7-1-2013

Next-Generation Sequencing for Biomedical Applications

Mingyan Xu

Mingyan Xu

Candidate

Biomedical Sciences---Molecular Genetics & 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

Scott Ness

Todd Thompson

Rebecca Hartley

NEXT-GENERATION SEQUENCING FOR BIOMEDICAL APPLICATIONS

by

MINGYAN XU

B.S., NANCHANG UNIVERSITY, JIANGXI, CHINA 2005
M.S., BEIJING NORMAL UNIVERSITY, BEIJING, CHINA 2008

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biomedical Sciences

The University of New Mexico
Albuquerque, New Mexico

July, 2013

ACKNOWLEDGMENTS

I heartily acknowledge Dr. Jeremy S. Edwards, my advisor and dissertation chair, for continuing to encourage me through the years of classroom teachings and the long number of months writing and rewriting these chapters. His guidance and professional style will remain with me as I continue my career. Dr. Edwards is very sharp. He always has solutions whenever we had problems during the experiments, especially for the research design. I will remember the hours and hours we spent in Jersey Jacks and Med II discussing the experiments, the ideas for grant, and eventually we awarded the funding.

I also thank my committee members, Dr. Rebecca Hartley, Dr. Scott Ness, and Dr. Todd Thompson, for their valuable recommendations pertaining to this study and assistance in my professional development. Dr. Hartley was very nice and knowledgeable; she spent hours and hours after class to help me understand the material because I could not catch it up in class. Dr. Ness is very gentle and humor. More importantly, he helped me to find my second rotation lab, which is the same lab I spent my years for Ph.D. I will never forget that he taught me to choose my own Ph.D life, not for others, not for thanksgiving. Dr. Thompson is very sharp. He can always point out my weakness, which helped a lot for research design. Especially during the first year when I had no background on Biology, Dr. Todd Thompson met me every week and helped to build up biological sense from essential dogmas. Without the suggestions, advices and helps from my committee, this work could never been done.

To colleagues Norah Martinez, Dr. Anthony Aragon, Dr. Antoine Ho, Dr. Hazel Higgins, Florence Sanchez, I appreciate all the help from them. Thanks Norah, the “lab machine” preparing lots libraries for everyone. Thanks Anthony for help with the DNA analysis and running Ion PGM. Thanks Antoine taught me to do ePCR and helped me especially from my rotation in the lab. Thanks Hazel preparing E-coli DNA for my experiments. Thanks Florence placing all the orders for me. Everyone contributed a lot to my Ph.D work, without him or her it cannot be done.

To BSGP, they are the greatest faculty. To Ignacio Ortiz, he knows everything that you asked about the program, all the regulations, rules, deadlines, and milestones. He has done great job as BSGP manager. We are very sorry for his leaving, but congratulates to his new job at College of Nursing. To Helen Hathaway, our great and nice program director. She is very helpful whenever you have problem, any problem about the study, about the rotation, about the class. To Natalie Robinson Bruner, she can always organize all your paperwork and arrange a meeting room for us. To Mary Fenton, she does great job since we submitted the application for BSGP program. She is very nice. To all the BSGP faculties, their great job is appreciated.

To KUGR facility, Dr. Scott Ness, Jamie Padilla, Dr. Gavin Pickett provided great services, such as Nanodrop, Agilent Bioanalyser, and especially Ion Proton. Thanks Jamie for running my samples on Proton.

And finally to my most important ones, to my wife, Yuan wen, your love is the greatest gift of all. To my sweaty daughter, Sophia Wenqing Xu, you are the best gift of all. To my younger brothers Mingrun and Mingfang, and younger sister Niannian, you are always on my side and give me support. To my parents, Yuanzhong Xu and Dong’e Wu, everything is impossible without you and your support as my backbone. I will never forget how hard you have been working on educating me, supporting me, and helping me, from the days I entered schools to the days I graduated. No matter how badly people discourage us, no matter how badly people discriminate poor and weak, we are all keeping together and pursuing what we believe.

NEXT-GENERATION SEQUENCING FOR BIOMEDICAL APPLICATIONS

By

MINGYAN XU

B.S., NANCHANG UNIVERSITY, JIANGXI, CHINA 2005
M.S., BEIJING NORMAL UNIVERSITY, BEIJING, CHINA 2008

ABSTRACT

Based on the development of current status of Next-Generation Sequencing technologies (NGS), I developed several technologies to improve NGS technologies and exampled to apply the NGS technology to biomedical sciences. In the first chapter, I overviewed most of the NGS technologies, described the characteristics of the technologies, and compared the pits and falls, as well as the applications. Based on the current development status, I developed Dual-Primer emulsion PCR (DPePCR) to simplify the library preparation and elongate the read length, and the details were illustrated in the chapter 2. To minimize the sequencing feature/template and improve sequencing throughput, I developed Rolony sequencing by Synthesis, which was elaborated in chapter 3. To utilize the NGS technologies on biological researches, Haplotype-resolved sequencing was developed to sequence haplotype genome in chapter 4. Haplotype-resolved sequencing technology provided haplotype information, which is very important for fully understanding human genome, completely describing genetic diversities, and identifying the association between genetic variations and diseases. In one word, this dissertation developed several sequencing technologies that improved the next-generation sequencing technologies, and will potentially contribute to biomedical applications.

TABLE OF CONTENTS

List of Figures

Abbreviations

AP = Anchor primer

ASE = Allele specific expression

ASM = Allele specific methylation

bp = base pair

BS^3 = Bis(sulfosuccinimidyl) suberate

CCD = Charge-coupled device

CNVs = Copy Number variations

cPAL = combinatorial probe anchor ligation

DNA = Deoxyribonucleic acid

dNTPs = deoxyribonucleotides

DPePCR = Dual Primer emulsion Polymerase Chain Reaction

dsDNA = double strand DNA

ePCR = emulsion Polymerase Chain Reaction

emPCR = emulsion Polymerase Chain Reaction

gDNA = Genomic DNA

HapMap = International HapMap project

Het = heterozygous

HGP = Human Genome Project

Hom = homozygous

Ion PGM = Ion Personal Genome Machine

MB Water = Molecular Biology Water

MDA = Multiple Displacement Amplification

MPC = Magnetic Particle Collector

NGS = Next Generation Sequencing

NPM = Nextera PCR Mix

PCR = Polymerase Chain Reaction

PNK = polynucleotide kinase

Polony = PCR Amplified Colony

PPC = PCR primer cocktail

RCA = Rolling Circle Amplification

RCR = Rolling Circle Replication

Rolony = Rolling Circle Amplified Colony

RSB = Resuspension buffer

SBL = Sequencing by Ligation

SBS = Sequencing by Synthesis

SMRT = Single molecule real time

SNPs = Single nucleotide polymorphisms

SNVs = Single nucleotide variations

SOLiD = Sequencing by Oligonucleotide Ligation and Detection

ssDNA = single strand DNA

TD = Tagmentation

TDE1 = Tagmentation Enzyme

UV = Ultraviolet

WGA = Whole Genome Amplification

WGRS = Whole Genome Re-Sequencing

CHAPTER 1

INTRODUCTION

Next-Generation DNA Sequencing Technologies

1.1 Introduction

Since the 20th century, human vision has extended to two opposite directions. One direction is the macrocosm, where scientists study the secret of the planets and Universe. On the other hand, the microcosm has become a more popular objective where scientists are studying cells, DNA, and other macromolecules. In 1953, DNA was proved as a double helix structure which contains all the genetic information of most organisms[1], which marked the first milestone of modern biological science. However, this milestone just opened the door to the amazing biological world. In the past several decades, the study of DNA has driven most aspects of human genome science research[2-4].

Since DNA codes for all the genetic information, scientists have been curious to figure out all the DNA sequence for most organisms, especially humans. In 1977, Frederick Sanger, a two-time Nobel laureate, published the first paper describing DNA sequencing with chain-terminating inhibitors and initiated the automated DNA sequencing era[5]. After that, capillary sequencing was developed, which was the technology used for the Human Genome Project (HGP)[2,6-8]. The conventional Sanger sequencing technology is, therefore, considered the "first-generation" sequencing and has played a significant role for over 30 years. The development of DNA sequencing technology has accelerated in the past decade. Since 2005, a number of innovative "next-generation" (or 2nd generation) ultra-high throughput sequencing technologies[9-14] (i.e. Roche/454[15], Polonator[16], Applied Biosystems SOLiD, Complete Genomics[17], Illumina/Solexa[18], Qiagen/Intelligent Bio-Systems[19], Helicos Biosciences[20,21], Pacific Biosciences[22], Ion Torrent Systems[23] etc.) have become commercially available for biomedical research. Over the course of the past 10 years, the HGP motivated an unbelievable reduction of sequencing costs, from $10 per finished base to less than $0.00001 per finished base. Additionally, there are a series of sequencing technologies under development, (e.g. Halcyon Molecular[24], ZS Genetics[24], Oxford Nanopore[25], IBM[26], GE (http://www.patents.com/us-7264934.html), Noblegen[27], Bionanomatrix[28], GnuBIO etc[24]) which will not be covered in this chapter. High-throughput technologies for DNA sequencing have succeeded by spatially and temporally increasing the amount of data that can be generated and have shown tremendous impact on biomedical research and medicine[11,29-40].

Taking an overview of the development of DNA sequencing technologies, there is a principle that is a common point to study the microcosm, which is miniaturizing of the processes, expediting sample processing, magnifying the smaller objectives making them visible and detectable. Based on the evolution of DNA features during the development of DNA sequencing technology, I classify the next-generation sequencing technologies as three groups. The first group contains Roche/454, Polonator, SOLiD, and Ion Torrent systems because they all employed emulsion polymerase chain reaction (emPCR or ePCR) and the sequencing templates are bead-based micron scale. The second group composes of Illumina/Solexa, Complete Genomics technologies because their sequencing features are nano-scale size and non-bead-based, and smaller than the first group. The third group involves Helicos and Pacific Biosciences due to their sequencing features, which are single molecule and PCR-free processes. The third group has also been described as the “third-generation sequencing” technologies[24]. Oxford Nanopore sequencing is not discussed here since it is not commercially available. In this chapter, I will focus on the well-developed technologies and discuss these three groups respectively, compare the pros and cons, summarize the main applications, and give a methodological idea to researchers and other potential readers.

1.2 Micron scale bead-based sequencing technologies

In 2005, the 454 sequencing paper was published on *Nature* on July 31st[15] and Polony sequencing paper was published on *Science* on August 4th[16] marked the beginning of the next-generation sequencing era. The improvements of these sequencing technologies were concentrated in miniaturization of the processes, multi-parallel, and rapid sample processing comparing to the first-generation sequencing technologies. During the development, minimizing the DNA sequencing feature, simplifying the whole procedure and increasing the throughput were the three main goals that the efforts were focused upon. This section aims to focus on the micron scale DNA sequencing feature technologies, which include Roche/454 pyrosequencing, Polony sequencing, SOLiD sequencing and Ion Torrent semiconductor sequencing. The common points for these technologies are micron scale bead-based and ePCR amplified templates. However, they are distinguishable from sequencing chemistry to signal detection instruments. 454 utilizes pyrosequencing, Polony employs sequencing by ligation (SBL), SOLiD takes advantages of two-base encoding SBL and Ion Torrent system exploits semiconductor chip to detect hydrogen ions that released during the polymerization of DNA.

1.2.1 Roche/454 sequencing technology

Roche/454 FLX was the first commercial next-generation sequencing system introduced in 2005[15]. 454 Sequencing uses a large-scale parallel pyrosequencing system capable of sequencing roughly 400-600 megabases of DNA per 10-h run on the Genome Sequencer FLX with GS FLX Titanium series reagents. This system relies on fixing nebulizer and adapter-ligated DNA fragment library (figure 1) to small DNA-capture beads in a water-in-oil emulsion. The DNA fixed to these beads is then amplified by PCR. Each DNA-bound bead is placed into a ~44 μm well on a PicoTiterPlate, a fiber optic chip. A mix of enzymes such as DNA polymerase, ATP sulfurylase, and luciferase are packed into the well. The PicoTiterPlate is then placed into the GS FLX System for sequencing.

1.2.1.1 DNA library preparation and emPCR

To generate fragment library for emPCR (figure 2), genomic DNA is sheared into small fragments (300-800 base pairs), blunt-end repaired and A-tailed (Adenylylation). Short adapters are then ligated onto the ends of the fragments. These adapters provide priming sequences for both amplification and sequencing of the sample-library fragments. One adapter contains a 5'-biotin tag for immobilization of the DNA library onto streptavidin-coated beads. After nick repair, the non-biotinylated strand is released and used as a single-stranded template DNA (sstDNA) library. The sstDNA library is assessed for its quality and the optimal amount (one DNA molecule per bead) needed for emPCR is determined by titration. The sstDNA library is immobilized onto beads. The beads containing a library fragment carry a single sstDNA molecule. The bead-bound library is emulsified with the amplification reagents in a water-in-oil mixture. Each bead is captured within its own micro-reactor where PCR amplification occurs. This results in bead-immobilized, clonally amplified DNA fragments.

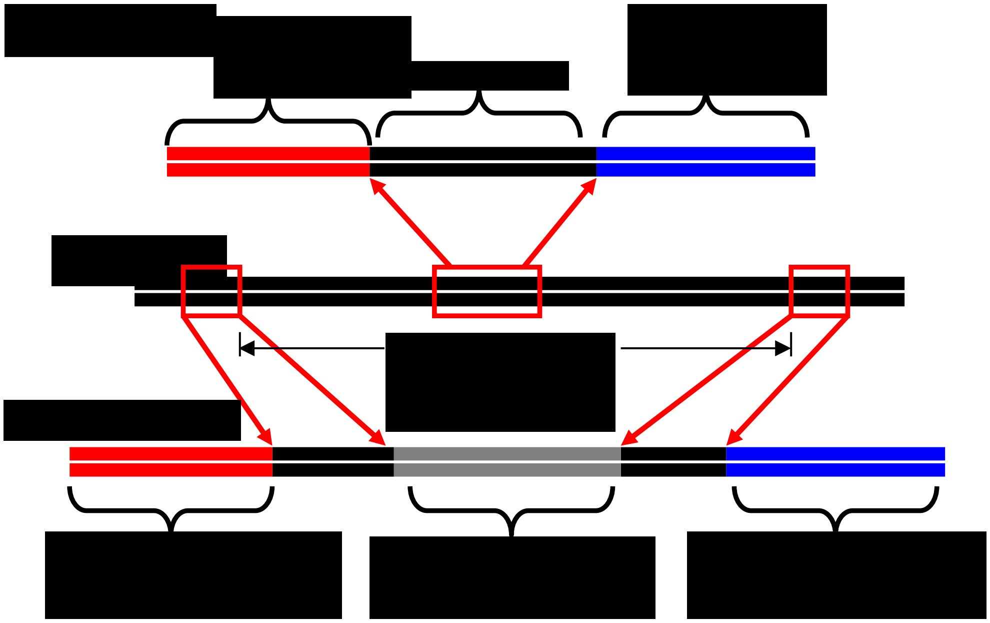


Figure 1. Structure of fragment library and mate-paired library. Fragment library and mate-paired library are two most used library structures for next-generation sequencing. For fragment library, genomic DNA is randomly sheared to small fragments, 100-1000bp. After end-repair and A-tail, two adapters can be ligated onto the ends of fragments. Each fragment is distributed randomly

from the genomic DNA. For mate-paired library, genomic DNA is randomly sheared to 1kb-3kb fragments. The fragments are then circularized with a universe linker. The restriction sites on the universe linker allow restriction enzyme digestion to cut out two genomic tags on each side of the universe linker. The two genomic tags are then end-repaired, A-tailed, and ligated to adapter ends to generate the mate-paired library. The two genomic tags are separated by a certain distance (i.e.1kb) in the genome, but not randomly distributed as with the fragment library.

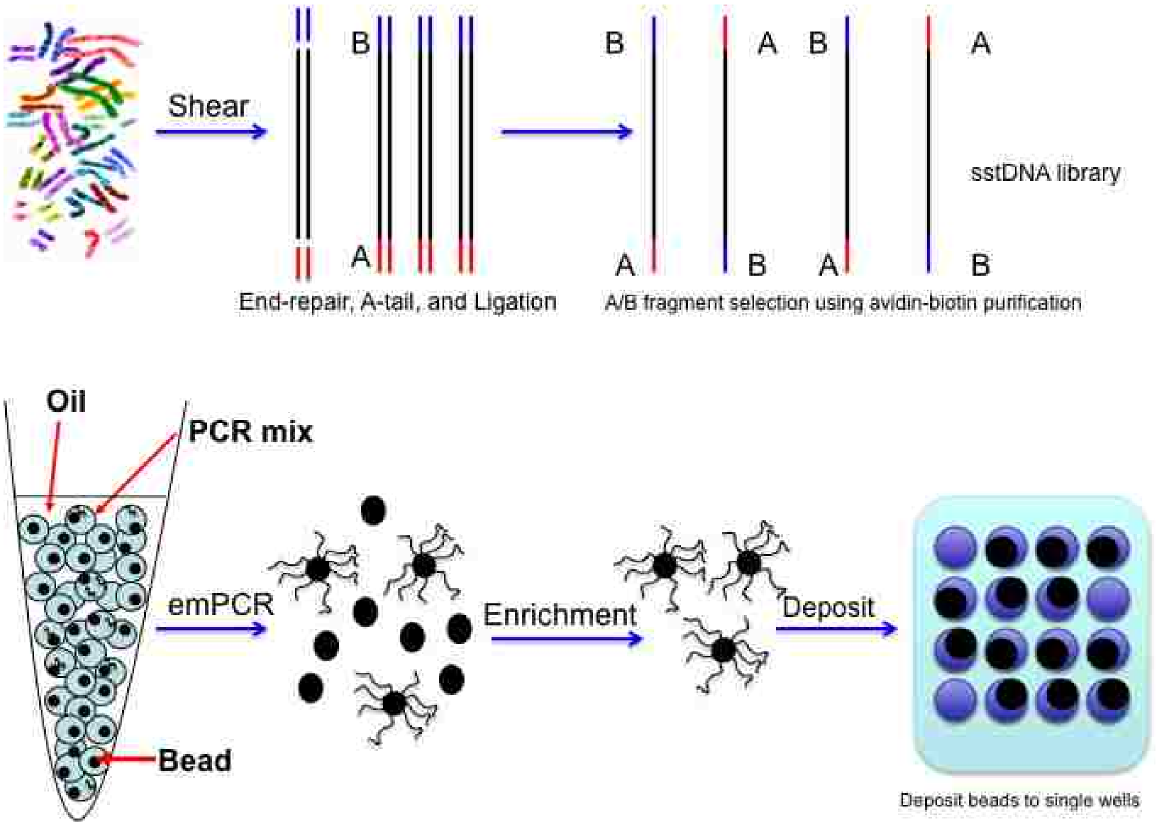

Shear
B
A
End-repair, A-tail, and Ligation
B
A
A
B
B
A
A
B
sstDNA library
A/B fragment selection using avidin-biotin purification
Oil
PCR mix
emPCR
Bead
Enrichment
Deposit
Deposit beads to single wells

Figure 2. Overview of the Roche/454 DNA sequencing technology. Genomic DNA is sheared, and then ligated to adapters A and B on two ends. Single-stranded template DNA (sstDNA) library is selected using avidin-biotin purification strategy. The sstDNA library is mixed with PCR

reagents, beads, and then emulsified with oil to form DNA molecule and bead (ideally, 1:1 ratio approximately) in water-in-oil microreactors. DNA positive beads can be enriched after PCR and emulsion breaking steps. Finally, the beads are coated with approximately 1 million copies of the original single-stranded fragment, which provides sufficient signal strength during the pyrosequencing sequencing and signal detection processes.

1.2.1.2 Sequencing

Single-stranded template DNA library beads are added to the DNA Bead Incubation Mix (containing DNA polymerase) and are layered with Enzyme Beads (containing sulfurylase and luciferase) onto a PicoTiterPlate device (figure 2). The device is centrifuged to deposit the beads into the wells. The layer of Enzyme Beads ensures that the DNA beads remain positioned in the wells during the sequencing reaction. The bead-deposition process is designed to maximize the number of wells that contain a single amplified library bead.

The loaded PicoTiterPlate device41 is placed into the Genome Sequencer FLX Instrument. The fluidics sub-system delivers sequencing reagents (containing buffers and nucleotides) across the wells of the plate. The four DNA nucleotides are added sequentially in a fixed order across the PicoTiterPlate device during a sequencing run. During the nucleotide flow, millions of copies of DNA bound to each of the beads are sequenced in parallel. When a nucleotide complementary to the template strand is added into a well, the polymerase extends the existing DNA strand by adding nucleotide(s). Addition of one (or more) nucleotide(s) generates a light signal that is recorded by the CCD camera in the instrument. This technique called pyrosequencing[41] is based on sequencing-by-synthesis. The signal strength is proportional to the number of nucleotides (figure 3); for example, homopolymer stretches, incorporated in a single nucleotide flow generate a greater signal than single nucleotides. However, the signal strength for homopolymer stretches is linear only up to eight consecutive nucleotides after which the signal falls-off rapidly.

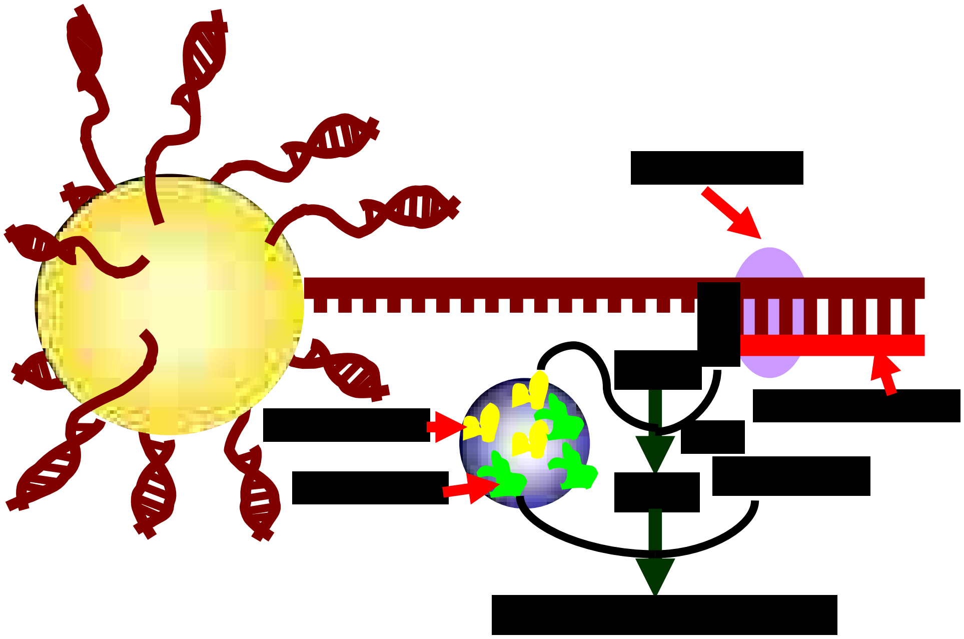


Figure 3. Principle of Pyrosequencing. Firstly, a sequencing primer is hybridized to a single-stranded amplified template, and incubated with DNA polymerase, ATP sulfurylase, luciferase,

adenosine 5' phosphosulfate (APS), and luciferin. Then, the first deoxribonucleotide triphosphate (dNTP) is added to the reaction. DNA polymerase catalyzes the incorporation of the complementary deoxyribo-nucleotide triphosphate into the DNA strand, and releasing pyrophosphate (PPi) in a quantity equimolar to the amount of incorporated nucleotide. PPi can be converted to ATP by sulfurylase in the presence of adenosine 5' phosphosulfate (APS). Then, ATP drives the luciferase-mediated conversion of luciferin to oxyluciferin that generates proportional amount visible light to the amount of ATP. Finally, the light is detected by a charge coupled device (CCD) chip and seen as a peak in the raw data output. The height of each peak (light signal) is proportional to the number of nucleotides incorporated. After degradation of unincorporated nucleotides and ATP, another nucleotide is added and cycles can be repeated to sequence the template.

