617544	InDel	1	C	2	C/-	62.5/37.5	25/15	40

2674489	2669128	InDel	1	G	2	G/-	58.3/41.7	28/20	48
2687610	2682247	InDel	1	G	1	-	39	16	41
2752151	2745441	InDel	1	C	2	C/-	59.0/41.0	23/16	39
3151608	3144402	InDel	1	C	1	-	36.6	15	41
3263011	3255607	InDel	1	C	2	C/-	63.6/36.4	28/16	44
3373843	3366109	InDel	1	C	2	-/C	52.9/47.1	9/8	17
3390890	3383153	InDel	1	C	1	-	36.5	19	52
3439905	3430872	InDel	1	G	2	G/-	62.5/37.5	25/15	40
3445805	3436769	InDel	1	G	2	G/-	63.6/36.4	35/20	55
3476168	3467127	InDel	1	C	2	C/-	62.2/37.8	23/14	37
3498858	3489811	InDel	1	G	2	G/-	61.9/35.7	26/15	42
3597163	3588103	InDel	1	G	2	G/-	52.4/40.5	22/17	42
3645036	3634970	InDel	1	G	2	G/-	63.8/36.2	30/17	47
3857922	3846406	InDel	1	A	2	A/-	53.8/46.2	21/18	39
3957957	3945536	SNV	1	C	1	T	96.8	30	31
3994039	3981615	InDel	1	G	2	G/-	55.6/44.4	15/12	27
4016754	4004326	InDel	1	G	2	G/-	60.4/39.6	29/19	48
4073620	4060662	InDel	1	C	2	C/-	51.5/48.5	17/16	33
4083065	4070107	InDel	1	G	1	-	42.9	18	42
4093467	4080506	InDel	1	G	2	G/-	57.6/42.4	34/25	59
4526301	4507791	InDel	1	C	2	-/C	57.9/42.1	11/8	19

Supplementary Table 4 – IonTorrent SNVs. Note, many are believed to be false positives. Despite CLC Bio's built-in homopolymer filter setting for 454 or Ion Torrent reads, it is assumed most of these are false positives.

References

1. Trapnell, C. and Salzberg, S.L. (2009) How to map billions of short reads onto genomes. *Nature biotechnology*, **27**, 455-457.
2. Imelfort, M. and Edwards, D. (2009) De novo sequencing of plant genomes using second-generation technologies. *Briefings in bioinformatics*, **10**, 609-618.
3. Li, Y., Hu, Y., Bolund, L. and Wang, J. (2010) State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human genomics*, **4**, 271-277.
4. Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y., Hirst, M., Schein, J.E. *et al.* (2009) De novo transcriptome assembly with ABySS. *Bioinformatics (Oxford, England)*, **25**, 2872-2877.
5. Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O., Buffalo, V., Zerbino, D.R., Diekhans, M. *et al.* (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, **21**, 2224-2241.
6. Miller, J.R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315-327.
7. Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X. and Song, Y.Q. (2011) Evaluation of next-generation sequencing software in mapping and assembly. *Journal of human genetics*, **56**, 406-414.
8. Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. and Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome biology*, **10**, R98.
9. Hernandez, D., Francois, P., Farinelli, L., Osteras, M. and Schrenzel, J. (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome research*, **18**, 802-809.
10. Bashir, A., Volik, S., Collins, C., Bafna, V. and Raphael, B.J. (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS computational biology*, **4**, e1000051.
11. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, **40**, 722-729.
12. Costantini, M. and Bernardi, G. (2009) Mapping insertions, deletions and SNPs on Venter's chromosomes. *PloS one*, **4**, e5972.
13. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, **7**, 461-465.
14. Rothenberg, S.M. and Settleman, J. (2010) Discovering tumor suppressor genes through genome-wide copy number analysis. *Current genomics*, **11**, 297-310.
15. Diguistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M. *et al.* (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome biology*, **10**, R94.