1.2.2 Polony sequencing

Polony Sequencing is a highly accurate multiplex sequencing technique that can generate millions of immobilized DNA sequences in parallel[16]. This technique was first developed by Dr. George Church at Harvard Medical School[42]. Unlike other sequencing techniques, Polony sequencing technology is an open platform with freely downloadable, open source software and protocols[43]. Also, the hardware of this technique can be easily set up with a commonly available epifluorescence microscope and a computer-controlled flowcell/fluidics system.

Polony sequencing is generally performed on paired-end Tags (PET) library (or mate-paired library, figure 1) that each molecule of DNA template is of 135bp in length with two 17-18bp paired genomic tags separated and flanked by common sequences. The current read length of this technique is 26 bases per bead and 13 bases per tag, leaving a 4-5 bases gap in each tag. The procedure of Polony sequencing can be broken down into three main parts, which are the paired end-tag library construction, template amplification and DNA sequencing.

1.2.2.1 Paired-end-tag library construction

This protocol begins by randomly shearing the tested genomic DNA into ~1kbp size fragments (figure 4). The sheared DNA molecules are then end repaired and A-tailed. During the next step, the DNA molecules are circularized with T-tailed 30bp long synthetic oligonucleotides (T30), which contain two outward-facing MmeI (type IIs restriction endonuclease) recognition sites and the resulting circularized DNA undergoes rolling circle replication. The amplified circularized DNA molecules are then digested with MmeI which cuts at a distance from its recognition site, releasing the T30 fragment flanked by 17-18bp tags (~70bp in length). The paired-tag molecules need to be end-repaired prior to the ligation of ePCR (emPCR) primer oligonucleotides (FDV and RDV) to their both ends. The resulting 135bp library molecules are size-selected and nick translated. Lastly, the 135bp paired end-tag library molecules are amplified with PCR to increase the amount of library material and eliminate extraneous ligation products in a single step. The resulting DNA template contains a 44bp FDV sequence, a 17-18bp proximal tag, the T30 sequence, a 17-18bp distal tag, and a 25bp RDV sequence.

1.2.2.2 Template amplification by ePCR

To perform ePCR (figure 4), firstly, the Monosized, paramagnetic streptavidin–coated beads are pre-loaded with dual biotin forward primer. Next, an aqueous phase is prepared with the pre-loaded beads, PCR mixture, forward and reverse primers, and the paired end-tag library. This mix is then vortexed with an oil phase to create the emulsion. Ideally, each droplet of water in the oil emulsion has one bead, one molecule of template DNA, and plenty PCR mix, permitting millions of non-interacting amplification within a milliliter-scale volume by performing PCR. After amplification, the emulsion from preceding step is broken using silicone oil, isopropanol, annealing and detergent buffer, following with a series of vortexing, centrifuging, and magnetic separation. The resulting solution is a suspension of empty, clonal and non-clonal beads, which arise from emulsion droplets that initially have zero, one or multiple DNA template molecules, respectively. The amplified bead can be enriched and 3'-capped before loading on the flowcell.

The enrichment of amplified beads is achieved through hybridization to a larger, low density, non-magnetic polystyrene beads that pre-loaded with a biotinylated capture oligonucleotides (figure 4). The mixture is then centrifuged to separate the amplified and capture beads complex from the unamplified beads. The amplified capture beads complex has a lower density and thus will remain in the supernatant while the unamplified beads form a pellet. The supernatant is recovered and treated with NaOH, which will break the complex. The paramagnetic amplified beads are separated from the non-magnetic capture beads by magnetic separation. This enrichment protocol is capable in enriching five times the amount of amplified beads.

The purpose of bead capping is to attach a "capping" oligonucleotide to the 3' end of both unextended forward ePCR primers and the RDV segment of template DNA. The cap is an amino group that prevents fluorescent probes from ligating to these ends and at the same time, helping the subsequent coupling of template DNA to the aminosilanated flow cell glass surface.

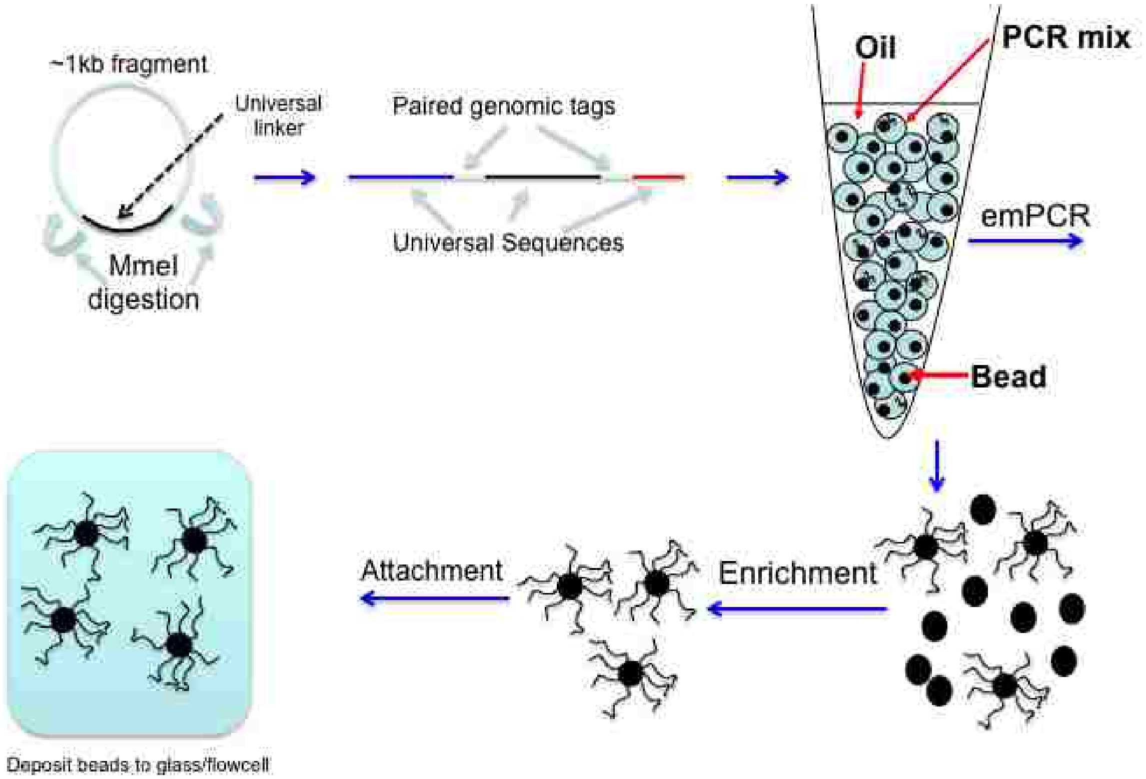


Figure 4. A multiplex approach to DNA sequencing. Genomic DNA is Sheared, size-selected (yellow), and then circularized with a linker (red) bearing MmeI recognition sites. After MmeI digestion and adapters (blue) ligation, the 134- to 136-bp mate-paired library can be achieved by rolling-circle amplification. EPCR generates clonal template amplification on 1-μm beads. Enrichment is performed by hybridizing beads to nonmagnetic, low-density ''capture beads'' (dark blue) to enrich the amplified fraction (red) of magnetic ePCR beads. The enriched Beads can be immobilized and mounted in a flowcell for automated sequencing. For each sequencing by ligation cycle, four-color imaging is performed across several hundred raster positions to determine the sequence of each amplified bead at a specific position in one of the tags.

The beads need to be attached on the glass surface of flowcell for sequencing. Firstly, the flowcell is washed and aminosilane-treated, enabling the subsequent covalent coupling of template DNA on it and eliminating any fluorescent contamination. Then the amplified, enriched and capped beads are mixed with a crosslinker BS^3 (Bis[sulfosuccinimidyl] suberate) in PBS and immediately loaded into flowcell, which is then inverted and incubated at room temperature for 1 h. The flowcell can be washed with Tris-HCl to quench the crosslinker reaction before sequencing.

1.2.2.3 DNA sequencing by ligation (SBL)

The biochemistry of Polony sequencing mainly relies on sequencing by ligation (figure 4). First, a series of anchor primers are flowed through the flow cell and hybridize to the synthetic oligonucleotide sequences at the immediate 3’ or 5’ end of the 17-18bp proximal or distal genomic DNA tags. Next, an enzymatic ligation reaction of the anchor primer to a population of degenerate nonamers that are labeled with fluorescent dyes is performed (figure 5). The fluorophore-tagged nonamers are selectively ligated onto the anchor primer, providing a fluorescent signal that indicates whether there is an A, C, G, or T at the query position on the genomic DNA tag. After four color imaging, the anchor primer/nonamer complexes are stripped off and a new cycle is begun by replacing nonamers. A new mixture of the fluorescently tagged nonamers is introduced, for which the query position is shifted one base further into the genomic DNA tag. Six bases from the 5’ to 3’ direction and seven bases from the 3’ end could be queried in this fashion. The ultimate result is a read length of 26 bases per run (13 bases from each of the paired tags) with a 4 to 5 bases gap in the middle of each tag.

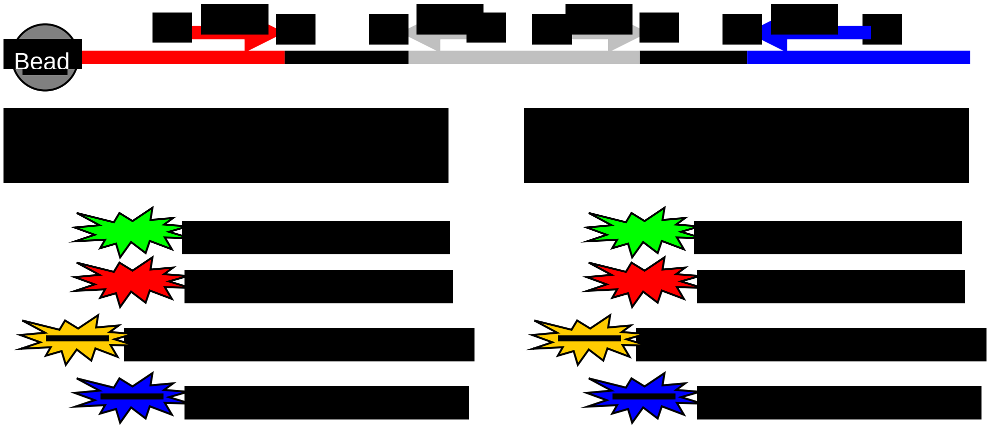


Figure 5. Sequencing by ligation (SBL). SBL can be performed from minus (3’ to 5’) and plus (5’ to 3’) directions, respectively. For Minus SBL, 5’-phosphorylated minus anchor primer (i.e. AP1 or AP3) is annealed to the template, then the complementary 3’ to 5’ query nonamer with one base fixed is ligated to the anchor primer. The fluorescent tags on the nonamers indicate the specific ligated nonamers, which interprets the sequence on the template. The anchor primer and ligated nonamer can be stripped. Another cycle can be proceeded by annealing anchor primer, ligating another set of 3’ to 5’ nonamers, which has another base fixed. For plus SBL, the approach is similar except using plus anchor primer (i.e. AP2 or AP4) and plus nonamers (5’ to 3’query)

1.2.3 Applied Biosystems SOLiD^{TM} sequencing

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is a next-generation sequencing technology developed by Life Technologies and has been commercially available since 2008. This technology can generate hundreds of millions to billions of small sequence reads at a single run [44].

The SOLiD sequencing technology also utilizes bead-coated and emPCR approach as Polony sequencing technology. However, instead of preparing mate-pair library for Polony sequencing, SOLiD uses fragment library (figure 1), which begins by shearing genomic DNA into small fragments and then two different adapters are ligated to the ends of fragments. The proper diluted fragment library can be clonal amplified on 1μm beads by emPCR. After emPCR, the beads can be recovered, enriched, deposited onto a glass surface using same strategy as polony sequencing.

Different to conventional SBL for Polony sequencing, SOLiD approach is based on 2-base encoding ligation sequencing. The sequencing step is basically composed of five rounds and each round consists of about 5-7 cycles (Figure 7). Each round begins with the addition of a P1-complementary universal primer. This primer has, for example, n nucleotides and its 5'-end matches exactly with the 3’-end of the P1. In each cycle, 8-mer probes are added and ligated according to their first and second bases. Then, the remaining unbound probes are washed out, the fluorescent signal from the bound probe is measured and the bound probe is cleaved between its fifth and sixth nucleotide. Finally, the primer and probes are all reset for the next round. In the next round a new universal primer anneals the position n-1 (its 5’-end matches to the base exactly before the 3’-end of the P1) and the subsequent cycles are repeated similar to the first round. The remaining three rounds are performed with new universal primers annealing positions n-2, n-3 and n-4 relative to the 3'-end of P1. A complete reaction of five rounds allows the sequencing of about 25 base pairs of the template from P1.

For decoding the data (Figure 7), which are represented as colors, we must first know two important factors. First, we must know that each color indicates two bases. Second, we need to know one of the bases in the sequence: this base is incorporated in the sequence in the last step. This known base is the last nucleotide of the 3’-end of the known P1. Therefore, since each color represents two nucleotides in which the second base of each dinucleotide unit constitutes the first base of the following dinucleotide, knowing just one base in the sequence will lead us to interpret the whole sequence.

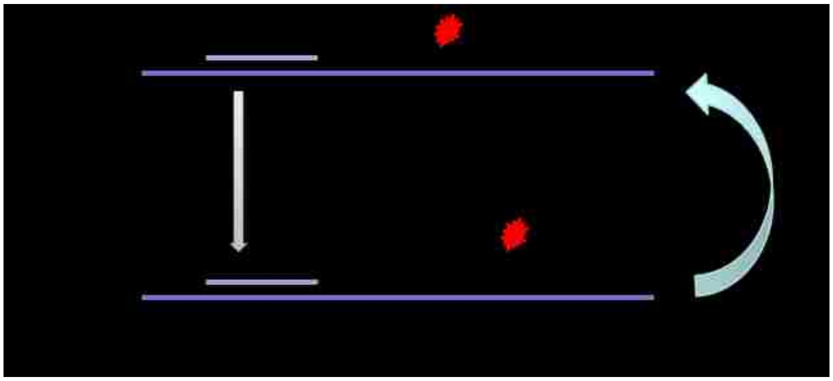


Figure 6. Process of 2-base encoded sequencing by ligation. The DNA sequencing feature is prepared using the same strategy as polony sequencing. Firstly, a primer is annealed to universe adapter sequences on each amplified fragment and then DNA ligase is provided along with specific fluorescent labeled 8-mers, whose 4th and 5th bases are encoded by the attached fluorescent group. Each ligation step is followed by fluorescence detection, after which a regeneration step removes bases from the ligated 8-mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation.

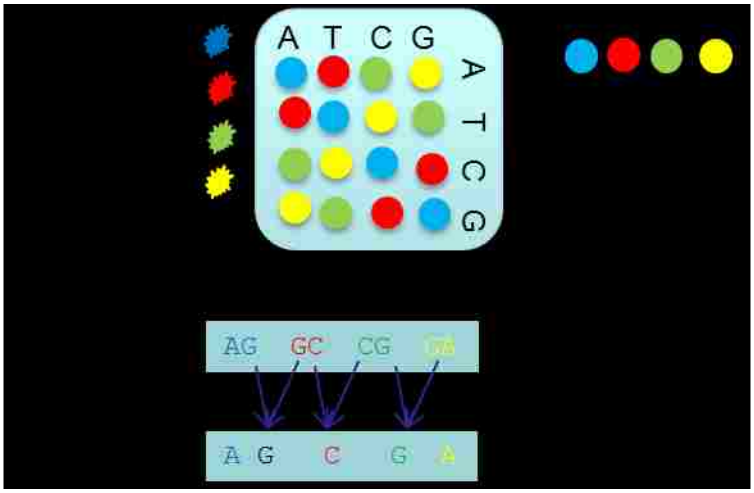


Figure 7.The 2-base encoded Applied Biosystem SOLiD sequencing technology. Because each fluorescent group on a ligated 8-mer identifies a two-base combination, the resulting sequence reads can be screened for base-calling errors versus true polymorphisms versus single base deletions by aligning the individual reads to a known high-quality reference sequence.

1.2.4 Ion Torrent Semiconductor Sequencing

Ion Torrent Semiconductor Sequencing is a DNA sequencing technique based on the detection of hydrogen ions, which are released during the polymerization of DNA using SBS chemistry[23]. Ion semiconductor sequencing may also be referred to as Ion Torrent sequencing, pH-mediated sequencing, silicon sequencing, or semiconductor sequencing. It was developed by Ion Torrent Systems Inc. and was released in February 2010. Ion Torrent has marketed their machine as a rapid, compact and economical sequencer that can be utilized in a large number of laboratories as a bench top machine. Ion Torrent has released 314, 316, and 318 chips with different ability of throughput, meeting various uses. Basically, a microwell containing a DNA-coated bead is flooded with a single species of deoxyribonucleotide (dNTP). If the introduced dNTP is complementary to the leading template nucleotide, it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers a hypersensitive ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence, multiple dNTP molecules will be incorporated in a single cycle. This leads to a corresponding number of released hydrogen and a proportionally higher electronic signal. This technology differs from other sequencing technologies in that no modified nucleotides or optics is used.

The whole process composes of library construction, template preparation on beads, sequencing on a chip, signal processing and base calling (Figure 9). The fragment library and ePCR are performed as previous sections. The beads used here are 3-μm size spheres. The special and key technology for Ion Torrent sequencing is the sequencing chemistry and signal detection system. In nature, the incorporation of a deoxyribonucleotide (dNTP) into a growing DNA strand involves the formation of a covalent bond and the release of pyrophosphate and a positively charged hydrogen ion. A dNTP will only be incorporated if it is complementary to the leading unpaired template nucleotide. Ion semiconductor sequencing exploits these facts by determining if a hydrogen ion is released upon providing a single species of dNTP to the reaction.

Microwells on a semiconductor chip that each contain one single-stranded template DNA molecule to be sequenced and one DNA polymerase are sequentially flooded with unmodified A, C, G or T dNTP (Figure 9). If an introduced dNTP is complementary to the next unpaired nucleotide on the template strand it is incorporated into the growing complementary strand by the DNA polymerase. If the introduced dNTP is not complementary there is no incorporation and no biochemical reaction. The hydrogen ion that is released in the reaction changes the pH of the solution, which is detected by a hypersensitive ion sensor (Figure 8). The unattached dNTP molecules are washed out before the next cycle when a different dNTP is introduced.

As illustrated in figure 9, beneath the layer of micro-wells is an ion sensitive layer, below which is a hypersensitive ISFET ion sensor. All layers are contained within a CMOS semiconductor chip (Figure 8), similar to that used in the electronics industry. Each released hydrogen ion triggers the ISFET ion sensor. The series of electrical pulses transmitted from the chip to a computer is translated into a DNA sequence, with no intermediate signal conversion required. Each chip contains an array of micro-wells with corresponding ISFET detectors. Because nucleotide incorporation events are measured directly by electronics, the use of labeled nucleotides and optical measurements are avoided.

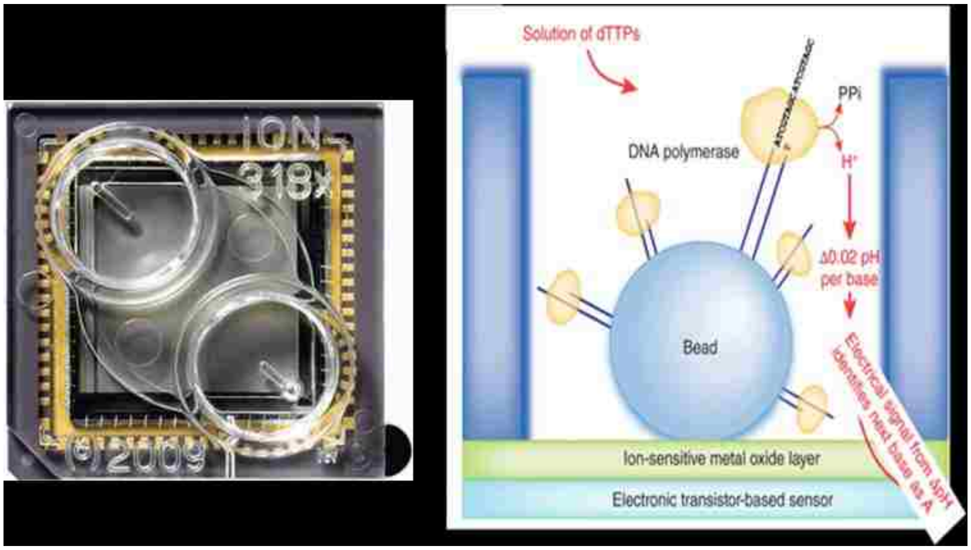

ION
318x
©2009
Solution of dTTPs
PPi
DNA polymerase
H+
Δ0.02 pH per base
Bead
Electrical signal from ΔpH identifies next base as A
Ion-sensitive metal oxide layer
Electronic transistor-based sensor

Figure 8. Ion semiconductor 318 chip and sequencing principle. The top figure shows a simplified drawing of a well, a bead containing DNA template, underlying sensor and electronics. Protons (H+) are released when nucleotides (dNTP) are incorporated on the growing DNA strands, changing the pH of the well (ΔpH). This induces a change in surface potential of the metal-oxide-sensing layer and a change in potential (ΔV) of the source terminal of the underlying field-effect transistor. Sequence on ion chip, sequencing primers and DNA polymerase is bound to the template-carrying beads, beads are pipetted into the chip's loading port. The chip is installed in the sequencing instrument; all four nucleotides cyclically flowed in an automated 2-h run. Signal processing, software converts the raw data into measurements of incorporation in each well for each successive nucleotide flow. After bases are called, each read is passed through a filter to exclude low-accuracy reads and per-base quality values are predicted.

1.3 Nano-scale bead-free sequencing

To simplify and minimizing the DNA sequencing feature, nano-scale bead-free features were developed, which conquered the limitation of size of beads, the complicated process of emulsion PCR and shorten the whole procedure time[45]. The most successful representatives are Illumina cluster sequencing [18] and Complete Genomics nanoball sequencing[17]. Moreover, Rolony (Rolling circle amplified colony) sequencing is an under-developed technology, which has greatly promising to be a significant member of next-generation sequencing technologies. Therefore, this section is aim to describe the features of the first two technologies, while Rolony sequencing by synthesis will be illustrated in chapter 3.

1.3.1 Illumina/Solexa Sequencing

Illumina sequencing technology utilizes clonal array formation and proprietary reversible terminator technology for rapid and accurate high-throughput sequencing. Due to the advanced property of this technology, Illumina platforms have been the biggest commercial system in the market since 2006, and hundreds of papers published using this technology. The main two characteristics are clonal cluster array generation and specific sequencing by synthesis (SBS) using reversible terminators.

Initially, diluted sequencing templates (fragment library) are immobilized on a proprietary flow cell surface (Figure 9) designed to present the DNA in a manner that facilitates access to enzymes while ensuring high stability of surface bound template and low non-specific binding of fluorescently labeled nucleotides. Solid-phase amplification (Figure 9) produces up to 1,000 identical copies of each single template molecule in close proximity (several hundred nanometers diameter) and giving ten million single-molecule clusters per square centimeter.

SBS technology employs four fluorescently labeled nucleotides to sequence the tens of millions of clusters on the flow cell surface in parallel (Figure 9). During each sequencing cycle, a single labeled deoxynucleoside triphosphate (dNTP) is added to the nucleic acid chain. The nucleotide label serves as a terminator for polymerization, so after each dNTP incorporation, the fluorescent dye is imaged to identify the base and then enzymatically cleaved to allow incorporation of the next nucleotide. Since all four reversible terminator-bound dNTPs (A, C, T, G) are present as single, separate molecules, natural competition minimizes incorporation bias (figure 9). Base calls are made directly from signal intensity measurements during each cycle, which greatly reduces raw error rates compared to other technologies.

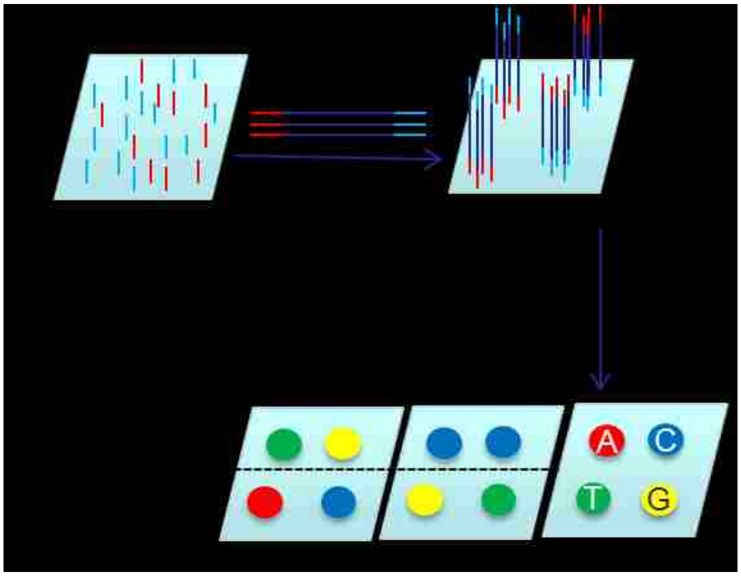


Figure 9. Overview of the Illumina cluster DNA sequencing template prepared by bridge PCR amplification and Illumina sequencing by synthesis. First, all four fluorescently labeled, 3-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation.

1.3.2 Complete Genomics DNA Nanoball Sequencing[17]

DNA nanoball sequencing takes advantage of rolling circle replication (RCR) to amplify single-strand circularized genomic DNA into DNA nanoball. The bases can be determined using conventional sequencing by ligation (SBL) and Unchained sequencing by ligation. Due to the nanoscale of DNA sequencing feature, this technology allows large numbers of DNA nanoballs to be sequenced per run at lower reagent costs compared to other next generation sequencing platforms. However, a limitation of this method is that SBL generates only short reads, which presents challenges when mapping its reads to reference genome, as well as having difficulty to perform *de novo* sequencing assembly.

1.3.2.1 Single-strand circularized template preparation

This method begins with randomly shearing genomic DNA and selecting out 400–500 base pair (bp) fragments, four rounds of ligating adapter sequences, PCR amplifying, restriction enzyme digesting,and circularizing to the fragments. The genomic DNA fragments are undergoing a serial of directional adapter ligation, circle ligation, rolling circle amplification and recursive cutting with type IIS restriction enzymes to form circular ssDNA library.

1.3.2.2 Rolling Circle replication (RCR)

Once a single strand circular DNA template with sample DNA four unique adapter sequences is generated, the template can be amplified into a long string of DNA by RCA using Phi 29 DNA polymerase. Due to strand displacement feature of Phi 29, the new synthesized strand is released from the circular template, and resulting a long single-stranded DNA comprising several head-to-tail copies of the circular template. The design of palindromic sequences inside of the adapters helps the single strand to fold onto itself, resulting in a tight ball of DNA approximately 300 nanometers (nm) across, which also allows the nanoballs to remain separated from each other and reduces any tangling between different nanoballs. The circular ssDNA library is replicated with Phi29 polymerase. Using a controlled, synchronized synthesis, hundreds of tandem copies of the sequencing substrate are obtained in palindrome-promoted coils of single-stranded DNA, referred to as DNA nanoballs (DNBs). DNBs are then adsorbed onto photolithographically etched, surface modified 25- by 75-mm silicon substrates with grid-patterned arrays of ~300-nm spots for DNB binding. B. High accuracy combinatorial probe anchor ligation (cPAL) sequencing chemistry was then used to independently read up to 10 bases adjacent to each of eight anchor sites, resulting in a total of 31- to 35-base mate-paired reads (62 to 70 bases per DNB).

1.3.2.3 DNA nanoball microarray and unchained sequencing by ligation

The DNA nanoballs are attached to a microarray flow cell. The flow cell is a 25 mm by 75 mm silicon wafer coated with silicon dioxide, titanium, hexamethyldisilazane (HMDS), and a photoresist material. The DNA nanoballs are added to the flow cell and selectively bind to the aminosilane in a highly ordered pattern, allowing a very high density of DNA nanoballs to be sequenced. Unchained sequencing reactions interrogate specific nucleotide locations in the nanoball by ligating fluorescent probes to the DNA. The color of the fluorescence at each interrogated position is recorded through a high-resolution CCD camera, and eventually generates 31- to 35bp mate pair reads.

1.4 Single molecule sequencing

Since instruments can detect smaller and smaller objectives, single molecule sequencing arises by expectations. So far, the DNA sequencing features has evolved from millimeters (original polonies), micrometers (bead-coated template), nanometers (Illumina cluster, and nanoball), and to single molecule scale. Helicos Biosciences and Pacific Biosciences announced the most successful single molecule sequencing technologies. Since they are single molecule scale and have very advanced signal detection systems, people have considered them as ‘third-generation’ sequencing technologies[24].

1.4.1 Helicos Biosciences single molecule sequencing

Helicos launched the first True Single Molecule Sequencing (tSMS)™ technology, which is capable of directly measuring single DNA molecules without amplification and biases, the cost and complexity that often accompany it[20,46,47]. Helicos tSMS technology signals a new era of quantitative, functional genomics by offering unparalleled quantitative accuracy, simplicity and scale. The tSMS technology is a powerful new method to accurately interrogate billions of single strands of DNA (or RNA) in parallel by directly detecting single nucleotide incorporations on each of the single strands. This unprecedented detection sensitivity is made possible by Helicos' proprietary ultra-low background surface chemistry, sequencing reagents and imaging technology.

Similar to the other sequencing technologies, tSMS approach also needs to fragment the original DNA samples. After size selection, the DNA double-helix fragments are melted into single strands and a polyA tail is added to these DNA molecules (Figure 10). Billions of these single DNA molecules are captured on a proprietary surface within a flow cell and serve as templates for the sequencing-by-synthesis process.

During the SBS process, based on the sequence of the template, fluorescent labelled nucleotides: C, G, A or T are added one at a time and incorporated into the growing complementary strand by a DNA polymerase (Figure 10). Unincorporated nucleotides are then washed away. Upon illumination with a laser, the incorporated fluorescent nucleotides emit light that is detected by the HeliScope™ Single Molecule Sequencer. The HeliScope Sequencer captures thousands of images across the flow cell surface and records the strands that have been incorporated with nucleotides. These images containing tens of thousands of single fluorescent molecules are akin to star fields. Once the imaging of the flow cell is complete, the fluorescent label is removed and next cycle is ready to proceed. Tracking nucleotide incorporation on each strand determines the exact sequence of each individual DNA molecule.

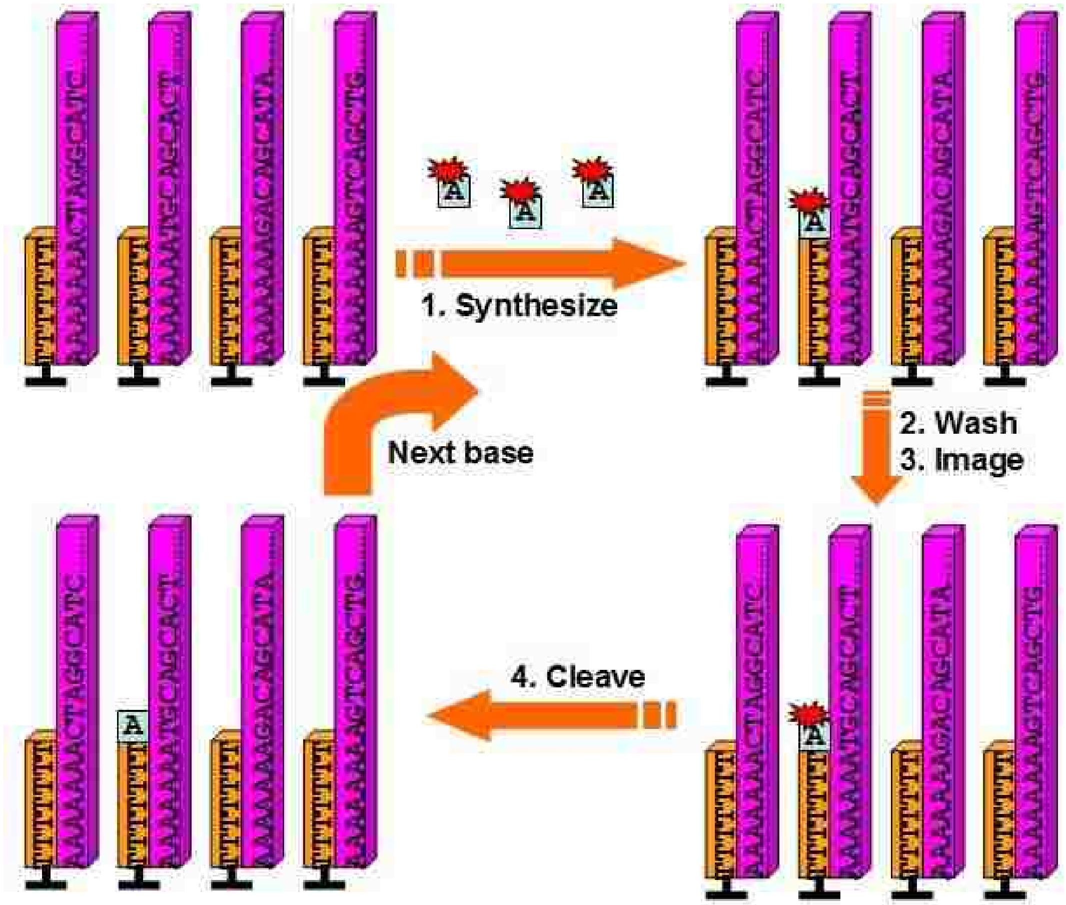


Figure 10. Overview of Helico BioSciences single molecule sequencing technology. After hybridization of modified genomic DNA to oligo dT capture primers inside the flow cell, acquisition of template positions, and cleaving the fluorescent templated lablel, the tSMS process is a cyclical process involving multiple rounds of (1) synthesis using labeled nucleotides, (2) washing, (3) imaging and (4) cleaving the fluorescent label until the desired read length is achieved.

Helicos tSMS technology offers significant advantages over both traditional and “next generation” sequencing technologies. It is the first PCR-free universal genetic analysis platform and pursues a single molecule sequencing strategy, which simplifies the DNA sample preparation process, avoids PCR-induced bias and errors, simplifies data analysis and tolerates degraded samples.

1.4.2 Pacific Biosciences single molecule real time (SMRT) sequencing

Single molecule real time sequencing (SMRT) is a parallelized single molecule DNA sequencing by synthesis technology developed by Pacific Biosciences[22]. SMRT utilizes the zero-mode waveguide (ZMW) [48], a structure that creates an illuminated observation volume that is small enough to observe only a single nucleotide being incorporated by DNA polymerase. Remarkably, a single DNA polymerase enzyme is affixed at the bottom of a ZMW with a single molecule of DNA as a template (Figure 11). Each of the four nucleotides is labeled with one of four different fluorescent dyes on terminal phosphate group. When a nucleotide is incorporated by the DNA polymerase, the fluorescent tag is cleaved off and diffuses out of the observation area of the ZMW where its fluorescence is no longer observable (Figure 11).

Briefly, the SMRT sequencing is performed on a chip that contains thousands of ZMW. Inside each ZMW, a single active Phi 29 DNA polymerase with a single molecule of single stranded DNA template is immobilized to the bottom through which light can penetrate and create a visualization chamber that allows monitoring of the activity of the DNA polymerase at a single molecule level. The signal from a phospho-linked nucleotide incorporated by the polymerase is detected as the DNA synthesis precedes which results in the DNA sequencing in real time.

For each of the nucleotide bases, there are four corresponding fluorescent dye molecules that enable the detector to identify the base being incorporated by the DNA polymerase as it performs the DNA synthesis. The fluorescent dye molecule is attached to the phosphate chain of the nucleotide. When the nucleotide is incorporated by the DNA polymerase, the fluorescent dye is cleaved off with the phosphate chain as a part of a natural DNA synthesis process during which a phosphodiester bond is created to elongate the DNA chain. The cleaved fluorescent dye molecule then diffuses out of the detection volume so that the fluorescent signal is no longer detected.

Pacific Biosciences was expected to commercialize SMRT sequencing in 2011. The prototype of the SMRT chip contains ~3000 ZMW holes that allow parallelized DNA sequencing. Each of the ZMW holes produces approximately 1,500bp (base pair) read lengths at a speed of 10bp per second.

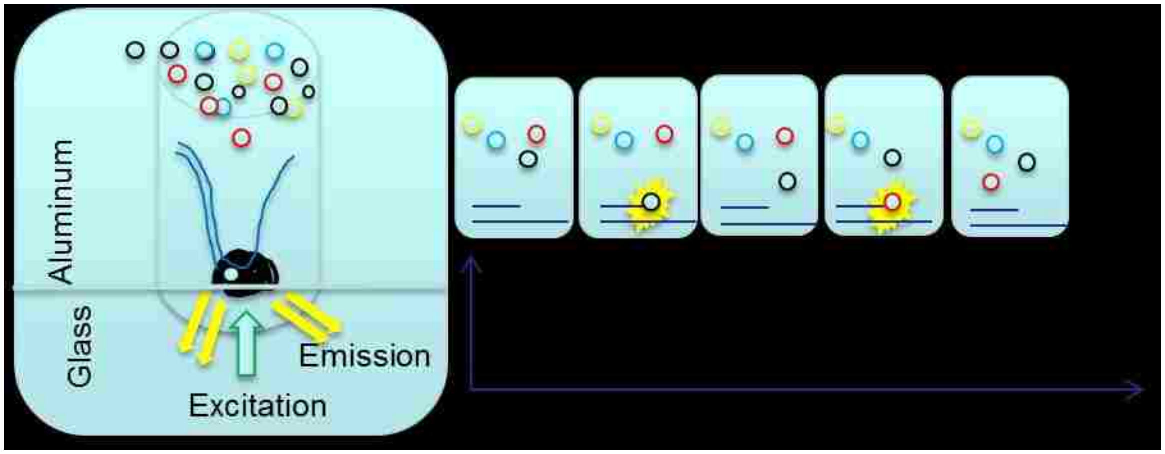


Figure 11. Principle of single-molecule real-time DNA sequencing. (A) Experimental geometry. A single molecule of DNA template-bound Φ29 DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement in the zeptoliter (10–21 liter) regime, enabling detection of individual phospholinked nucleotide substrates against the bulk solution background as they are incorporated into the DNA strand by the polymerase. (B) Schematic event sequence of the phospholinked dNTPs incorporation cycle, with a corresponding expected time trace of detected fluorescence intensity from the ZMW. (1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocate to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse.

1.5 Summary

All these NGS technologies have their own features. 454 sequencing provides the longest reads, up to 1kb, but it is limited by low throughput and very high cost. Illumina sequencing technology gives highest throughput, but the reads are short, 2x100 on Hiseq, or 2x250 on MiSeq. SOLiD and Polony sequencing are limited by short reads, but the cost is relative cheaper. Nanoball array sequencing is also limited by short reads and complicated library construction. Helicos and Pacific Biosciences single molecule sequencing technologies are promising, but the low quality data has to be improved. The price for sequencing human genome has dropped to $4000, which is near the $1000 goal.

However, next generation sequencing has driven the whole biomedical sciences moving forward to a next level[49]. In the past few years, next generation sequencing technology has been applied to all kinds of biomedical research[35,50-64]. In 2008, 454 sequencing was employed to sequence James Watson's genome[65], and just a few months before that, Crag Venter's genome was published, to announce the first human genome[66]. Since then, there have been several more human genomes been sequenced[17,18,67-71].

Cancer genome research is a major application for NGS technology[32,33,37-39,59,72-114]. In the last few years, there are thousands cancers have been sequenced[32,33,37-39,59,73-114]. A couple of national cancer genome projects are undergoing, such as 'Cancer Genome Project' leading by Wellcome Trust Sanger Institute and 'Cancer Genome Atlas' launched by national human genome research institute. All this will improve our understanding of human cancer, paling the road to eventually cancer therapy.

Exome sequencing is another significant application of NGS technologies. In the last few years, thousands of Exomes have been sequenced as well as exome sample across a variety of diseases[29-31,51,52,59,63,108,110,115-185]. The exome contains all coding regions and it composes about 1% of whole genome, ~180,000 exons. Remarkably, this 1% region contains 85% of the genetic changes that may be the cause of a variety of diseases. Due to the fact that there is minimum technology to study the non-coding region to date and the exome is much smaller than whole genome, exome sequencing is getting more popular.

Moreover, there are other applications. ChIP-seq is the example of using NGS technology to study protein-DNA interactions[186-190]. RNA-seq utilizes NGS technology to get higher resolution of RNA study[60,190-198]. NGS has also been used for research that could not been done before[34,35,199-202].

In a word, NGS technology will drive all kinds of research to personalized medicine, as well as improving other aspects of human life.

Therefore, to better serve the biomedical research and clinical applications, there are lots spaces we can improve the next generation sequencing technology. For this dissertation, I focused on the following three projects to improve the NGS technology and its applications. Firstly, dual primer ePCR technique was developed to obtain longer reads for Polony sequencing; more remarkably, it provided paired-end reads which is very helpful for chromosome structure translocation study and genome mapping. This technique also simplifies the library construction process, which shortens the sequencing turnover time. Secondly, Rolony sequencing by synthesis technology was developed in chapter 3. Rolony sequencing not only simplifies the library construction process, it also dramatically increases the throughput. Both SBL and SBS can be performed on Rolony patterned on grid array flowcell. Finally, I demonstrated the application of NGS technology by developing whole genome amplification technique for haplotype-resolved sequencing. Ion Torrent, Ion Proton, HiSeq 2000 and MiSeq sequencers were utilized to sequence human genome library, over 300 whole genome amplification libraries, and human exome library.

CHAPTER 2

DUAL PRIMER EMULSION PCR FOR NEXT-GENERATION SEQUENICNG

Ming Yan Xu [1], Anthony D. Aragon [1], Monica R. Mascarenas [1], Norah Torrez-Martinez [1],

Jeremy S. Edwards [1, 2, 3]

[1] Molecular Genetics and Microbiology, University of New Mexico Health Sciences Center, Albuquerque, NM 87131

[2] Chemical and Nuclear Engineering, University of New Mexico, Albuquerque, NM 87131

[3] Cancer Research and Treatment Center, University of New Mexico Health Sciences Center, Albuquerque, NM 87131

Key Words: Next-generation Sequencing, Emulsion PCR, Bridge PCR.

Note: this chapter was published on Biotechniques 2010

Abstract

We have developed a highly sensitive single molecule clonal amplification method called Dual Primer Emulsion PCR (DPePCR) for next-generation DNA sequencing. The approach is similar in concept to standard emulsion PCR, however, in DPePCR both primers are attached to the beads, therefore following PCR amplification, both strands of the PCR products are attached to the beads. The ends of each strand can be freed for analysis by restriction digestion of the bridged PCR fragments, which allows efficient paired-end sequencing of fragment libraries.

The concept of performing single molecule biology in high-throughput using emulsions is not new and has been used for single molecule evaluation for a number of years [203] before the approach was first applied for next-generation sequencing [16]. However, for emulsion technology to be applicable to next-generation sequencing, a method for capturing the contents from the droplets is required. To accomplish this, Dressman el al [204] reported an approach to amplify single DNA molecules onto beads for detection and enumeration of genetic variation. They named their technology BEAMing: beads, emulsion, amplification and magnetics. This approach has now evolved to "emulsion PCR" or ePCR. The DNA bound to beads generated from ePCR provides an excellent template for high-throughput sequencing because of the ability of PCR amplifies a single molecule of DNA resulting in clonal molecules per bead[16]. Due to that fact a number of next-generation sequencing approaches utilize emulsions and beads for sequencing[10-13,205,206].

An alternative approach for amplification for DNA for next-generation sequencing is the Bridge Amplification strategy (used by Illumina) which was originally reported by Bing et al (Bridge Amplification: A Solid Phase PCR System for the Amplification and Detection of Allelic Differences in Single Copy Genes http://www.promega.com/geneticidproc/ussymp7proc/0726. html). Bridge amplification is a technology that uses a single aqueous compartment, however, the individual amplicons are constrained by primers bound to a solid phase that are extended and amplified [40]. As the name implies, the extension product from one bead-bound primer forms a bridge to the other bead bound primer.

In this paper we describe a novel approach called Dual Primer emulsion PCR (DPePCR) which combines concepts from both emulsion PCR and Bridge Amplification for the generation of simple fragment libraries for paired-end next-generation sequencing. The DPePCR strategy can amplify ~200 base pair fragments (with the adaptors) and enable sequencing of both ends of the DNA fragment. This effectively shortens library preparation time, and increases the library complexity[16] when compared to the construction of a mate paired DNA library, and will definitely contribute to the $1,000 genome goal.

Normal ePCR has been extensively applied to next-generation DNA sequencing[43]. Most of the next-generation sequencing approaches are restricted to short read lengths [205], and therefore to optimize resequencing of a human genome, mate-paired or paired-end sequencing is important. However, the construction of mate-paired libraries for next-generation sequencing is difficult and time-consuming[43]. Fortunately, fragment library construction is quite simple, and the described DPePCR strategy enables paired-end sequencing of fragment libraries with essentially no modification of existing sequencing approaches.

To perform DPePCR, both forward and reverse primers are attached to 1 μm Myone beads (Figure 12B) (see Supplemental Material) that are included in a modified ePCR protocol[43]. Additionally, since the amplicons are confined to the droplets, the amplification efficiency is increased by including free primers in the aqueous phase (Figure 12C). After 120 PCR cycles, the single DNA fragment in the emulsion can be amplified effectively. After amplification, we have found the DNA to be highly stable in the double stranded state. Extreme conditions will denature the DNA; however, since both strands are present on the bead, the double stranded state immediately reforms, which inhibits the ability to sequence the DNA. To overcome this issue, type IIs recognition enzyme sites (i.e. BceAI and AcuI) were placed on two ends of the amplicons being amplified immediately adjacent to the unknown sequence during library construction (Figure 12A). The DPePCR product can then be digested with restriction enzymes (i.e. BceAI and AcuI), a capping adaptor is ligated to the free end of the dsDNA (Figure 12D). This gives us the ability to sequence from both strands of the DPePCR using standard sequencing by ligation (Figure 12D) from paired ends of 200-300bp.

The sequencing strategy for DPePCR beads is identical to sequencing from standard ePCR beads (Figure 12D). The difference is that since there are two paired-end fragments, both can be sequenced independently from both the 3’->5’ and 5’->3’ directions using four different anchor primers. Therefore, in addition to allowing paired-end sequencing, the DPePCR strategy increases the total number of bases read per bead.

To validate the formation of strong double strand DPePCR product, the beads were treated with 0.1M NaOH (without restriction enzyme digestion). After denaturing, a Cy5 labeled oligo was annealed to the bead-bound DNA fragment (Supplement Material). The results indicated that the Cy5 Oligo could not hybridize to the DPePCR product, which suggests the formation of double strand DPePCR product was in a “bridged” confirmation, since we were able to sequence the DNA on the beads following digestion with restriction enzymes.