16. Gardner, S.N., Lam, M.W., Smith, J.R., Torres, C.L. and Slezak, T.R. (2005) Draft versus finished sequence data for DNA and protein diagnostic signature development. *Nucleic acids research*, **33**, 5838-5850.
17. Glasner, J.D., Rusch, M., Liss, P., Plunkett, G., 3rd, Cabot, E.L., Darling, A., Anderson, B.D., Infield-Harm, P., Gilson, M.C. and Perna, N.T. (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. *Nucleic acids research*, **34**, D41-45.
18. Yandell, M. and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature reviews. Genetics*, **13**, 329-342.
19. Zerbino, D.R. (2010) Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **Chapter 11**, Unit 11 15.
20. Kumar, S. and Blaxter, M.L. (2010) Comparing de novo assemblers for 454 transcriptome data. *BMC genomics*, **11**, 571.
21. Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, **14**, 1394-1403.
22. Darling, A.E., Tritt, A., Eisen, J.A. and Facciotti, M.T. (2011) Mauve assembly metrics. *Bioinformatics (Oxford, England)*, **27**, 2756-2757.
23. Darling, A.E., Mau, B. and Perna, N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one*, **5**, e11147.
24. Rissman, A.I., Mau, B., Biehl, B.S., Darling, A.E., Glasner, J.D. and Perna, N.T. (2009) Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics (Oxford, England)*, **25**, 2071-2073.

Chapter 7

Conclusion

The projects undertaken have evolved significantly during the course of its completion; however, the main aims have always involved next-generation sequencing. Each completed aim held a tangible goal that was reached, and they all related to next-generation sequencing.

First, the cSBL variation involving deoxyinosine and endonuclease V was a project to develop and improve an aspect of next-generation sequencing, specifically, the acquisition of sequence data. This new biochemistry was successfully demonstrated, and provides an alternative to other proprietary methods. A lesson learned from this project was recognizing the sheer speed of the field. There are significant commercial pressures, and subsequently, a powerful and large private sector presence. The cSBL variation, while demonstrated to viable, could not compete with contemporary methods offered by private companies such as Roche 454, Illumina / Solexa, or ABI Solid in terms of pure read length. ABI Solid uses a similar methodology, employing SBL but utilizing a chemical, as opposed to enzymatic, cleavage. Nevertheless, although this cSBL variation may not become a mainstream technique in a field that advances as quickly as Moore's Law for computer processors, this particular cSBL variation was later used in a paper that came out in August of this year (Endonuclease V-assisted accurate cleavage of oligonucleotide probes controlled by deoxyinosine and deoxynucleoside phosphorothioate for sequencing-by-ligation (1)). Despite the accomplishments in developing and optimizing such an cSBL variation, while more evolutionary than transformative, still provides a useful alternative and may have niche practical uses in any application that involves cycles of DNA digestion.

In a standard sequencing pipeline, after the sequences are acquired, they must be analyzed, and our work focus shifted to this step. The amount of data being generated by next-generation sequencing technologies were growing, especially in the context of discovering meaningful insight into the human genome. The development of SAWTooth (Sequence Analysis Workbench Tool) was an attempt to address this problem. The "shape" of the data varied, depending on the platform used to obtain the sequencing information. For example, the data could be mate-paired or singled, and there were tendencies toward different error types in the different codes used for analysis. Software algorithms vary, making different codes more or less useful for each type of sequencing project. At the time, there was a potential need in the laboratory to accumulate many Polonator reads, which were mate-paired with a sizable separation and contained short reads. The SAWTooth code was developed to specifically take reads in this configuration, and match them up to a reference genome as efficiently and accurately as possible. The results were a success, as SAWTooth outperformed its more popular contemporaries at the time, NovaAlign and Bowtie 2. Further work on the SAWTooth project to needed, because while the demonstrated ability to map to a reference more quickly was impressive, the code cannot yet generate SNP or Indel information for downstream biological analysis. Currently, SAWTooth can provide only re-sequencing metrics, or reveal copy number variations. As mentioned earlier regarding the cSBL biochemistry, the field moves extremely fast and the shape of the data being generated for use by the code changed within the course of a year. Nevertheless, the basic algorithms that form the basis of SAWTooth proved highly efficient, and clearly powerful, when utilized for the specific purpose of

mapping short, mate-paired reads. They have the potential to be adapted to analyze reads from other technologies should this prove useful.