Theoretically, the DNA fragment on each bead arose from a single molecule in an emulsion and should therefore be clonal, which is the critical requirement for the DPePCR. To validate the clonal amplification, we used sequencing by ligation to sequence the DNA fragment; the clonal amplified beads showed the

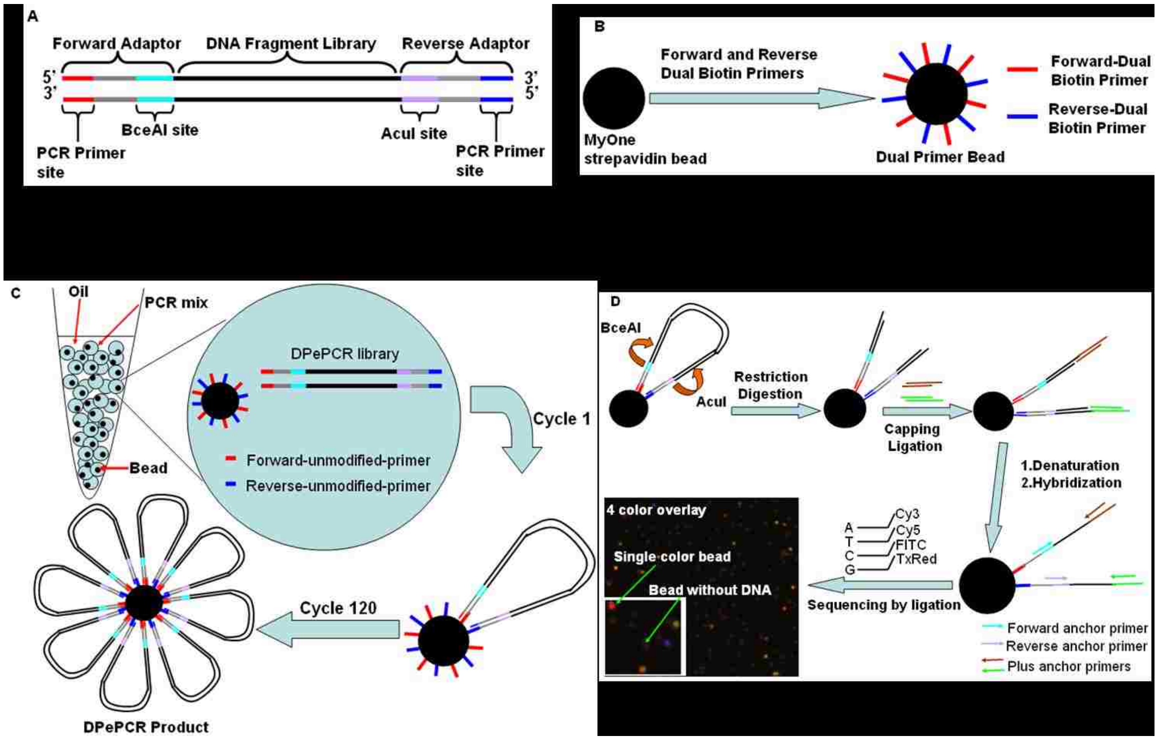


Figure 12. Dual Primer emulsion PCR approach. A. The DPePCR library**:** 100-200 base pair DNA fragments were adapted by FDV2-BceAI-Primer-F/R and RDV2-AcuI-Pimer-F/R. Restriction enzymes BceAI and AcuI recognition sites were designed at the ends of the two primers for digestion of the DPePCR product. **B.** Dual Primer beads: equal amount of FDV2-dualbiotin and RDV2-dualbiotin were attached to MyOne strepavidin beads at the same time. **C.** Dual primer beads, diluted DNA fragment, ePCR reagents and two kinds of free short unmodified primers were mixed with silicone oil phase. Each microdroplet contains single bead, single DNA molecule and enough PCR reagents will be clonal amplified. After the first cycle, one FDV2-dualbiotin strand and one RDV2-dualbiotin strand were amplified and fold into a bridge because of the strong covalent bond of the two complementary DNA fragment. After 120 cycles, millions of the bridges were formed. **D.** Digestion and sequencing of the DPePCR beads: After the emulsion

amplification, the DPePCR beads were digested with BceAI and AcuI. The digestion will leave 12 (on BceAI strand) and 16 (on AcuI strand) bases of unknown DNA fragment for sequencing. Specific anchor primer was hybridized at the end of the FDV2 strand or RDV2 strand, following by ligating a single base fixed fluorescent nonamers. The clonal amplified beads showed the different single fluorescent color under fluorescent microscope.

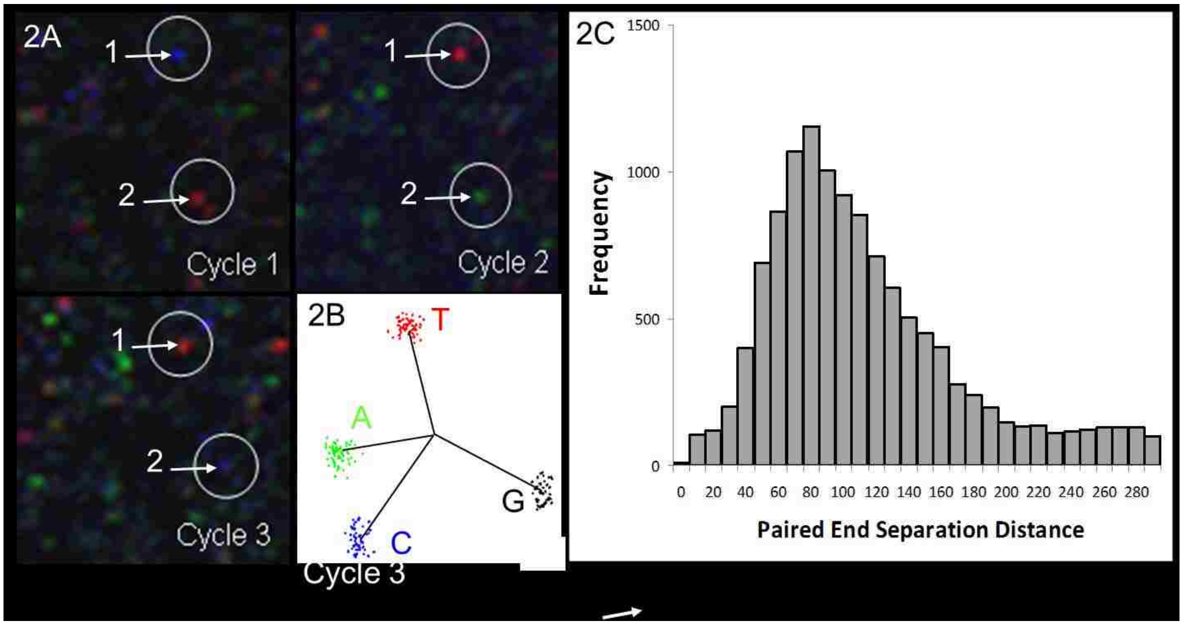


Figure 13. Validation of Polonator DNA sequencing image. 2A. The DPePCR beads were sequenced by ligation with annealing anchor primer and ligated fluorescent nonamers. Green, red and blue beads relate to A, T and C respectively. The sequence for bead 1 is CTT; bead 2 is TAC from the three cycles. **2B**.The tetrahedral plot is a two dimensional projection of the 4 fluorescent intensities plotted. Since each of the beads is a clonal amplicon, the data points are clustered into 4 groups. **2C**. A random sampling of ~21,000 reads was mapped uniquely to the

Streptococcus pyrogenes genome (AE014074). The histogram shows the average separation in the genome is about 100 bases

single color during each cycle (Figure 13), which indicated the beads were clonal amplified. Additionally, a random sampling of ~21,000 reads was mapped uniquely to the Streptococcus pyrogenes genome (AE014074). The reads were of the form NNNNNNN (~100bases) NNNNNN.NNNNNNN. The average separation in the genome was about 100 bases (Figure 13), which is consistent with the size range that selected in the gel purification step. Additionally, the library complexity was high, with 99.3% of the reads being unique, thus improving upon the traditional mate-pair library production protocol [16,43].

Acknowledgments

The work was supported by NIH/NHGRI $1,000 Genome Grant (R21HG004350) to JSE.

Competing interests

The authors declare no competing interests.

Supplement material

Materials and Methods

The DPePCR library was prepared by fragmenting the genomic DNA by sonication and PAGE purifying 100-200 base pair genome fragments. The DNA was blunt ended repaired and A-tailed. Then FDV2-BceAI-Primer-F/R and RDV2-AcuI-Primer-F/R (Supplementary Material) were ligated to the DNA fragments and PCR amplified for 13 cycles. The PCR products were then gel purified to isolate the DPePCR library.

To PCR amplify the library for sequencing, 2μL of appropriate concentration (to allow single molecule per drop) DPePCR library was mixed with 265 μL DPePCR master mix (Supplement Material), 13 μL of JumpStar Taq (20,000 U/mL), 0.1 μM FDV2-Short-Unmodified-Primer and 0.1 μM RDV2-Short-Unmodified-Primer and 20 μL Dual Primer beads (Supplement Material). The mixture was added to 400 μL Silicone Oil Phase in a 2 mL tube and put on a TissueLyser II (Qiagen) to emulsify (19 Hz for 90 seconds). The well mixed DPePCR emulsion mixture was aliquoted to three 650 μL tubes and thermocycled (Supplementary Material). The emulsion beads were broken and cleaned using the protocol in the supplementary information. Finally, the beads were suspended in 100 μL TE with 0.01%Triton.

The DPePCR beads were digested with 5 μL AcuI (5000 U/mL) (NEB), 0.8 μL S-adenosylmethionine (SAM) (32 mM) and 394.5 μL 1x NEBuffer 4 by incubating at 37°C for 60 min. After AcuI digestion, the beads were washed 3x with NXS and TE with 0.01% Triton, and digested with 4μL BceAI (1000 U/mL) (NEB), 4 μL Bovine Serum Albumin (10 mg/mL) and 392 μL 1x NEBuffer 3 by incubating at 37°C for 60 min. After BceAI digestion, the beads were washed 3 times with NXS and TE with 0.01% Triton. Before ligating capping adaptors, the beads were treated with 2 μL Antarctic Phosphatase (5000 U/mL) (NEB) and 198μL 1x Antarctic Phosphatase Reaction Buffer at 37°C for 60 min to remove 5´ phosphoryl groups from nucleic acids. After Dephosphorylation, the beads were washed 3 times with NXS and TE with 0.01% Triton. For capping adaptors ligation, the beads were incubated with 2.5 μL AcuI-PlusPrimer-3'over-F/R (50 μM), 2.5 μL BceAI-PlusPrimer-3'over-F/R (50 μM), 2 μL T4 DNA ligase (600,000 U/mL) (Enzymatics) and 193 μL 1x quick ligation buffer at room temperature for 60min. The beads were treated with 0.1M NaOH and washed three times with 1x SSPE (Fisher Scientific).

The DPePCR beads were sequenced from the FDV2 and RDV2 strand respectively using the same strategy from both the 3'->5' and 5'->3' directions. For test sequencing of the FDV2 strand in 3'->5' direction, the beads were resuspended in 135 µL 1xSSPE with 15 µL anchor primer FDV2-PM (100 µM), incubated at 70°C for 5 min, 50°C for 5 min and 25°C for 5 min, and washed three times with 1E (see supplemental Material). Then 2 µL T4 DNA ligase and 200 µL ligation buffer (2 µL of each nonamer (300 µM) (IDT) and 200 µL 1x T4 DNA quick ligase buffer) were added to the beads and incubated at room temperature for 30 min. The beads were then washed with 1E buffer and resuspended in 50 µL 1E for observation on fluorescent microscope. For test sequencing of the RDV2 strand in 3'->5' direction, the beads were treated with 0.1M NaOH and then hybridized anchor primer RDV2-PM and ligated nanomers using the same protocol as FDV2 strand. For 5'->3' direction sequencing, the beads need to be capped with dideoxynucleotide (Supplemental Material) before hybridization and ligation.

Bind forward PCR primer to microbeads

To bind forward PCR primer to microbeads in a 1.5 mL tube add 240µl B&W buffer to 240 µL MyOne streptavidin (Invitrogen) beads. Mix sample by inversion. Place on MPC to remove liquid by pipetting. Wash beads two times with 480 µL of B&W buffer. Resuspend beads in 480 µl of B&W then add 4.8 µL of 1mM FDV2- and RDV2-dual biotin, vortex, incubate at room temperature for 20 min with mixing. Wash beads three times with 600 µL of B&W buffer. Resuspend beads in 640 µL of TE.

Silicone Oil Phase is prepared by mixing 10 g Silicone oil AR20 (Sigma), 7.5 g 749 (Dow Chemicals) and 7.5 g 5225 (Dow Chemicals).

Bind and wash buffer

5 mM Tris·Cl, pH 7.5

0.5 mM EDTA, pH 8.0

1.0 M NaCl

NXS buffer

10 mM Tris·Cl, pH 7.5

1 mM EDTA, pH 8.0

100 mM NaCl

1% Triton X-100

Wash 1E wash buffer

10 mM Tris·Cl, pH 7.5

50 mM KCl

2 mM EDTA, pH 8.0

0.01% (v/v) Triton X-100

Preparation FDV2-BceAI-Primer-F/R (50 µM), RDV2-AcuI-Primer-F/R (50 µM), AcuI-PlusPrimer-3’over-F/R (50 µM), BceAI-PlusPrimer-3’over-F/R (50 µM)

Anneal Primers：

Mix 20 µL FDV2-BceAI-Primer-F (100 µM) and 20 µL FDV2-BceAI-Primer-R (100 µM); 20 µL RDV2-AcuI-Primer-F (100 µM) and 20 µL RDV2-AcuI-Primer-R (100 µM); 20 µL BceAI-PlusPrimer-3’over-F (100 µM) and 20 µL BceAI-PlusPrimer-3’over-R (100 µM); 20 µL AcuI-PlusPrimer-3’over-F (100 µM) and 20 µL AcuI-PlusPrimer-3’over-R (100 µM) respectively and then follow the parameters:

1. 80°C-----2 min
2. 70°C-----2 min
3. 60°C-----5 min
4. 50°C-----5 min
5. 40°C-----5 min
6. 30°C-----5 min
7. 20°C-----5 min

8. 4°C-----Hold

Library PCR cycle Parameters

1. 94°C-----2 min

2. 94°C-----30 sec

3. 58.8°C---30 sec

4. 72°C----- 90 sec

5. Go to step 2-----13 times

6. 72°C-----5 min

7. 4°C-----Hold

Library PCR Master Mix

Reagent	Final Concentration
10x Jumpstart Buffer	1x
dNTPs	0.2 mM
FDV2-BceAI-PCR	0.5 µM
RDV2-AcuI-PCR	0.5 µM
Jumpstart Taq	0.01 U/µl

M B Water	-

Dual Primer ePCR Master Mix:

Reagent	Final Concentration
10x EXT	1.2X
MgCl2 (50 mM)	7.0 mM
dNTPs (10 mM)	1.0 mM
Triton (10%)	0.1%
BSA (10 mg/ml)	1 mg/ml
FDV2 short unmodified primer	0.1 µM
RDV2 short unmodified primer	0.1 µM
Glycerol (50%)	5%
JumpStart Taq	0.1 U/µl
M B Water	-

DPePCR Cycle Parameters

1. 94°C -----2 min

2. 90°C -----15 sec

3. 56°C -----2 min

4. 72°C -----45 sec

5. Go to Step 2----99 times

6. 90°C -----15 sec

7. 61°C -----5 min

8. Go to 6—19 times

9. 72°C -----5 min

10. 4°C -------Hold

DPePCR emulsion breaking procedure

Add 400 µL silicon oil to original PCR tubes, vortex, and pool into 1.5 mL tubes. Spin 1 min at top speed remove aqueous wash two more times with 500 µL silicon oil. Wash the beads three times as follows: 1) 300 µL 100% isopropanol (place tubes on MPC allow beads to collect), 2) 300 µL 80% EtOH/annealing buffer, 3) 300 µL 1X NXS (pool all samples into new tubes at first wash), 4) 300 µL 1x TE/0.01% Triton. After final wash resuspend beads in 100 µL 1x TE/0.01% Trition.

Capping with dideoxynucleotide reaction

Mix the beads with:

15 µL $CoCl_2$ (2.5mM)

114 µL ddH_2O

15 µL 10x tailing buffer

4 µL 1.25 mM (each) dideoxynucleotide mix

2 µL terminal transferase (20000U/mL) (NEB)

Incubate at 37°C for 60min. Wash 3 times with NXS, 1xTE with 0.01%Triton, and PBS.

Oligos (all purchased from IDT)

Primers	Sequence of the primers
FDV2-BceAI-Primer-F	5'-AAC CAC TAC GCC TCC GCT TTC CTC TCT ATG TCT ACT AGT CAA CGG CT -3'
FDV2-BceAI-Primer-R	5' -/5Phos/ GCC GTT GAC TAG TAG ACA TAG AGA GGA AAG CGG AGG CGT AGT GGT T/3AmMO/-3'
RDV2-AcuI-Primer-F	5'-ATA TGT CAA CTG CCC CGG GTT CCT CAT TCA CTC ATG ACT ACT GAA GT-3'
RDV2-AcuI-Primer-R	5'-/5Phos/CTT CAG TAG TCA TGA GTG AAT GAG GAA CCC GGG GCA GTT GAC ATA T/3AmMO/-3'
FDV2-BceAI-PCR	5'-AAC CAC TAC GCC TCC GCT TTC-3'

RDV2-AcuI-PCR	5’-CTG CCC CGG GTT CCT CA -3’
FDV2-dualbiotin	5’-/52-bio/TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT AAC CAC TAC GCC TCC GCT TTC CTC TCT ATG-3’
RDV2-dualbiotin	5’-/52-Bio/ TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTA TAT GTC AAC TGC CCC GGG TTC CTC ATT C-3’
FDV2-short-unmodified-primer	5’-AAC CAC TAC GCC TCC GCT TTC CTC-3’
RDV2-short-unmodified-primer	5’-A TAT GTC AAC TGC CCC GGG TTC CT-3’
BceAI-PlusPrimer-3’over-F	5’-/5Phos/NNT ATC ACT CAG ACT ATA CTC TAA ACC A-3’
BceAI-PlusPrimer-3’over-R	5’-TAG AGT ATA GTC TGA GTG ATA/3AmMO/-3’
AcuI-PlusPrimer-3’over-F	5’-/5Phos/TCT ACA TAT TCT ACC ATC TAC TTA TCA CAC A-3’
AcuI-PlusPrimer-3’over-R	5’-GAT AAG TAG ATG GTA GAA TAT GTA GAN N/3Phos/-3’
FDV2-PM	5’-/5Phos/AGC CGT TGA CTA GTA GAC-3’
RDV2-PM	5’-/5Phos/ACT TCA GTA GTC ATG AGT-3’
BceAI-Plus-AchorPrimer	5’-TAG AGT ATA GTC TGA GTG ATA-3’

AcuI-Plus-AchorPrimer	5’-GAT AAG TAG ATG GTA GAA TAT GTA GA-3’

-1 sequencing nonamers	**+1 sequencing nonamers**
5' Cy5-NNNNNNNNT	5’-/5Phos/TNN NNN NNN/3Cy5/-3’
5' Cy3-NNNNNNNNA	5’-/5Phos/ANN NNN NNN/3Cy3/-3’
5' TexasRed-NNNNNNNNC	5’-/5Phos/CNN NNN NNN/3TxRed/-3’
5' FAM-NNNNNNNNG	5’-/5Phos/GNN NNN NNN/3FAM/-3’

CHAPTER 3

ROLONY SEQUENCING BY SYNTHESIS

3.1 Introduction

High-throughput sequencing technologies have succeeded in spatially and temporally increasing the amount of genetic information that can be gathered through miniaturization and/or rapid sample processing[9,11,15-18,23,28,46,207,208]. However, limitations still exist in terms of cost, speed, and throughput[209-211]. 454 sequencing, SOLiD, Ion Torrent and Polonator sequencing are all emulsion bead based sequencing technologies, with sequencing feature range from 1 μm to 28μm diameter. Because the throughput is limited for Ion Torrent and 454 sequencing technologies, also pyrosequencing for 454 sequencing, make it more expensive than other sequencing technologies. Although Illumina sequencing occupied the biggest sequencing market due to its successful sequencing by synthesis chemistry and high throughput data generation, the price and sample turnover time are still obstacles for most potential applications[10-12]. HeliScope[46] and Pacific Bioscience SMRT sequencing were supposed to sequence single DNA molecule without template amplification, which and have been classified as the third generation sequencing technology, but the immature system has delayed the applications for biomedical researchers[212]. Nanopore sequencing[25,27,213-216] was promising when they proposed a flash drive sequencer in early 2012, but the announcement date has been delayed several times, which make it hard to be considered as a viable sequencing technology.

Almost all the next generation sequencing technologies focus on minimizing the sequencing feature, extending the read length, improving the throughput, and reducing the cost. A Self-assembling DNA nanoarray method was developed to sequencing human genome using unchained base reads[17]. Genomic DNA undergoes a serial fragmentation, ligation, restriction enzyme digestion, PCR, and rolling circle amplifications to construct a circularized single strand DNA library. Then the circularized DNA library was amplified by rolling circle amplification using Phi29 DNA polymerase. The DNA nanoball generated was patterned on glass surface and sequenced using high-accuracy combinatorial probe anchor ligation (cPAL) sequencing. As shown in figure 14, the nanoball was smaller than bead/polony; meanwhile, the throughput was much higher since every circularized DNA molecule formed a sequencing feature—nanoball, while beads/emulsion PCR strategies can only enrich partial monoclonal beads as final sequencing features. The nanoballs were highly density patterned on the array surface which also increased the throughput comparing to the random beads array[45]. Although nanoarray sequencing provided the greatest sequencing feature, the extremely complicated library preparation process and read length limitation of ligation based cPAL sequencing hindered its widespread application.

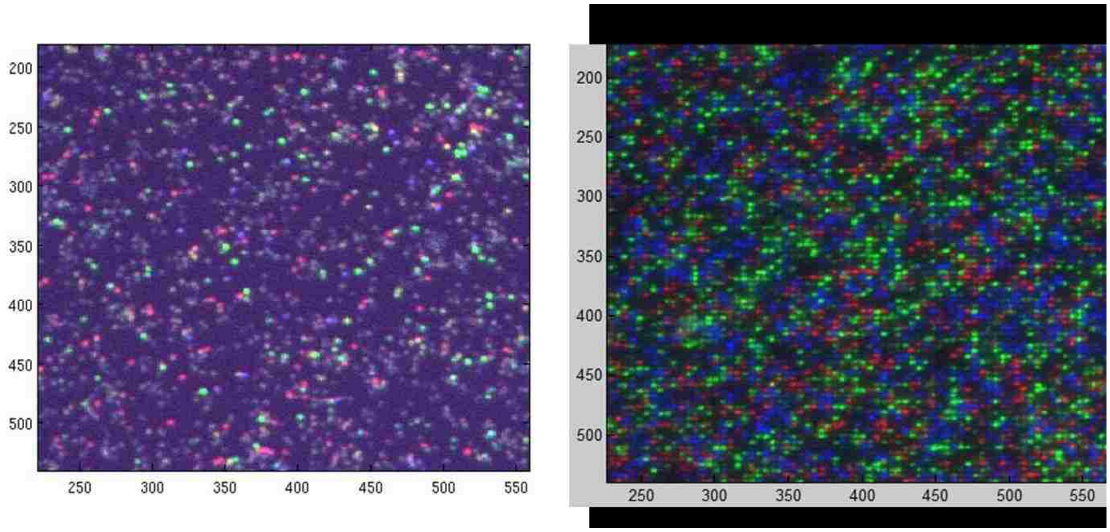


Figure 14. Comparison of bead sequencing and nanoball sequencing. The left image shows the beads are bigger than nanoballs which are shown on the right image. More importantly, there are lots blank beads with no DNA template, which reduces the throughput, while for nanoball image, all the features are monoclonal add patterned on grid surface which increased throughput at least 100x time comparing to the bead sequencing.

Considering all above, the primary aim of our research was to build on and develop faster, more robust, and inexpensive sequencing tools, which can then be applied to study human genetic diseases as well as cellular evolution[35,53-64,217-223]. The Church lab at Harvard Medical School and Edwards lab[224] at the University of New Mexico have been developing and improving polony sequencing technology[16,42,43,225-229]. The fact that polony technology utilizes only a single step to generate billions of “distinct clones” for sequencing has led to a paradigm shift in DNA sequencing technology[16,224]. The primary aim of this paper was to build on our strengths and drive the technology further by creating higher density sequencing features---Rolony (Rolling Circle Amplified Colony) technology. Furthermore, a novel sequencing by synthesis[230-236] was employed to sequence rolony on random array surface and patterned array surface.

In this chapter, the methods of preparation fragment DNA library, the formation of Rolony, and sequencing Rolony by ligation and synthesis are elaborated. Finally, a brief discussion about the potential applications of Rolony technology and summary were given at the end of this chapter.

3.2 Methods and Experiments Design

Rolony was rolling circle amplified colony. Rolonies were created by rolling circle amplification (RCA) of a circularized DNA fragment (Figure 15). The critical approach was to construct a circularized single strand DNA. We have extensive experienced constructing rolonies and sequencing rolonies using sequencing-by-synthesis (SBS) and sequencing by ligation. Figure 15 shows rolonies randomly immobilized on a glass surface. The properties of Rolony are under developed.

3.2.1 Experiment Design

To prepare circularized library, genomic DNA was fragmented, blunt-end repaired, A-tailed and then adapters were ligated to two ends (figure 15). Size selection and a special primer set PCR were performed to generate 5' phosphorylated single strand DNA. The linear DNA was then circularized by CircLigase and rolling circle amplified by Phi29 DNA polymerase to obtain rolony. The rolony were attached on random array patterned surface and sequenced by ligation or synthesis (figure 16).

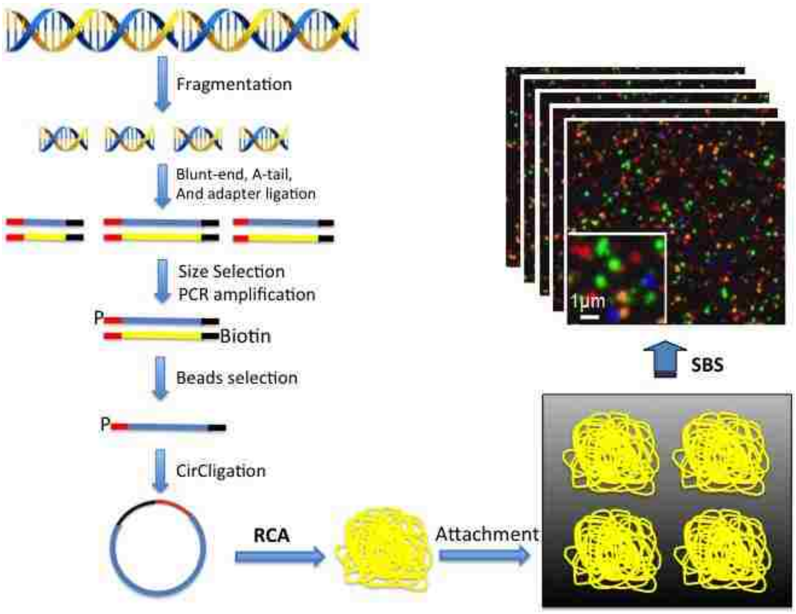


Figure 15. Overview of Rolony sequencing technology. Rolony preparation begins with circularization of 5'phosphorylated ssDNA template, and then follows up a rolling circle

amplification step to generate a compact rolony. The rolony can be arrayed on surface by attachment chemistry and sequenced by ligation and synthesis respectively.

3.2.2 Single strand DNA fragment library preparation

Genomic DNA was fragmented by NEBNext dsDNA fragmentase to ~100-500bp. Blunt-end repaired and A-tailed steps were performed after fragmentation and then adapters were added by T4 DNA ligase. 10 cycles of PCR were performed to amplify the fragments and the ~150bp PCR product was selected by running on a 6% PAGE gel and cleaned up by Phenol chloroform isoamylalcohol. For PCR primers, forward primer was 5’ phosphorylated modified and reverse primer was 5’ biotin modified to select single strand DNA fragments. Single strand DNA fragments were obtained by attaching the biotin related strand of double strand DNA fragments to Myone streptavidin beads, and then 0.15N of NaOH was used to elute the complementary strand with 5’ phosphorylation. Then 5’ P-ssDNA fragments were neutralized using Tris-HCl and sodium acetate buffer and cleaned up by Qiagen MINI elute PCR column. Qubit 2.0 was used to measure the DNA concentration.

3.2.3 Rolony preparation by rolling circle amplification

To perform rolling circle amplification (RCA), first, the 5’P-ssDNA fragments need to be circularized. To do this, 10 pmol of 5’P-ssDNA was mixed with 1x reaction buffer, 2.5 mM of MnCl2, 50 mM of ATP and 5 U/uL of CircLigase (Epicenter CL9021K) to a total of 50 µL reaction mixture. After incubating the reaction mix at 60°C for 2 h, the exceed CircLigase was inactivated at 80ºC for 10 min. The linear DNA was digested by incubating with 20U of Exonuclease I and 100U of Exonuclease III at 37ºC for 45 min. Extra Exonuclease I and Exonuclease III were heat inactivated at 80ºC for 15 min. Then circularized DNA template was purified using Qiagen Mini Elute Purification PCR kit, and Qubit 2.0 was used to measure the DNA concentration. The circularized DNA template was stored in -20ºC freezer and was good for at least 3 month.

To perform rolling circle amplification, 1 pmol of circularized DNA template was add the a RCA mixture which contains 1x Phi29 reaction buffer, 10 nM of RCA-primer, 10 µM of AA-dUTP, and 1 mM of dNTPs to reach a total volume of 99 µL. After incubating the RCA mixture with DNA template at 95ºC for 1 min, 60ºC for 1 min, 40ºC for 1 min, the mixture was paused on ice, then 10 U of Phi29 DNA polymerase (Enzymatics) was added, mixed well and incubated at 30 °C for 6 h.

After RCA, 400 uL of 1x PBS was added to the RCA tube and stored at 4ºC. Before loading Rolony on surface, additional 1: 10 dilution is recommended using 1x PBS.

3.2.4 Rolony attachment on surface

Amino-salinization treatment of binding surface. Amino-saline coding surface was freshly prepared as following. 1. Thoroughly wash with dH_2O, isopropanol, dH_2O and acetone respectively and air-dry the glass surface to be coated. 2. Coat surface with Amino-silane (Perform acetone steps in a fume hood). 1% of 3-Aminopropyltriethoxysilane in dH_2O was freshly prepared and immersed the glass surface for 5 min. Then rinsed the surface with dH_2O and Acetone respectively, and air-dry the surface before use.

Bind Rolony to amino glass slide. A flow cell was assembled using HybriWell^{TM} (GRACE Bio-Labs, Inc.). 100 µL of Rolony was loaded in the flowcell, and incubated at room temperature for 1 h. Then the flowcell was then washed twice using Wash 2 (5x SSC+0.01%Tween20) as follow by capillary action: first, tiled the flowcell at 45 degree; then added a drop liquid to the inlet of the chamber, the capillary action should displace the drop and accumulation of wash should be seen at the outlet of the chamber, avoid bubble formation. Repeated wash 3 times and the rolonies were ready for sequencing.

The patterned array flowcell was gift from Wyss Institute at Harvard Medical School. Basically, silicon dioxide was coated on silicon surface. Then the silicon dioxide was overlaid by a layer of titanium. Under a serial of photolithography, dry etching, vapor deposition, deep-UV, positive-tone photoresist coating, the silicon surface was able to be treated by aminosilane and finally remove the photoresist. 2x rolonies were loaded on the flowcell and incubated at room temperature for 1h. Followed by 3 times of Wash 2 to remove the un-binding rolonies, the flowcell was ready for sequencing by ligation and/or synthesis.

3.2.5 Rolony Sequencing by Ligation

First 5 cycles were done by conventional SBL. The hybridization of sequencing anchor primer was performed as follow: first, prepared and filled flow cell with the primer solution (1 µL of 1 mM Anchor primer-1 in 200 µL of Wash 2 buffer). The flowcell was incubated at 50ºC for 5min, then room temperature for 5 min. After hybridization, the flowcell was washed with Wash 2 for 3 times.

A ligation mix, containing 100 µL of 2x Quick ligation buffer, 1 µL of fluorescent nonamer mix, and 4 µL of T4 DNA ligase, was added to the flowcell and incubated at 37ºCfor 5 min, then room temperature for 5 min. Then 3 times of Wash 2 and 1E wash were performed before imaging the flowcell.

The image was taken under Polonator G.007 and fluorescent microscope. After imaging, the signal was stripped by adding 200 µL of 65% Formamide (in TE) into the chamber, incubate 5 min. Then the flowcell was washed with 3 times of 65% Formamide and 3 times of Wash 2. Then the flowcell is ready for second cycle sequencing.

To sequence 6-10 bases, Rolony-minus Anchor Primer-2 was hybridized to rolony (the process is the same as hybridizing Anchor primer-1). Then 200 µL of 7-mer-GA (30 uM) in 1x Quick ligase buffer with T4 DNA ligase was added to flowcell and incubated at 37ºC for 7 min, then at room temperature for 7 min, and then repeat 37ºC and RT for 7 min to ligate Rolony-minus Anchor Primer-2 and 7-mer-GA. 1E wash was performed after ligation. 200 µL of phosphorylation mixture (4 µL T4 polynucleotide kinase (PNK) in 196 µL 1x PNK buffer) was then added to flowcell, and the flowcell was incubated at 37ºCfor 30 min to perform

5'Phosphorylation. After phosphorylation, the nonamers were ligated to 7-mer-GA for imaging. When finished imaging, the anchor primer strand were stripped by Formamide before next round of sequencing.

3.2.6 Rolony Sequencing by Synthesis

The sequencing by synthesis was performed using the following four steps. Firstly, a fluorescent extend step incorporated the fluorescent reversible terminators to anchor primers, which was followed by saturation step to incorporate non-fluorescent reversible terminators to the un-extended primers. Wash to remove the excess reagents, four channel images was then taken using CCD camera. The terminators and fluorescence were then cleaved and 3'OH group was recovered to the sequencing strand. With the free 3'OH group, another sequencing cycle was ready to proceed.

Sequencing by synthesis reagents were provided by Intelligent Bio-System (acquired by Qiagen in 2012). Basically, working solutions were prepared from the sequencing kit provided:

Extend A: Mix full contents of A1, A2, and A3*;

Extend B: Mix full contents of B1, B2, and B3**;

Cleave: Mix full contents of Cleave 1 and Cleave 2***;

Cap: Fully dissolve solid Cap 2 into solution Cap 1****;

Wash Buffer: Dilute 10X stock to 1X with water.

All reagents were stored at 4°C

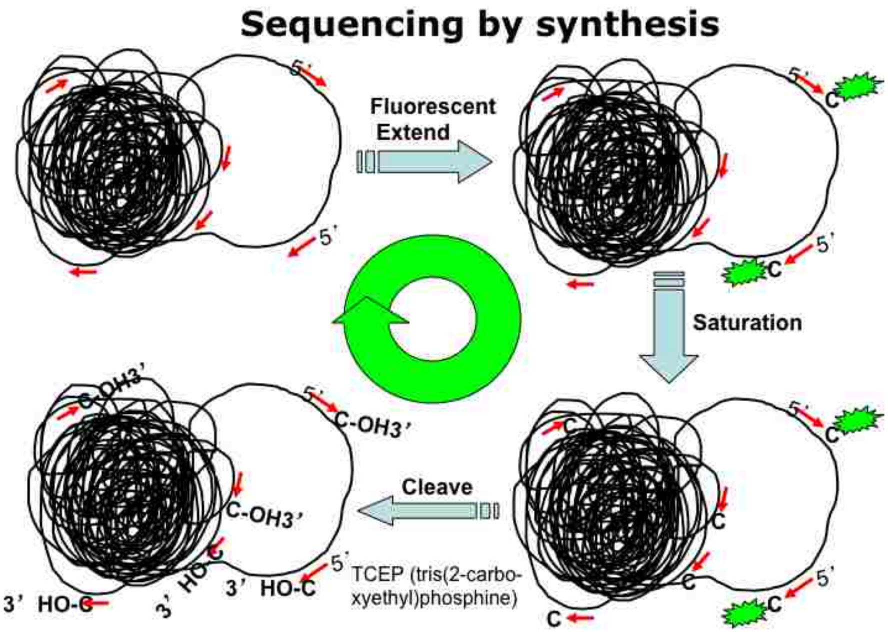


Figure 16. Process of Sequencing by synthesis on Rolony

* This solution was stable for up to 3 days at +4 °C. If smaller amount is needed, then mix 48 volumes of 'Extend A1' with 1 volume of 'Extend A2' and 1 volume of 'Extend A3'. ** This solution was stable for up to 3 days at +4 °C. If smaller amount is needed, then mix 48 volumes of 'Extend B1' with 1 volume of 'Extend

B2’ and 1 volume of ‘Extend B3’. *** This solution was stable for a week at room temperature. If smaller amount is needed, mix 19 volumes of ‘Cleave 1’ and 1 volume of ‘Cleave 2’. **** This solution was stable for a week at room temperature. If smaller amount is needed, then store solution at +4 °C.

Delivery Method: Inject 500 uL of solution, followed by 500 uL of Wash Buffer, using pump speed 30. Reaction Method: Inject Extend A into flowcell, incubate for 5 min at 55°C. Wash flowcell with 1 mL of Wash Buffer, gradually decreasing pump at speed 30. Inject Extend B into flowcell; incubate for 5 min at 55°C. The flowcell was washed with 10 mL of Wash Buffer, gradually decreasing pump speed from 22 to 30 (5x 1 mL at 22, 1 mL at 24, 26, 28, 29, 30). Image and then Inject Cleave into flowcell, incubate for 10 min at 55°C. Flowcell was washed with 5 mL of Wash Buffer, gradually decreasing pump speed from 22 to 30 (1 mL at 22, 24, 26, 28, 30). Cap was injected into flowcell, incubated for 5 min at 37°C. The flowcell was washed with 10 mL of Wash Buffer, gradually decreasing pump speed from 22 to 30 (5x 1 mL at 22, 1 mL at 24, 26, 28, 29, 30). And the repeat the above process to perform cycles sequencing.

3.2.7 Paired–end Rolony sequencing by synthesis

One of the major limitations of rolonies (and many other next gen sequencing approaches) was that there was no obvious way to obtain paired-end reads. Paired-end reads are very useful for genome assembly and read mapping. For example, structural polymorphisms are very difficult to identify without paired-end reads. Herein, we developed an innovative paired-end sequencing strategy for rolonies. Extensive preliminary data suggested that our paired-end sequencing strategy was feasible, and potentially the approach will attract significant attention from many sequencing groups.

To generate paired-end reads from rolony, the critical requirement for DNA polymerase was strand displacement. To test the ability of strand displacement, three DNA polymerases, Bst, Phi29, and IBS Modified N9^{o} were selected to extend a hairpin oligo as shown figure 22. Since the full length of the oligo was ~ 60bases and this ~ 60bp product band should only appear when the polymerase has the strand displacement ability. Otherwise, the product will stay around 40bp with the anchor primer.

Paired-end Rolony was prepared as the process illustrated as figure 17. Basically, the 5' phosphorylated single strand DNA library was prepared as regular single library (see details in section 3.2.2). The single DNA was amplified and double digested before ligating the hairpin adapters. With the dual hairpin adapters, the double strand linear DNA was constructed into a circularized template and this template was rolling circle amplified to generate paired-end rolony. The paired-end rolony was attached on surface and sequenced using two anchor primers from each of the hairpin.

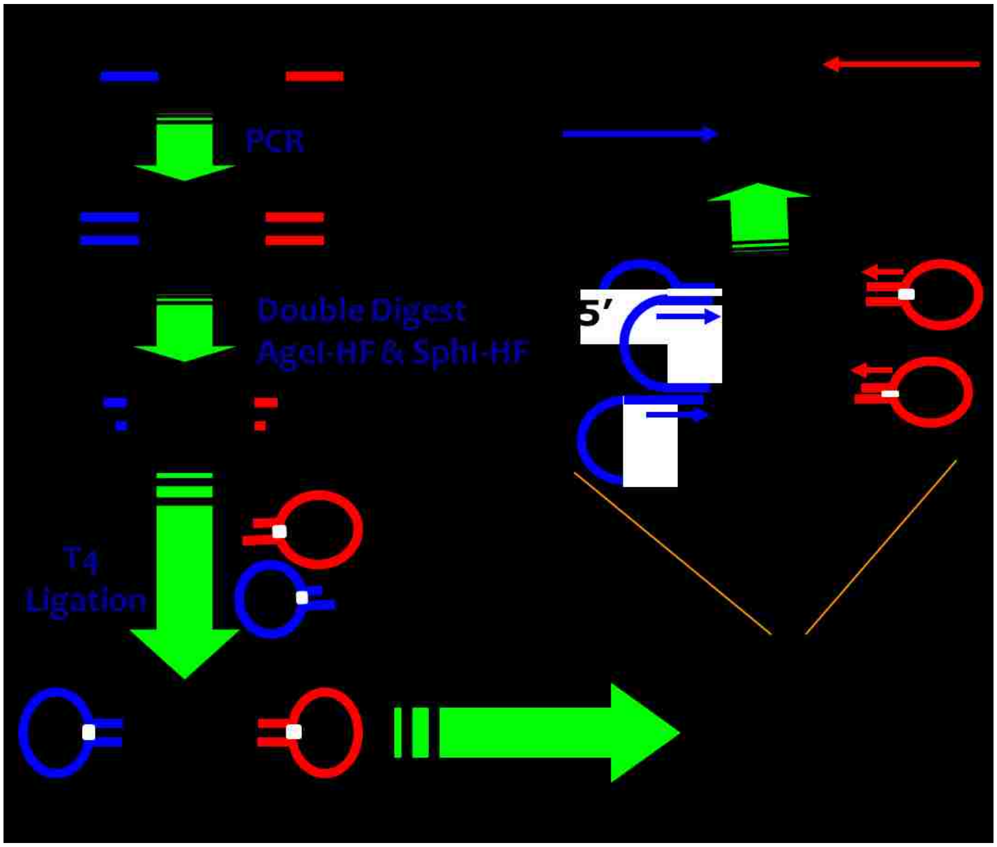


Figure 17. Process of Paired-End Rolony sequencing

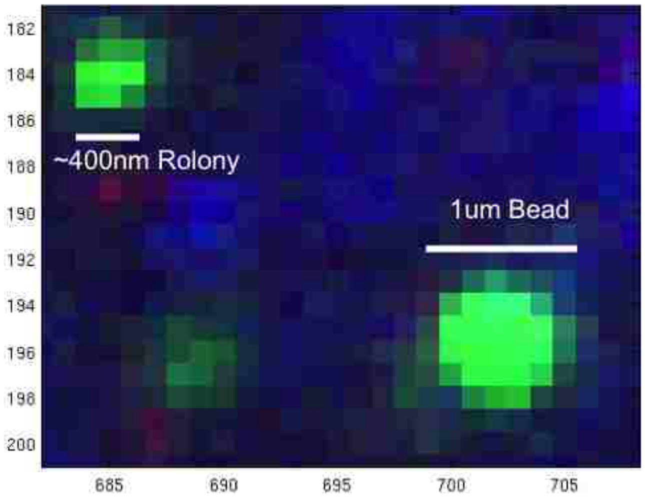


Figure 18. Size comparison of Rolony and bead

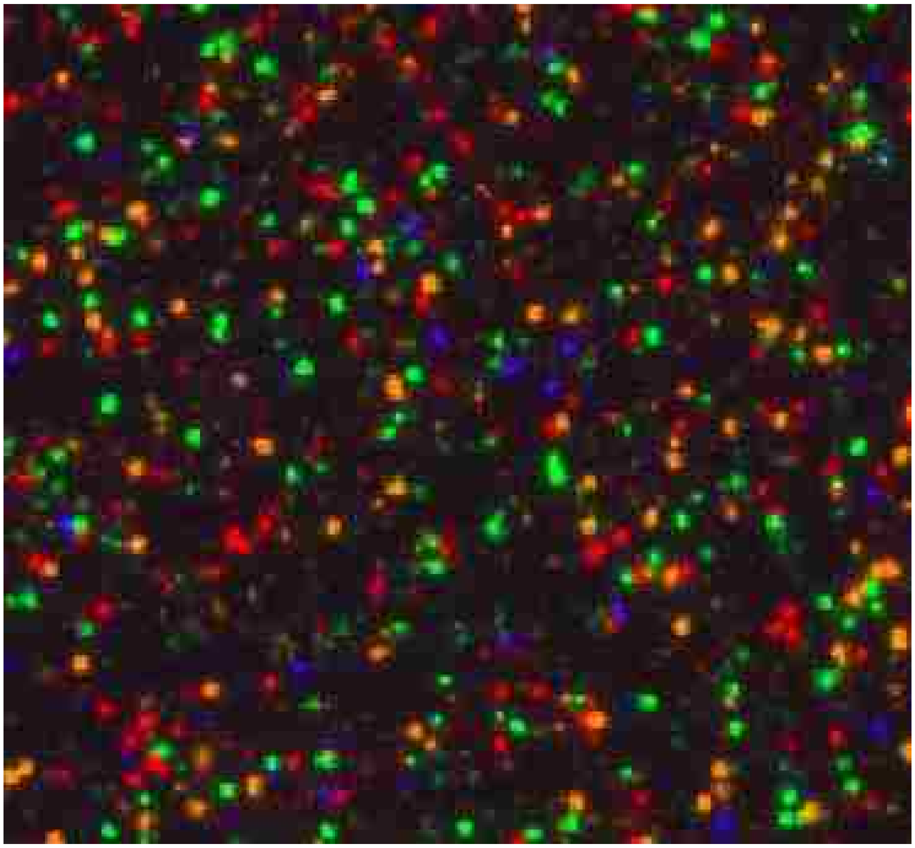


Figure 19. Rolony Sequencing by Ligation

3.3 Results

3.3.1 Fragment library and rolony characterises

5’Phosphorylated single strand DNA library was successfully prepared. Qubit ssDNA reagents were used to measure the concentration of single strand DNA. ~10 ng of DNA library was always obtained with average size ~150 bases. The linear DNA was circularized and ~3 ng of circularized DNA was always achieved after purification. The circularized DNA library was stable for at least 3 months under -20ºC. Rolony was successfully prepared per the protocol, and quality controlled by loading on slide and sequence one base, or using fluorescent probe hybridization to visualize rolony. Rolony was stable for at least 6 month when suspended in PBS buffer and stored at 4ºC. Rolony was ~400 nm diameter, and the size was smaller than Myone beads which was 1μm diameter as shown in figure 18.

3.3.2 Rolony Sequencing by Ligation

Rolony was successfully sequenced by ligation. Both 5’ to 3’ direction and 3’ to 5’ direction were preceded, and for the 3’ to 5’ direction, bases 1-5 were sequenced by regular SBL, while bases 6-12 were sequenced using cPAL method. The following image showed an example of Rolony sequencing by ligation. The image was taken under 40x +1.5x setting HC_Plan_APO (4 colors shown, FAM-blue, Cy3-green, Cy5-red, TxRed-yellow).

3.3.3 Rolony Sequencing by Synthesis on patterned array flowcell

Rolony was successfully sequenced using SBS reagents from Intelligent Bio-System. 25bpwas performed, and longer read length can be achieved potentially. With patterned flowcell, near 100% of the surface was attached rolony, which increased the throughput up to 100-fold as the regular polony sequencing. The following image was one frame of the 320 μm x 320 μm region surface which captured hundreds thousands of rolonies (as shown figure 20). The SBS single was very good with high single-noise ratio as shown in figure20, 21, which was very helpful for base calling.