The last project in this work was one of meta-analysis; to compare the tools that can be used to generate a *de novo* assembly. The concept of a *de novo* assembler comparison is not new. Next-generation sequencing is a large and diverse field, and while this is often a strength as a whole, it leads to numerous and diverging options when it comes to acquisition, analysis, and assembly. Every step has multiple valid options, depending on the scientific application, therefore, testing and assessment of which software package should be paired with which sequencing platform for a given application becomes extremely important. Overall, this work may serve as a useful guide for other individuals in the field to consider when choosing assemblers for their bacterial *de novo* sequencing projects. Unfortunately, the relevancy of such findings may be limited, due to the quick speed of the field.

With that in mind, perhaps it would be prudent for next-generation sequencing work within an academic lab to consider the large private sector presence and rapid evolution of the technology. This is reflected in the shift in this dissertation's projects from optimizing biochemisties and developing software, to using available data from tested methods and analyzing the manner in which draft genomes are created.

The work required to create a brand new method or a new biochemistry for sequencing relies upon a series of intelligent guesses, and a process of elimination that does not guarantee success. Completion of such a task is a manner of diligence, laboratory competence, and time. Time can be shortened by utilizing larger and larger staff,

something which academic laboratories must consider in evaluating competitiveness with the private sector. The same principles, to a lesser extent, apply to the development of software for the field, though the private sector presence in such endeavors is smaller.

In the course of this work, the quickest path was to take already tried and tested methodologies and generate data. The data is of great interest and can be specialized for academic labs, looking to investigate specific organisms. This reduces competition that may be present from trying to wholesale improve aspects of the entire field, and creates data where publication can be more certain.

This is not to slight any of the previous accomplishments described in this dissertation, which were hard-fought and well-earned. It is, of course, easier to utilize useful tools, rather than attempting to improve the tools themselves. It is for this reason that some focus in the Edwards laboratory changed from improving the tools to generating or just obtaining data and using Genome-Wide Association studies to analyze interesting biological questions.

However, the development of this particular cSBL biochemistry or the SAWTooth software was still worth the effort despite the seemingly modest impact of these advances in the field. From a personal and localized level, the efforts of creating these methods are learning experiences that instructed those involved and gave valuable understanding. Additionally, these advances still contribute to the growing body of knowledge that exists in these fields. These tools exist for future scientists to use, which are a simple Pubmed search away. Seemingly minor biochemistries or apparently niche and overly specialized code may be useful with the right adjustments, tweaks, or the right application.

Philosophically speaking, trying to use the best tool available may be the easiest option; perhaps what is right, though not necessarily most expedient, is to create an optimized tool.

Genome sequencing holds great potential for altering lives in real and tangible ways, particularly in healthcare. This has fostered a very fertile ground for private sector presence, causing the technologies to improve and change. It is only a matter of time before a new technology or method, whether nanopore technologies or Transmission Electron Microscopy methods, changes the field. This will render current "next-generation" sequencing technologies obsolete, and industrial goals will shift from improving read lengths to increasing electric field sensitivity. The entire game will change, and it is only a matter of time, but the possibilities are as exciting as they are difficult to predict.

References

1. Li, Y., Pan, Z., Tang, J., Pu, D., Xiao, P. and Lu, Z. (2012) Endonuclease V-assisted accurate cleavage of oligonucleotide probes controlled by deoxyinosine and deoxynucleoside phosphorothioate for sequencing-by-ligation. *The Analyst*, **137**, 4421-4424.