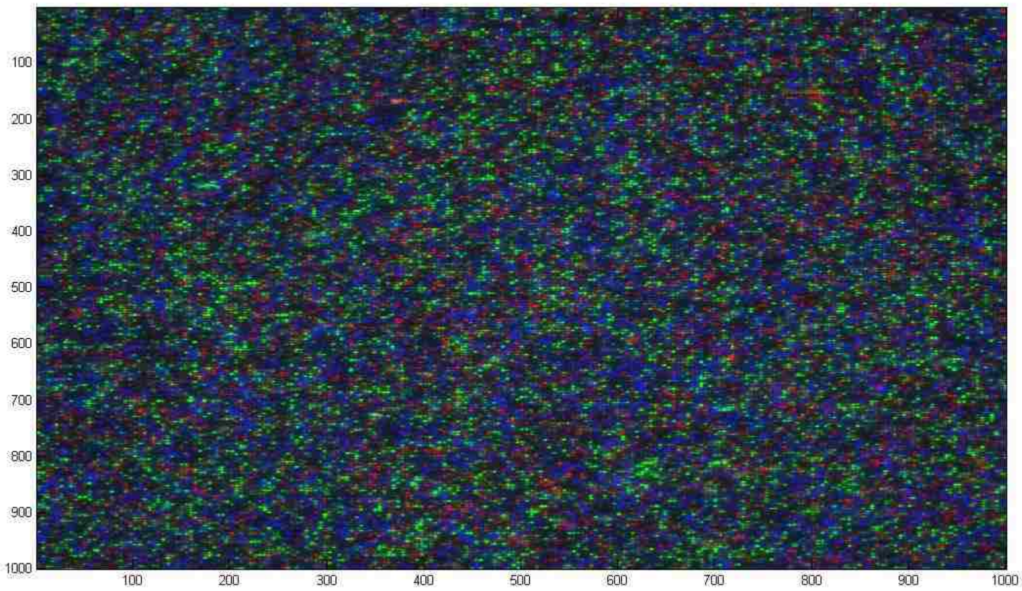


Figure 20. Rolony sequencing by synthesis on patterned array surface.

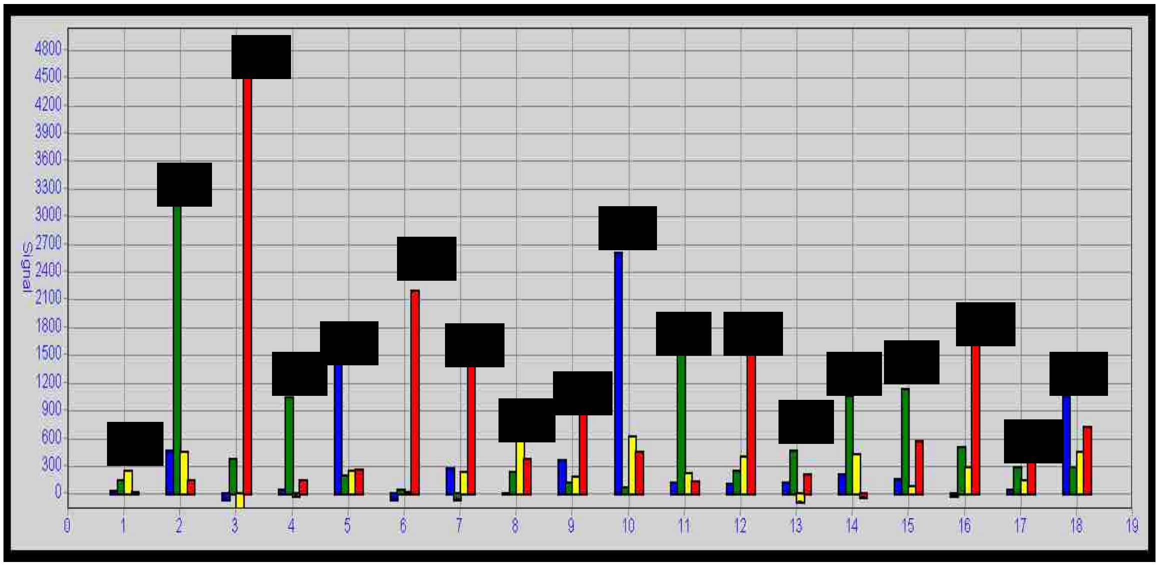


Figure 21. Base calling for Rolony Sequencing by synthesis.

3.3.4 Paired-end Rolony sequencing by synthesis

Polymerase strand displacement ability was tested by extending hairpin oligo. As shown figure 22, IBS enzyme and Bst were very good for strand displacement, while Phi29 was able to do strand displacement with low efficiency.

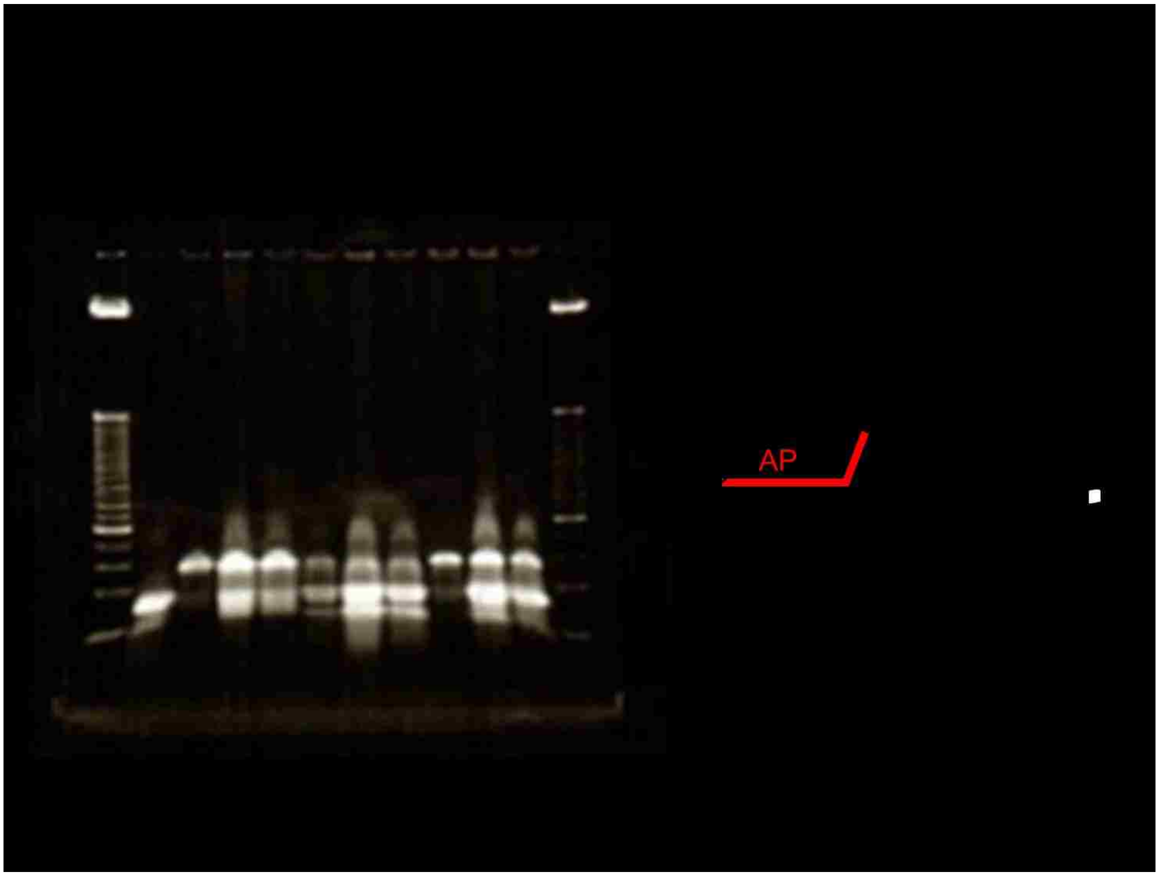


Figure 22. Strand displacement test for polymerase Bst, Phi29 and IBS enzyme.

Paired-end rolony was successfully sequenced from two hairpin end. 2 bases of each end were performed, and longer read length was potentially feasible. The paired-end reads was very helpful for assessment genome structure recombination, genome assembly and reads mapping.

3.4 Discussion

Fragment library was successfully constructed for preparing Rolony. Rolony is only about 400nm diameter, which is smaller than any bead-based sequencing feature. Rolony was successfully prepared and sequenced using SBL and SBS chemistry. Remarkably, 25bp of SBS was achieved, and longer reads could be obtained potentially in the future. Meanwhile, paired-end rolony strategy was developed to get paired-end reads from rolony, which was difficult for most of the sequencing technologies. Longer reads could be achieved from Rolony SBS for better genome assembly, structure rearrangement and reads mapping.

3.5 Summary

In this chapter, a smaller DNA sequencing feature—rolony, was described. A straightforward library construction was developed. Additionally, rolonies were successfully prepared from circularized DNA library and sequenced using SBL and SBS chemistry on random array and patterned array flowcell. To obtain paired-end reads, a pared-end rolony preparation technology was developed, and paired-end rolony was successfully sequenced from two hairpin ends. The small feature and super high throughput of Rolony sequencing will potentially drive sequencing technology to a novel level and contribute to biomedical sciences and clinical applications.

CHAPTER 4

HAPLOTYPE-RESOLVED SEQUENCING

4. 1 Introduction

A haplotype is the combination of genes at adjacent locations (loci) on a chromosome that are inherited together. Therefore, haplotype phase information is extremely important for truly understanding human genomics[237,238]. Specifically, haplotypes are important for identifying the association between genetic variations and disease[70,239,240]. Widespread allele-specific expression (ASE)[241,242] and methylation (ASM) have shown that the specific DNA sequence and epigenetic context can influence the function of diploid state[243]. ASE is evident among autosomal genes and has been shown to be associated with many diseases[244] such as cancer[90,245] and neurodevelopmental disorders[246]. Therefore, acquiring haplotype information is critical to understanding disease associations[247,248].

With the completion of Human genome project[3,249] and the development of next generation sequencing technologies[9,10,15-17,23,49,217,224,238,250] we have improved our understanding of human genome as several human genomes have been sequenced[3,23,65-67,71,219,251] to date. However, the vast majority of them are diploid type and with heterozygous SNPs that cannot be assigned to specific allele. Although HapMap project[3,247,252,253] was launched to identify and catalog genetic similarities and differences between humans it is based on statistic and computational methods. Additionally, the 1000 Genomes Project was carried out to identify most genetic variants that have frequencies of at least 1% in the populations studied[70]. The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, which provides a comprehensive resource on human genetic variation. However, these population-based genotype data contain high error rates and therefore have difficulty phasing rare variants[254]. Both of these projects have provided valuable data to understand human genetics, however, these projects do not include haplotype phasing[254].

Recently, in order to obtain haplotype information from the sequencing of human genome a fosmid cloning method[71] and two dilution-amplification methods[255,256] were developed to address this issue using the next generation sequencing technologies[17,238]. The critical technique was whole genome amplification (WGA) using minimum amount of DNA, such as a couple of cells, or even less as single cell. There were a couple of commercial WGA kits, but none of them was able to amplify down to 1/10 of single genome, which is 0.3pg for haplotype genome. The three papers addressed different ways to amplify minimum amount of DNA. The fosmid cloning method separated DNA fragments into fosmid, then cloned DNA fragments and barcoded the fosmid DNA to be sequencing libraries. The other papers were dilution-based methods, which diluted DNA to desired concentration to acquire only a partial of genome per WGA reaction. Then multiple-displacement amplification method was utilized to amplify the certain amount of DNA for each WGA reaction, and then barcoded libraries were constructed using WGA DNA, and sequenced and analyzed on their private system.

Here we described an accurate and rapid method that can be used to directly phase haplotype. Basically, genomic DNA was diluted and aliquoted to hundreds of whole genome amplification (WGA) reactions. Each WGA reaction only contained a fraction of the human genome. The WGA was performed by multiple displacement amplification (MDA) using Phi29 polymerase to generate ~300ng DNA for libraries preparation. The WGA products were used to construct barcode libraries using Illumina Nextera technology[238] and Ion Plus library preparation system[23]. The barcoded libraries were pooled together and sequenced on HiSeq 2000 and Ion Torrent/Proton systems respectively. Whole genome re-sequencing (WGRS) data was obtained with the same genomic DNA, and heterozygous SNPs from WGRS data was used as phasing reference. The het SNPs from WGA data and WGRS data were combined to generated het-het linkage graph to phase haplotype. Exome sample was sequenced and used to validate the phase results. Finally, the phase results were compared with HapMap data.

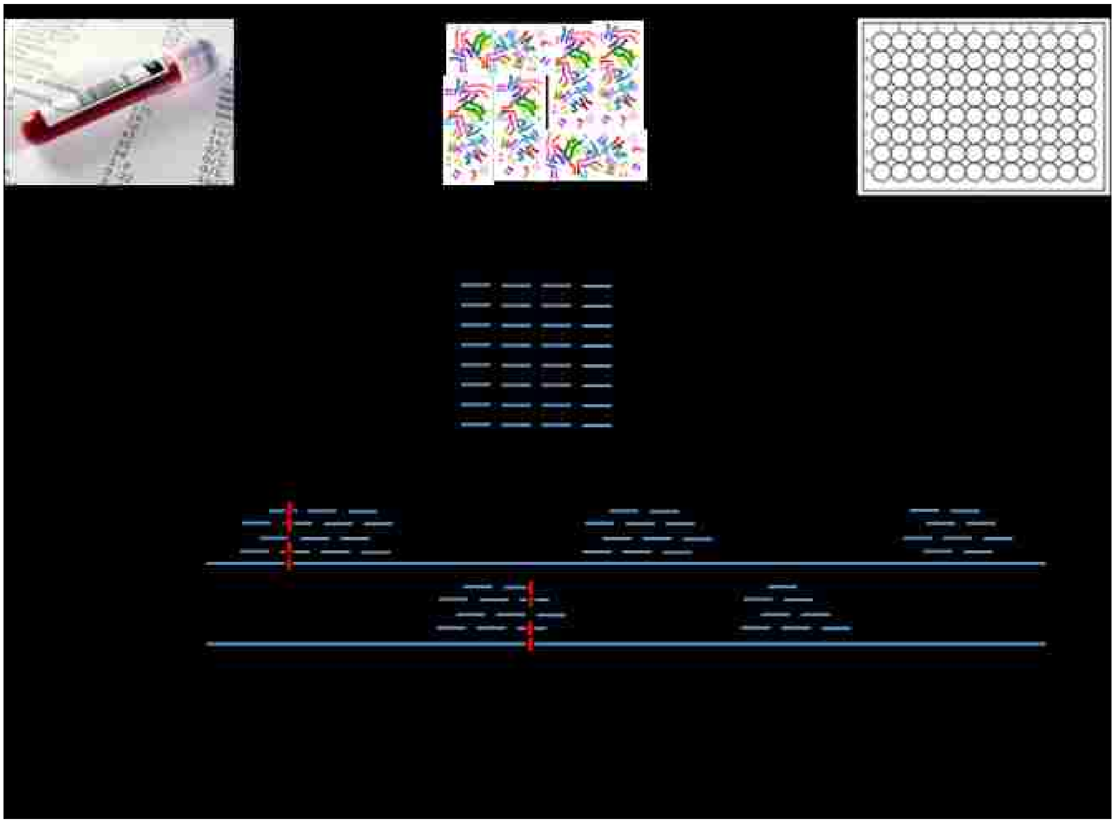


Figure 23. The principle of haplotype phasing. The pairs of chromosomes contain one maternal chromosome and one paternal chromosome. The purpose for Haplotyping is to separate the information from mixed chromosomes, and assign the alleles to specific parent. Due to

heterozygous SNPs distribute ~1.3kb between each; the heterozygous SNPs can be used to phasing haplotype. Basically, 5% of diploid genome was amplified and barcoded, and then pooled together and sequenced. The reads can be mapped to reference to assemble haplotype chromosomes.

4. 2 Methods and Experiments

There are some difficulties that arise when gathering haplotype information. First, the ability to separate haploid alleles; second amplifying the minimum amount of haplotype genomic DNA[257,258]. To overcome these difficulties, the genomic DNA was diluted and aliquoted into hundreds of whole genome amplification reactions. Therefore, each WGA reaction only contained a fraction of one haploid fragment. As shown in figure 24, if it happened to have two diploid alleles in one WGA reaction, then it will appear as heterozygous allele. Each aliquot was amplified by MDA to generate enough DNA material for library preparation. Each WGA product was then sequenced and used to phasing haplotype.

4.2.1 Whole genome Amplification input DNA test

All experiments were conducted in a Cleanspot A PCR/UV work Station with a dedicated set of pipettes. To eliminate DNA contamination all reagents, except for the primers, dNTPs, and polymerase, were treated with UV for 20min.

Biological grade mineral oil (Bio-Rad) was used to prevent evaporation, and the tubes were centrifuged briefly and checked visually to make sure the aqueous phase formed a small sphere at the bottom of the PCR tube. The PCR tube was incubated at 30 °C for 18 h, and the enzyme was deactivated at 65 °C for 10 min, and then held at 4 °C. After PCR, 50 µL of H_2O was added to each tube, mixed well and briefly centrifuged. The mixture was pipetted onto a Parafilm, where the oil spread out while the sample was remained as a droplet. The samples were collected with a PR-20 pipette and placed in new 0.2-mL PCR tubes. The DNA concentration of each WGA product was measured on Qubit 2.0.

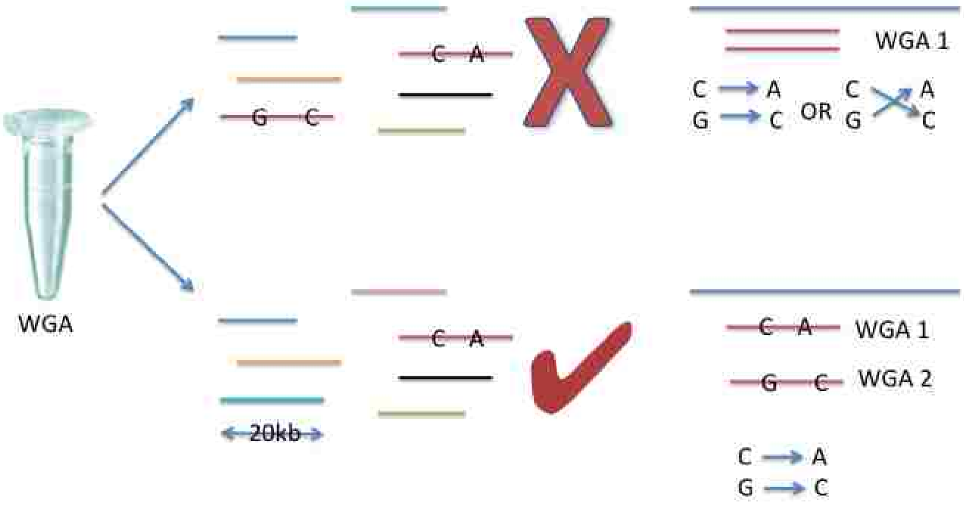


Figure 24 The principle of dilution based Haplotyping method. The goal was to amplify ~20kb fragments which only covered partial of human genome, to limit the fragments from same region of both parents to present in single WGA reaction.

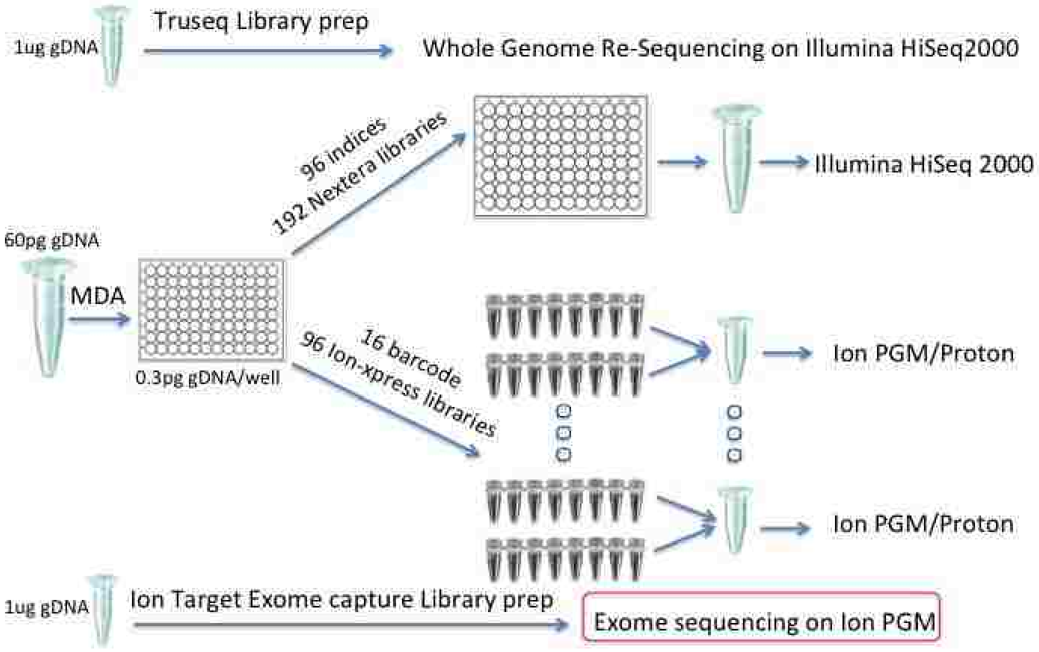


Figure 25 Overview of experiments design. A WGRS data was designed to call all the homozygous and heterozygous SNPs. Then ~300 WGA products were barcode constructed to sequencing libraries, and sequenced on HiSeq and Ion Torrent/Proton simultaneously. The WGRS set of data was used to call all the heterozygous SNPs, and phasing heterozygous SNPs using WGA data. Finally, Exome sequencing data was used to validate the heterozygous and all SNPs from WGRS data.

To perform WGA, two master mixes were made. For Master Mix I, 8 µL of 10x Phi29 Buffer, 20 µL of 1 mM dNTPs, 8 µL of 0.1 mM dUTP, 3 µL of H_2O, and 1 µL of Phi29 Polymerase (10 U/µL, Enzymatics) were freshly mixed together, and kept in ice. For Master Mix II, the desired concentration of DNA was mixed with 40 µM of octamer. 1 µL of Master Mix I and 1 µL of Master Mix II were added to a 0.2 mL PCR tube, without mixing. Aqueous was overlaid with 20 µL of bio-technology grade mineral oil (Bio-Rad) to prevent evaporation, then centrifuged briefly and checked visually to make sure the aqueous phase formed a small sphere at the bottom of the PCR tube. The PCR tube was Incubated at 30 cent °C for 18 hours, and deactivated at 65 °C for 10min, then held at 4 °C until analysis. Then the tube was removed from thermo cycler, and 50 uL of H2O was added to each tube, mixed well and briefly centrifuged. The mixture sample was transferred on a square of Parafilm, and the oil was spread out while the sample was remained as a droplet. Collected samples with a PR-20 pipette to new 0.2mL PCR tubes. 1uL of each WGA product was used to measure DNA concentration on Qubit 2.0.

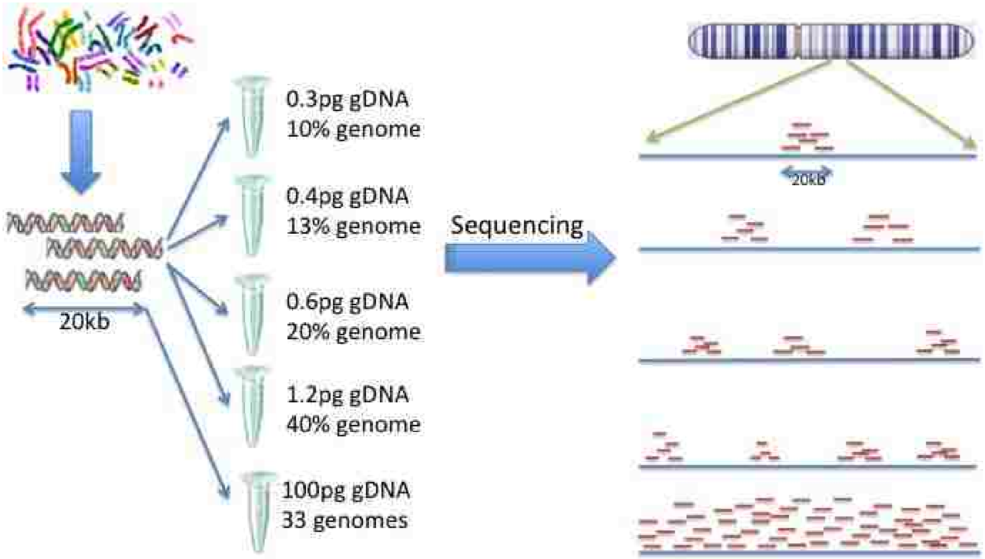


Figure 26 Whole genome Amplification input DNA test. A series of DNA concentration were tested. As increasing genomic DNA for WGA, the gaps between the 20kb fragments was getting smaller.

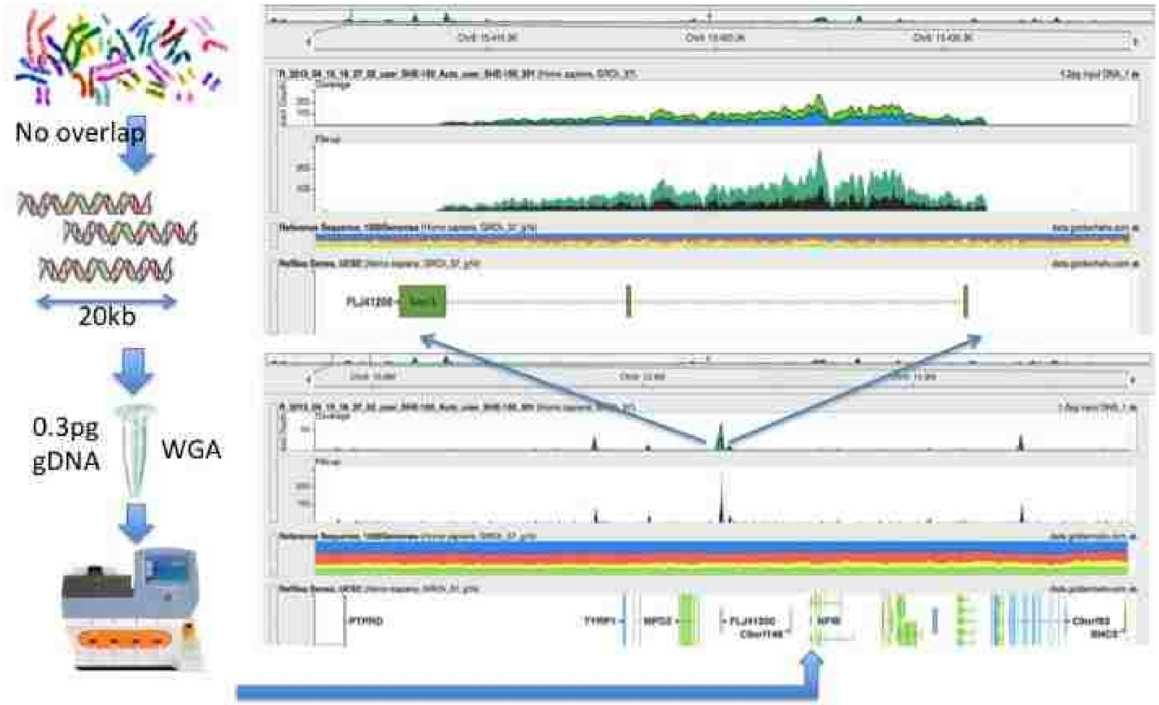


Figure 27 Amplify and sequence 20kb long fragments. Genomic DNA was extracted from a blood sample using Qiagen kit as per manufacture instructions. The average size of DNA fragments post extraction was ~20kb. A series of input DNA concentrations was used to perform WGA reaction, and then the WGA products were sequenced on Ion Torrent using Ion-318 chips. ~99% of the reads were mapped back to human genome reference (GRch37). Reads piled up to cover ~ 20kb region, which then spaced between next 20kb region. The space between “uncovered” regions decreased as the input DNA increased.

4.2.2 Whole genome Re-Sequencing

Whole genome re-sequencing was performed on HiSeq 2000 at Beijing Genomic Institute. The library was constructed as per Illumina TruSeq HT library preparation protocol with slight modification. Briefly, 1ug of genome DNA was sheared; blunt-end repaired, and then purified using Agencourt AMPure XP II reagents. The End repaired DNA was then mixed with A-tailing mix to add adenylate 3’ ends at 37°C for 30min. Ligation Mix and Adapter index 1 (AD001, ATCACG) were added directly to A-tailed DNA reaction and incubated at 30°C for 10 min then 5 µL of Stop Ligation Mix was added to quench ligation. The DNA was purified with Agencourt AMPure XP II reagent. The purified DNA was size-selected on 2% SizeSelect E-Gel system and ~800bp product was collected. Then 7 cycles of PCR was performed to amplify DNA product. The PCR mixture was purified using Agencout AMPure XP II reagent and then 1 uL of DNA library was used to measure the size and concentration on Agilent Technologies 2100 HS Bioanalyzer system.

4.2.3 Nextera DNA library preparation

WGA reaction using 0.3 pg genomic DNA was done to generate ~300 ng DNA for library preparation. Using 50 ng from each whole genome amplified DNA was purified and used to construct Nextera DNA libraries. WGA DNA was diluted to 20 µL of 2.5 ng/uL using molecular biological grade water and then 25 µL of Tagmentation (TD) buffer and 5 µL of Tagmentation Enzyme (TDE1) were added to each DNA sample. The mixture was pipetted up and down 10 times to mix well and then quick spun at 280 x g for 1 min at room temperature. Then the mixture was incubated at 55°C for 5 min and then held at 10°C. The mixtures were cleaned up using Qiagen Mini Elute PCR column as per manufacture instructions. To measure the size distribution of DNA fragments, 1 µL of undiluted DNA was loaded on an Agilent Technologies 2100 HS Bioanalyzer chip. All the DNA fragments distributed a broad size range from ~150bp to 1kb.

The second step for this library preparation was employed PCR to insert index 1 (i7) and index 2 (i5), and P5 and P7 for cluster generation and sequencing. Using 5 µL of index 2 primers, 5 µL of index 1 primers, 15 µL of NPM (Nextera PCR Master Mix), 5 µL of PPC (PCR Primer Cocktail), and 20 uL of purified tagmented DNA which were mixed and pipetted up and down 3-5 times, then quick spin at 280xg at room temperature for 1 min. PCR was performed using the following parameters on a thermal cycler: 72 °C for 3 min; 98°C for 30 sec; 5 cycles of 98°C for 10 sec, 63°C for 30 sec, and 72°C for 3 min; and then hold at 10°C. After PCR, clean-up was performed to purify DNA for sequencing. First, 50 µL of post-PCR mix was transferred to a 1.5 mL non-stick tube, then 30 µL of AMPure XP beads (for 2 x 250 runs on the MiSeq, add 25 µL of AMPure XP beads) were added, pipetted up and down 10 times. The tube was incubated at room temperature without shaking for 5 min and then placed on a Magnetic Particle Collector (MPC) for 2 min. The supernatant was removed and discarded carefully without touching the beads. The beads were then washed with freshly prepared 80% ethanol as follows: 200 µL of fresh 80% ethanol was added to the tube and incubated on for 30 sec; then carefully remove and discard the supernatant. With the tube on MPC, perform a second ethanol wash. Residual ethanol was removed with a pipette and the beads were air-dried for 15min. The tube was removed from MPC, and 32.5 µL of Re-suspension Buffer (RSB) was added. The mixture was gently pipetted up and down 10 times and incubated at room temperature for 2 min. After placing the tube on MPC for 2min, 30 µL of the supernatant was transferred from the tube to library collecting tubes.

The libraries were validated using Qubit 2.0 and Agilent Technologies 2100 Bioanalyzer on a High Sensitivity DNA chip. Based on an average size of 500bp, 1ng/µL =3nM.

All sample libraries were normalized to 2 nM final from 10 µL of original concentration using Tris-HCl 10 mM, pH 8.5 with 0.1% tween 20, and then 5 µL of each sample was transferred to a 1.5 mL non-stick tube, mixed well and stored at -20°C. Two sets of 96 barcode libraries each were PE100 sequenced on 6 lanes at HiSeq 2000 at Beijing Genomic Institute.

4.2.4 Ion-xpress DNA library preparation

Using 0.3 pg genomic DNA ~300 ng of DNA was generated from every WGA reaction. For DNA for library preparation, 100 ng of WGA DNA was sheared, blunt-end repaired and A-tailed using Ion-Xpress Plus Shear kit. Briefly, 100 ng of each WGA amplified DNA was mixed with 5 μL of Ion Shear Plus 10X Reaction Buffer, water to reach a total volume of 40 μL. Then 10 μL of Ion Shear Plus Enzyme Mix was added the mixture was pipetted up and down 10 times and the enzyme mix was immediately added to DNA and buffer. The mixture was incubated at 37 °C for 15 min to prepare 200-300bp fragments. Then 5 μL of Ion Shear Stop Buffer was added immediately after incubation and mixed thoroughly by vortexing for at least 5 sec. The reaction tube was then placed on ice.

The fragmented DNA was purified using Agencourt AMPure XP reagents according to the manufacture's protocol. Briefly, 1.8x sample volume of Agencout AMPure XP Reagent was added to the sheared DNA sample, and then pipetted up and down 5 times to thoroughly mix the bead suspension with the DNA, then pulse-spin and incubated at room temperature for 5 min. After incubation the tube was pulse-spun and placed on the MPC for 3. The supernatant was then removed without disturbing the bead pellet. To wash the beads, 500 µL of freshly prepared 70% ethanol was added and incubated for 30 sec with the tube on MPC. The tube was turned around twice MPC to move the beads around. When the solution appeared clear, the supernatant was carefully removed and discarded without disturbing the bead pellet. The ethanol wash was done one more time. To remove residual ethanol, the tube pulse-spun and then placed back on the MPC and any residual supernatant was removed without disturbing the bead pellet. The beads were air dried on the MPC at room temperature for 5 min. Then the tube was removed from MPC and 25 µL of Low TE was added and mixed well to elute DNA. The tube was placed back in MPC for 1 min and then the supernatant containing the eluted DNA was transferred to a new 0.2-mL PCR tube without disturbing the pellet.

The sheared DNA was ready for adapter ligation. For non-barcode library, ~25 µL of sheared DNA was mixed with 10 µL of 10X Ligase Buffer, 2 µL of Adapters, 2 µL of dNTPs Mix, 2 µL of DNA Ligase, 8 µL of Nick Repair Polymerase, 51 µL of H2O and incubated at 25°C for 15 min; 72°C for 5 min, and then hold at 4°C. For barcode library, ~25 µL of sheared DNA was mixed with 10 µL of 10X Ligase

Buffer, 2 µL of P1 Adapter, 2 µL of Ion Xpress Barcode, 2 µL of dNTP Mix, 2 µL of DNA Ligase, 8 µL of Nick Repair Polymerase, 51 µL of H2O and incubated at 25°C for 15 min; 72°C for 5 min, and then hold at 4°C. Post ligation, the mixture was purified using Agencourt AMPure XP reagent according to manufacturer's protocol.

The library was size-selected with the 2% E-Gel SizeSelect Agarose Gel on E-Gel IBase$^{\text{TM}}$ unit and E-Gel Safe Imager$^{\text{TM}}$ tranilluminator combo kit according to manufacturer's instruction. Briefly, 20 µL of the DNA sample was loaded in the sample loading well and 50bp DNA ladder was loaded in middle ladder well to direct size selection. The 330bp DNA product was collected from the DNA collection well for 200 base-reads.

Size-selected DNA was amplified by mixing with 5 µL of Library Amplification Primer Mix, and 100 µL of Platinum PCR Super Mix High Fidelity to a total volume of 130 µL. The PCR was carried out as follows: 95°C for 5 min; then 5-8 cycles of denaturation at 95°C for 15 sec, 58°C for 15 sec, and 70°C for 1 min; and hold at 4°C. After PCR, the DNA sample was purified the same as above using Agencourt AMPure XP reagent.

The concentration of DNA library was determined using Qubit 2.0. All the libraries were diluted to 5 pg/μL to perform sequencing template generating. Every 16-barcode libraries were pooled together and sequenced on Ion Torrent using Ion-318 chips or Ion Proton using P1 chip.

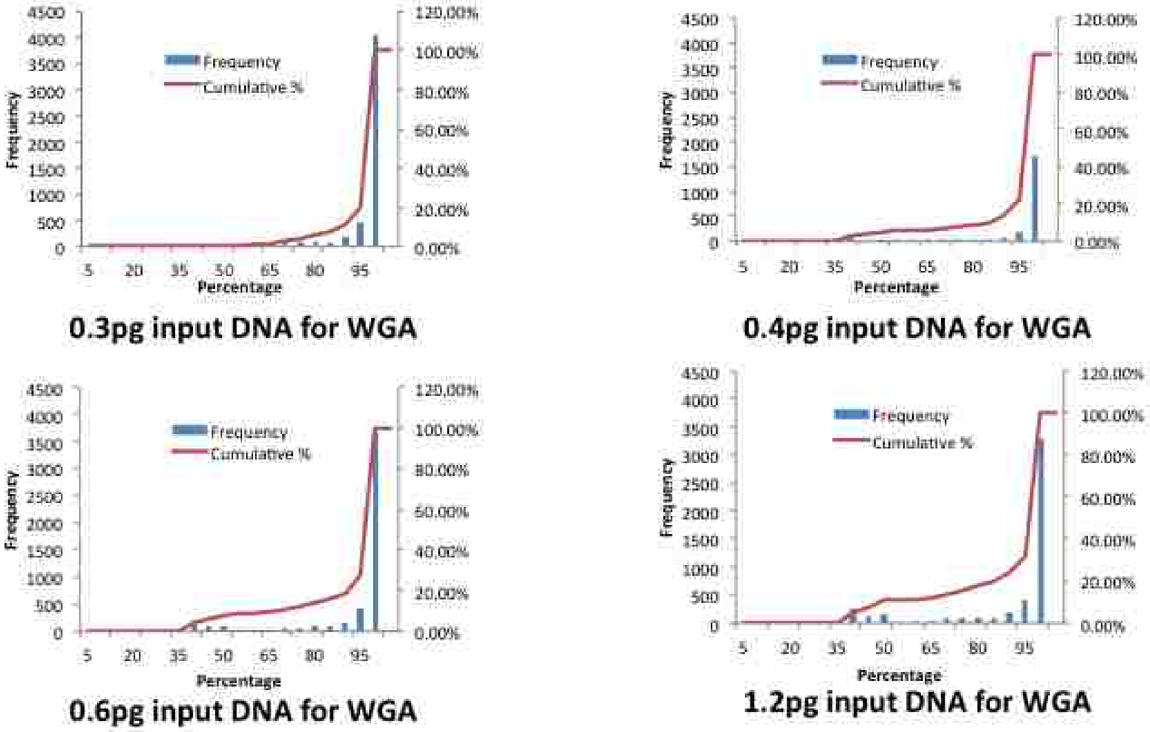
4.2.5 Ion TargetSeq$^{\text{TM}}$ Exome Enrichment DNA library preparation

The Exome library was constructed according to the protocol of Ion TargetSeq$^{\text{TM}}$ Exome Enrichment for the Ion Proton$^{\text{TM}}$ System with slightly modification. Briefly, 1 μg of genomic DNA was sheared by Ion Shear Plus kit and purified using Agencourt AMPure XP reagent. Then adapters were ligated to fragments and followed by size-selection on E-Gel system to collect ~220bp product. The 220bp DNA was amplified and quantitated and qualified using Qubit 2.0 and Agilent Bioanalyzer system. Then 500 ng of purified DNA was used to perform Exome enrichment at 55°C for 72 h. Then Exome library was purified and amplified to prepare exome-enriched library for template preparation and sequencing on Ion Torrent system.

4. 3 Results

4.3.1 WGA input DNA test

A series of input DNA were used to perform WGA, and then sequenced on Ion Torrent. As shown in figure 28, With increasing the input DNA amount from 0.3 pg/WGA to 1.2 pg/WGA, the heterozygous SNPs increased from <1% to ~11%, which was indicated by 50% scale as heterozygous SNPs.



0.3pg input DNA for WGA

0.4pg input DNA for WGA

0.6pg input DNA for WGA

1.2pg input DNA for WGA

Figure 28 Heterozygous SNPs increased as input DNA increased for WGA. A series of input DNA were used to perform WGA reaction, and then sequenced the WGA products on Ion Torrent system, and called all SNPs. With increasing the input DNA amount from 0.3 pg/WGA to 1.2

pg/WGA, the heterozygous SNPs increased from <1% to ~11%, which was indicated by 50% scale as heterozygous SNPs.

4.3.2 Whole Genome Re-Sequencing

~124 G WGRS data was achieved from 3 lanes sequencing on HiSeq 2000, and mapped back to human genome reference GRch37 using bowtie 2. The mapped reads were then taken to call SNPs. Total 3,844,198 SNPs were found with 2,388,901 heterozygous SNPs, spanning average 1.3kb per het. 2,388,901 het SNPs were used to generate a Het table. Meanwhile, 2,873,359 SNPs and 1,531,015 heterozygous SNP were identified from dbSNP and used to generate rsHet table.

4.3.3 Whole genome amplification libraries

~200 G data was generated from Ion Torrent system and HiSeq 2000, and mapped back to human genome reference GRch37 using bowtie 2 to generate a BAM file. All the mapped reads were then taken search for long reads fragment using Samtools Targetcut to generate a second BAM file, and aligned to genome reference. 573,282 of long fragment reads were generated, and 100% mapped to reference genome, with total 7,584,238,399 mapped bases. The mean size of long fragment was 13.23kb with maximum length of 127kb and length deviation of 5.8kb. ～2.5x coverage was obtained and used to call SNPs to reference dsSNP37. The average space for the het SNPs was ~1.3kb. All the heterozygous SNPs were used to phase haplotype chromosomes. The base ratios were: A 24.3%, C 16.9%, G: 16.9%, T 24.3% and N 17.7%. Ns were the gaps between the reads. A contig table was generated by going through each line of the BAM file to read 20kb fragments and find het from Het table (from the WGRS data), and a rsContig table was generated similarly by going through each line of the BAM file to read 20kb fragments and find het from rsHet talbe (from the WGRS data). Het table and contig table, rsHet table and rsContig table were then combined respectively in MatLab to assemble contigs, and finally assembled contigs were generated with an average length of 36kb (Figure 29, 30, 31). ~85% of heterozygous were phased, giving 50% of assembled genome was in block of length 48.4kb or greater.

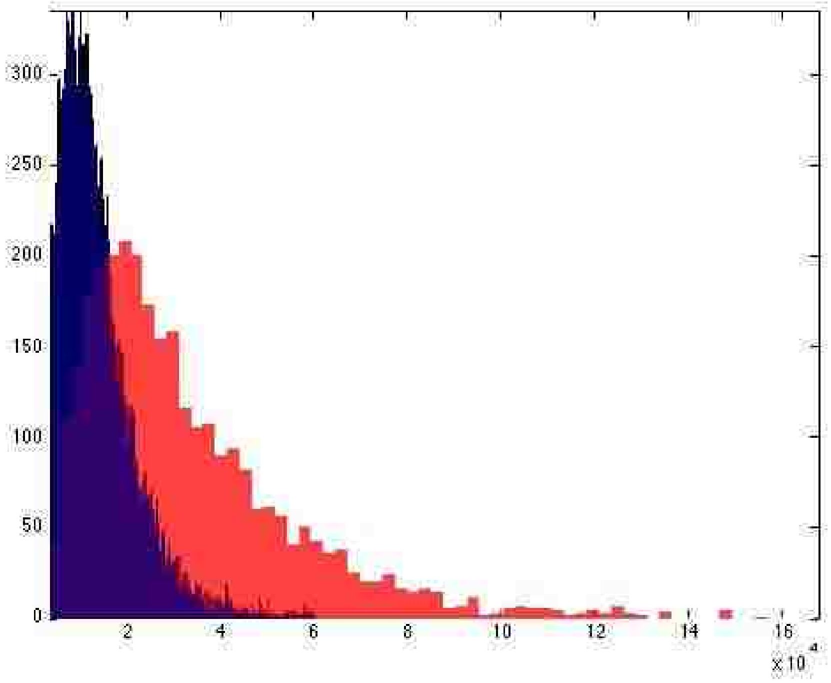


Figure 29 Histogram of the long reads (blue) and the assemble contigs (red) for chr1.The average assembled contig length for chr 1 was 33kb.

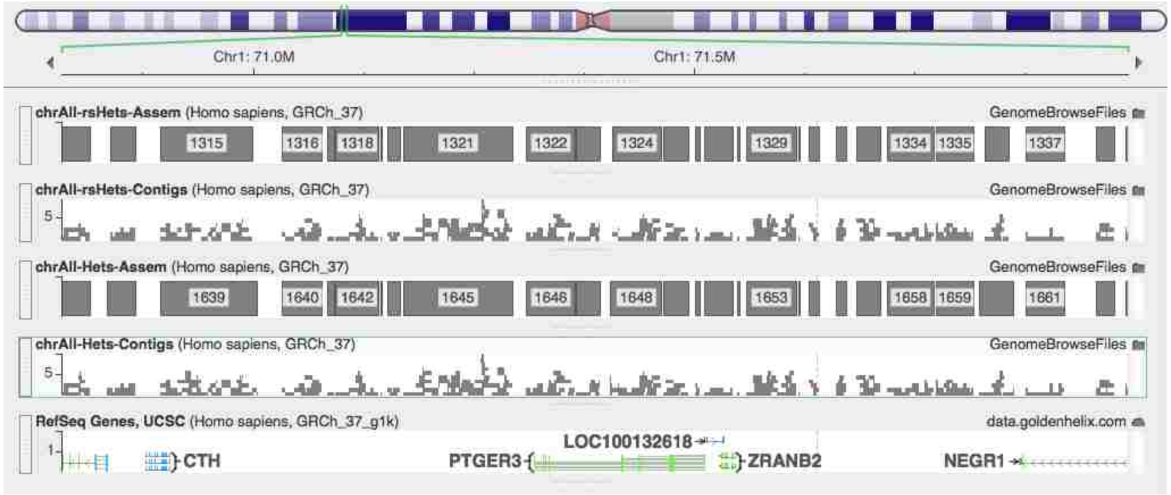


Figure 30 haplotypes spanning of chr1. The first row was assembled contigs generated by rsHet table and rsContig table. The second row was all the rsHets contigs. The third and fouth rows were assembled contigs and het contigs generated by Het table and Contig table.

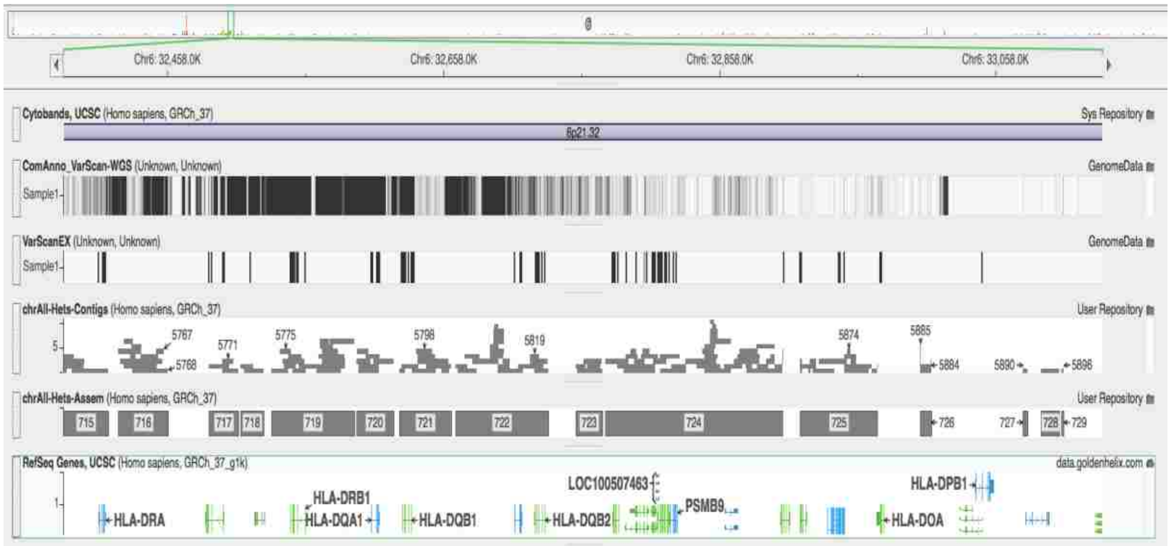


Figure 31 haplotypes spanning the HLA region of chr6. The HLA gene family provides instructions for making a group of related proteins known as the human leukocyte antigen (HLA) complex. The HLA complex helps the immune system distinguish the body's own proteins from proteins made by foreign invaders such as viruses and bacteria[259-279].

4.3.4 Validation by exome sequencing

4.9G exome data was generated and passed Q30 from Ion Torrent, and mapped to reference GRch37. The base ratios were: A 25.7%, C 24.4%, G 24.4% and T 25.5%. The exon region plus 100 bases upstream and 100 bases downstream region were selected to call SNPs. Total 64,996 SNPs were called with 24,309 homozygous SNPs and 39,690 heterozygous SNPs. While from the same region, whole genome re-sequencing data called 67,294 SNPs, with 28,417 homozygous SNPs and 38,877SNPs. The low number of het SNPs call from WGRS data may be caused by the lower coverage of WGRS, which miscalled some het SNPs to hom SNPs. 45,106 of SNPs were found from both Exome and WGRS data, while 22,188 of SNPs were only found from WGRS data and 18,890 of SNPs were only found from Exome data. 22,625 of heterozygous SNPs were found from both Exome and WGRS data, while 16,252 of het SNPs were found from WGRS data and 17,065 of het SNPs were found from Exome data. Most of the false-negative missing may be caused by low coverage of Exome and WGRS sequencing.

4. 4 Haplotyping Algorithm

Firstly, WGRS data was mapped to human genome reference GRch37 using Bowtie 2 and generated a BAM file. Then Samtools was used to convert this BAM file to an mpileup file, which was then processed with varsan2 to make vcf file. All the variants were called in the vcf file and annotated with SnpEFF. Secondly, the het SNPs was used to generate a Het table (Figure 32).

Meanwhile, all SNPs were identified from dbSNP and generated rsHet table (Figure 33). Thirdly, the whole genome amplification sequencing data were mapped to reference GRch37 on Bowtie 2, and generated a BAM file. This BAM file was then processed using Samtools Targetcut to identify 20kb fragments and generated a new BAM file, and then aligned to genome. Then het was found from Het table by going through each line in BAM file to read 20kb fragments to generate a Contig table (Figure 32). To generate rsContig table, the het was found from rsHet table by going through each line in the BAM file to read 20kb fragments (Figure 33). Lastly, Het table and Contig table, rsHet table and rsContig table were combined respectively in MatLab to assemble contigs, and finally it generated assembled contigs respectively.

The algorithms in MatLab were processed as Peters, B., *et al*[255] with modification. Firstly, a linkage table was made for every two-linked hets (Figure 34). The number of the count indicated the frequency of the two hets. For example, in Figure 34, A-C linkage was found 15 times, while C-T linkage was found 11 times from the Het table and Contig table or rsHet table and rsContig talbe. Secondly, a graph of het SNPs linkage was generated by equation $score = (maximum - 1.5 \times mimum - junk - wrong)/normalized \times 100$ (Figure 34). To optimize the linkage graph, the linkages with lowest scores were removed to keep the unique linkage for all the hets. Finally, the hets were assigned to alleles, and then generated assembled contigs (Figure 34).

4. 5 Discussion

Whole genome amplification was successfully performed using as little as 0.3 pg human genomic DNA. The WGA input DNA concentration was tested by increasing the input to 0.4, 0.6, and 1.2pg genomic DNA. With increasing the input DNA amount from 0.3 pg/WGA to 1.2 pg/WGA, the heterozygous SNPs increased from <1% to ~11%, which was indicated by 50% scale as heterozygous SNPs. We found 1.2pg genomic DNA input was not giving too much heterozygous especially they can be filtered out easily, but it would to reduce the total of need samples.

WGA produced high quality DNA for Ion xpress libraries and Illumina Nextera libraries. Both sets of libraries were successfully sequenced on Ion PGM and HiSeq2000 respectively. Combining with the WGRS data from HiSeq2000, WGA data were able to perform haplotype phasing using the linkages of heterozygous SNPs. Due to the low sequencing depth, only ~50% of genome was covered with assembled genome in block of length ~84kb or greater, which meant more WGA sample sequencing was needed in order to obtain higher coverage and give longer haplotype blocks.

The work was not compared to HapMap project, because HapMap was mapped using reference hg18 and GRch36, while the all the data in this paper were mapped to updated reference hg19 and GRch37, which made it very complicated to compare with the different format, setting of files.

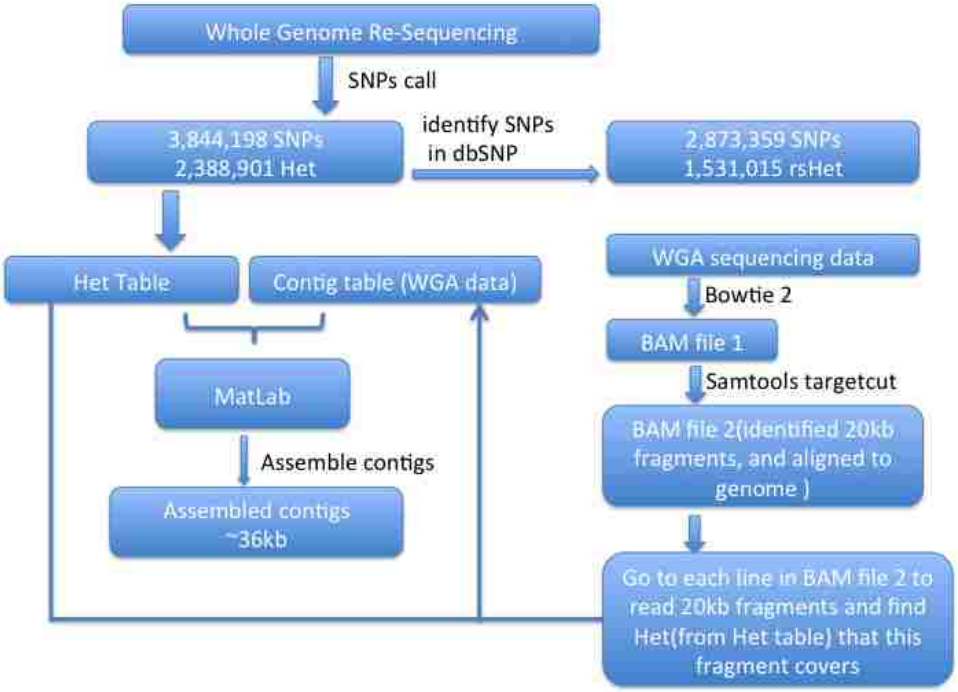


Figure 32 Process of assembly contigs using het table and contig table.

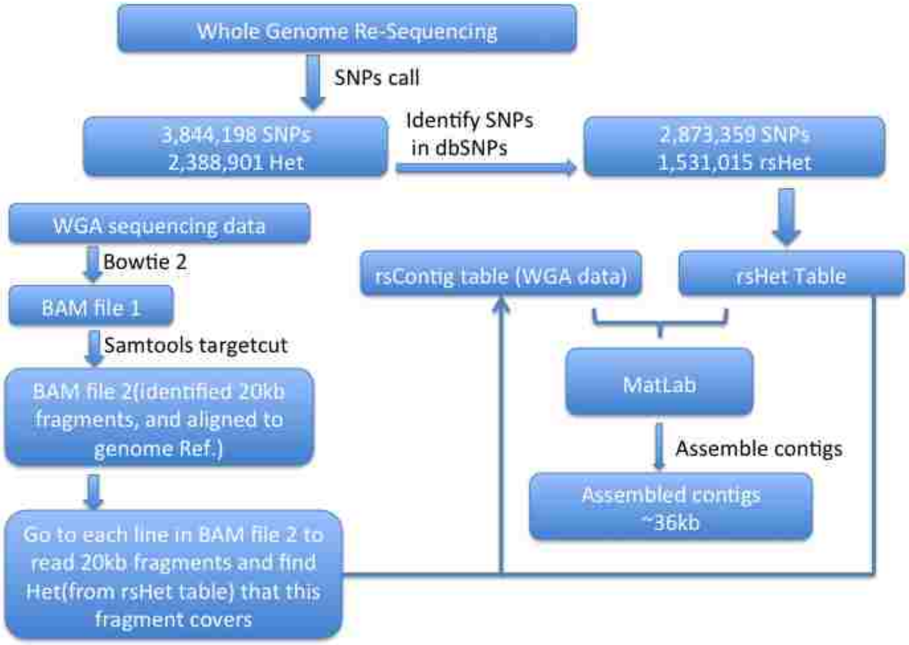
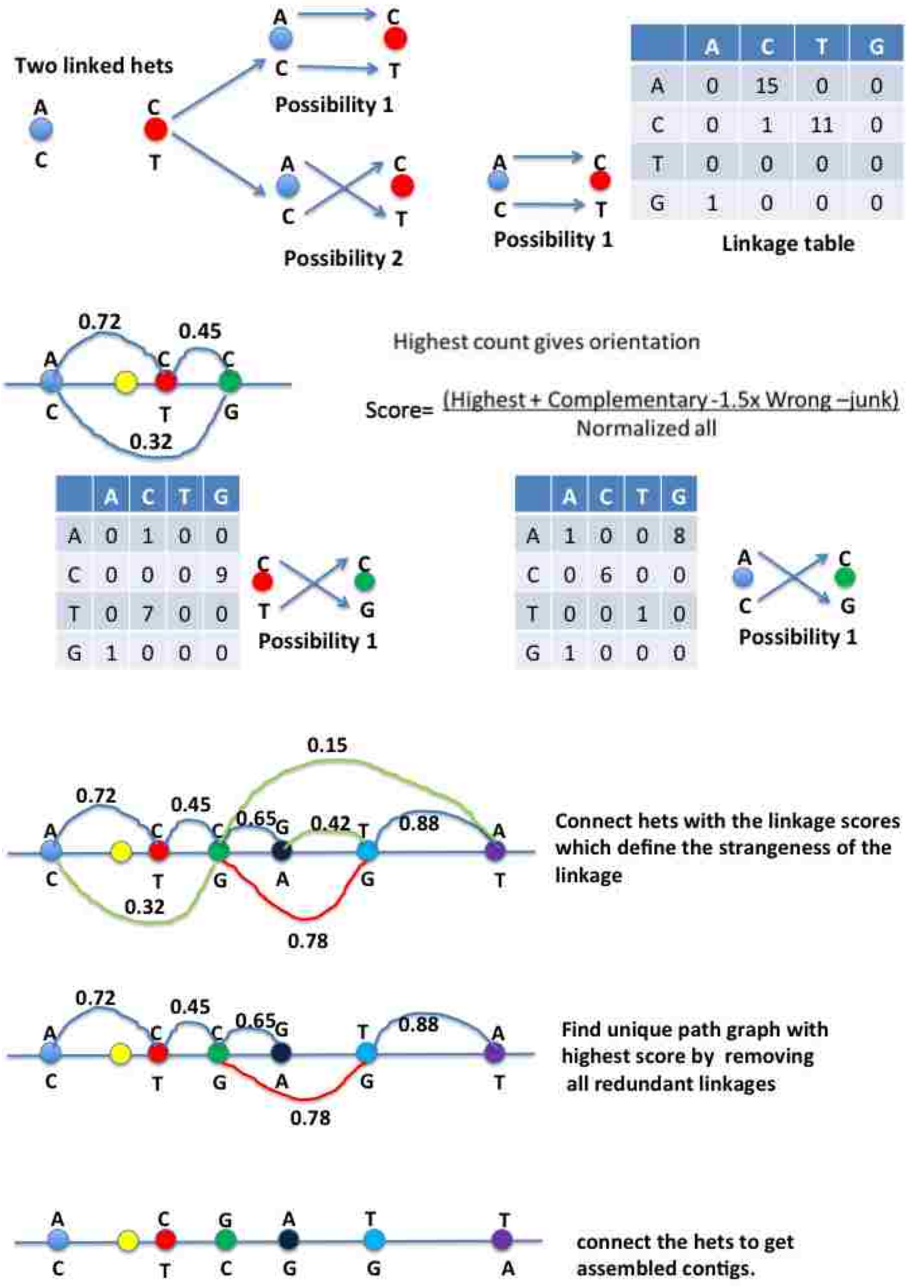

Whole Genome Re-Sequencing
SNPs call
3,844,198 SNPs
2,388,901 Het
Identify SNPs
in dbSNPs
2,873,359 SNPs
1,531,015 rsHet
WGA sequencing data
Bowtie 2
BAM file 1
Samtools targetcut
BAM file 2(identified 20kb fragments, and aligned to genome Ref.)
Go to each line in BAM file 2 to read 20kb fragments and find Het(from rsHet table) that this fragment covers
rsContig table (WGA data)
rsHet Table
MatLab
Assemble contigs
Assembled contigs
~36kb

Figure 33 Process of assembly contigs using rshet table and rsContig table.



Two linked hets
A
C
C
T
A
C
C
T
Possibility 1
A
C
C
T
Possibility 2
A
C
C
T
Possibility 1
Linkage table
0.72
0.45
A
C
C
C
T
G
0.32
Highest count gives orientation
Score= (Highest + Complementary -1.5x Wrong –junk) / Normalized all
C
T
C
G
Possibility 1
A
C
C
G
Possibility 1
0.15
0.72
0.45
0.65
0.42
0.88
A
C
C
G
T
A
C
T
G
A
G
T
0.32
0.78
Connect hets with the linkage scores which define the strangeness of the linkage
0.72
0.45
0.65
0.88
A
C
C
G
T
A
C
T
G
A
G
T
0.78
Find unique path graph with highest score by removing all redundant linkages
A
C
G
A
T
T
C
T
C
G
G
A
connect the hets to get assembled contigs.

	A	C	T	G
A	0	15	0	0
C	0	1	11	0
T	0	0	0	0
G	1	0	0	0

	A	C	T	G
A	0	1	0	0
C	0	0	0	9
T	0	7	0	0
G	1	0	0	0

	A	C	T	G
A	1	0	0	8
C	0	6	0	0
T	0	0	1	0
G	1	0	0	0

Figure 34 Algorithms of processing the linkages between every two linked hets and optimization of the graph and phasing the contigs by assigning the hets to chromosomes.

CHAPTER 5

SUMMARY

In this chapter, the significances and impacts of this dissertation were summarized, future work was discussed, and potential applications were anticipated.

5. 1 Summary of significances and impacts

In order to better serve the biomedical research and clinical applications, we demonstrated to improve the next generation sequencing technology in this dissertation.

Firstly, dual primer ePCR technique was developed to obtain longer reads for Polony sequencing. Dual primer ePCR technology simplified the library preparation by using fragment library instead of complicated mate-paired library. More remarkably, DPePCR technology provided paired-end reads which is very helpful for chromosome structure translocation study and genome mapping. This technique simplifies the library construction process, which shortens the sequencing turnover time and reducing the sequencing cost. The paired-end reads from DPePCR technology could be applied to study chromosome structure variations. For example, many cancer genome studies have shown that most cancer genome involves genome structure changes, chromosome rearrangements, which are very difficult to deal without paired-end reads. In addition, paired-end reads also simplify the mapping algorithms. DPePCR can be used to sequence small genomes, such as bacteria genomes. Also, it is capable for targeted sequencing, such as RNA-Seq, Exome sequencing, ChIP-Seq, etc.

Secondly, Rolony sequencing by synthesis technology not only simplified the library construction process, it also dramatically increased the throughput. Both SBL and SBS can be performed on Rolony patterned on grid array flowcell. The great feature of Rolony will innovatively change the sequencing field. Rolony has great feature comparing to polony and other beads based sequencing technology. 1. Rolony simplified the library construction by using fragment library. 2. The rolony is about 400nm diameter, which is much smaller than 1um bead. The smaller the sequencing template is, the higher density could be achieved on the same size sequencing flowcell, which means the throughput can be easily increased. 3. Rolony is very easy to pattern on grid flowcell which makes the sequencing and imaging process much easy and efficient. 4. Rolony can be sequenced by ligation and synthesis, which widen the potential applications for different platforms.

Finally, I demonstrated the application of NGS technology by developing whole genome amplification technique for haplotype-resolved sequencing. Haplotype information is very important for fully understanding human genome. It is also very important for completely describing genetic diversities and genetic ancestry. Moreover, it is essential for identifying associations between lots genetic variations and diseases. The conventional sequencing technologies were not able to provide haplotype information, while with our strategy; more detailed haplotype information can be exploited. The haplotype-resolved sequencing technology will open an era of studying associations between genetics and diseases.

5. 2 Future work

However, there are always potential to be improved for all kinds of technologies. The followings are potential aspects for improving.

First of all, dual primer ePCR technology was performed for sequencing by ligation, which could be adapted to sequencing by synthesis to increase the read length. The whole process is still labour consuming and experienced scientists to operate. It needs to be automated and simplified the operation before we can move and apply it to daily clinic sequencing.

Secondly, there are also limitations for Rolony sequencing technology. 1. Rolony employs rolling circle amplification which can only generate several thousands of original copy, limited the signal intensity, and thus eventually limited the read length. 2. The read length is also limited by library size. Circligase I and Circligse II can only circularize short fragments, which will constrict the library size, and eventually limited the read length. To achieve long reads and maintain good sequencing feature, other circularization process needs to be figured out. For example, T4 DNA ligase could be used to ligate longer linear DNA before getting rid of the complementary strand. For obtaining more copies of template, an extra Rolony amplification strategy may be performed post the RCA reaction.

For Haplotype-resolved sequencing technology, the limited WGA sequencing samples hindered us to assemble longer contigs. To this point, more WGA samples need to be sequenced to get Mb long contigs. After this, more individual sample needs to be sequenced as validation of this technology. Individuals with family generations are ideal for study the associations between genetics and diseases.

5. 3 Potential applications

Next generation sequencing has driven the entire biomedical sciences moving forward to next level[49]. In the last decade, next generation sequencing technology has been applied to all kinds of biomedical research[35,50-64]. There are several human genomes being sequenced[17,18,67-71]. In the last few years, there are thousands cancers have been sequenced[32,33,37-39,59,73-114]. All this will improve our understanding of human cancer, paling the road to eventually cancer therapy. Exome sequencing is another significant application of NGS technology. In the last few years, thousands of Exomes have been sequenced as exome sample across a variety of diseases[29-31,51,52,59,63,108,110,115-185]. Moreover, there are other applications. ChIP-seq is the example of using NGS technology to study protein-DNA interactions[186-190]. RNA-seq utilizes NGS technology to get higher resolution of RNA study[60,190-198]. NGS has also been used for research that could not been done before[34,35,199-202]. In a word, NGS technology will drive all kinds of research to personalized medicine, as well as improving other aspects of human life.

DPePCR technology could be used to sequence small genomes, such as bacteria genomes, targeted sequencing. Also, chromosome structure variations could be done by DPePCR sequencing technology based on the paired-end reads.

Rolony sequencing technology could potential change the sequencing field with its great features---bead-free, ePCR-free, small size, compatible with SBS and SBL. Also, Rolony could be used to perform electrical sequencing, which could change the sequencing world. While, there is lots optimization work to be done at this point. With the super high throughput, Rolony sequencing could be used to sequence human genome, cancer genome, exome, as well as other targeted sequencing.

Finally, for haplotype-resolved sequencing technology, it has tremendous potential for disease associated applications. Haplotype-resolved sequencing offers directly measurement of haplotype information, which is not possible for current massive parallel sequencing technologies. There are lots evidences showing that the haplotype information is essential for study most diseases, such as cancer. The super long reads/contigs generated from haplotype-resolved sequencing technology will benefit all the diseases researches. For example, the cancer genome will be more detailed analysed with super long contigs, potentially *de novo* cancer genome could be assembled. Allele specific expression caused diseases could also take advantage of haplotype information.

In a word, the innovative sequencing technology will reform the aspects of biomedical sciences and clinics, and play a more significant role to improve human life.

REFERENCE

1. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-8 (1953).

2. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-45 (2004).

3. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).

4. Gingeras, T.R., Higuchi, R., Kricka, L.J., Lo, Y.M. & Wittwer, C.T. Fifty years of molecular (DNA/RNA) diagnostics. *Clin Chem* **51**, 661-71 (2005).

5. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-7 (1977).

6. Venter, J.C. *et al.* Shotgun sequencing of the human genome. *Science* **280**, 1540-2 (1998).

7. Weissenbach, J. The Human Genome Project: from mapping to sequencing. *Clin Chem Lab Med* **36**, 511-4 (1998).

8. A nanobiosensor to detect single hybridization events supplemental material.

9. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46 (2010).

10. Ansorge, W.J. Next-generation DNA sequencing techniques. *N Biotechnol* **25**, 195-203 (2009).

11. Mardis, E.R. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387-402 (2008).

12. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-45 (2008).

13. Shendure, J.A., Porreca, G.J. & Church, G.M. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* **Chapter 7**, Unit 7 1 (2008).

14. Rothberg, J.M. & Leamon, J.H. The development and impact of 454 sequencing. *Nat Biotechnol* **26**, 1117-24 (2008).

15. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80 (2005).

16. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-32 (2005).

17. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* (2009).

18. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).

19. Guo, J. *et al.* Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc Natl Acad Sci U S A* **105**, 9145-50 (2008).

20. Harris, T.D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106-9 (2008).

21. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**, 847-50 (2009).

22. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-8 (2009).

23. Rothberg, J.M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348-52 (2011).

24. Schadt, E.E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum Mol Genet* **19**, R227-40 (2010).

25. Rhee, M. & Burns, M.A. Nanopore sequencing technology: nanopore preparations. *Trends Biotechnol* **25**, 174-81 (2007).

26. Luan, B. & Aksimentiev, A. Control and reversal of the electrophoretic force on DNA in a charged nanopore. *J Phys Condens Matter* **22**, 454123 (2010).

27. McNally, B. *et al.* Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano Lett* **10**, 2237-44 (2010).

28. Das, S.K. *et al.* Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res* **38**, e177 (2010).

29. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-5 (2010).

30. Ng, S.B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**, 790-3 (2010).

31. Hedges, D.J. *et al.* Exome sequencing of a multigenerational human pedigree. *PLoS One* **4**, e8232 (2009).

32. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685-96 (2010).

33. Aburatani, H. Cancer genome analysis through next-generation sequencing. *Gan To Kagaku Ryoho* **38**, 1-6 (2011).

34. Liu, G.E. Applications and case studies of the next-generation sequencing technologies in food, nutrition and agriculture. *Recent Pat Food Nutr Agric* **1**, 75-9 (2009).

35. Ekblom, R. & Galindo, J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**, 1-15 (2011).

36. Morozova, O. & Marra, M.A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255-64 (2008).

37. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719-24 (2009).

38. Chin, L., Andersen, J.N. & Futreal, P.A. Cancer genomics: from discovery science to personalized medicine. *Nat Med* **17**, 297-303 (2011).

39. Cheung, H.H., Lee, T.L., Rennert, O.M. & Chan, W.Y. DNA methylation of cancer genome. *Birth Defects Res C Embryo Today* **87**, 335-50 (2009).

40. Pemov, A., Modi, H., Chandler, D.P. & Bavykin, S. DNA analysis with multiplex microarray-enhanced PCR. *Nucleic Acids Res* **33**, e11 (2005).

41. Ahmadian, A., Ehn, M. & Hober, S. Pyrosequencing: history, biochemistry and future. *Clin Chim Acta* **363**, 83-94 (2006).

42. Mitra, R.D. & Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* **27**, e34 (1999).

43. Porreca, G.J., Shendure, J. & Church, G.M. Polony DNA sequencing. *Curr Protoc Mol Biol* **Chapter 7**, Unit 7 8 (2006).

44. Bormann Chung, C.A. *et al.* Whole methylome analysis by ultra-deep sequencing using two-base encoding. *PLoS One* **5**, e9320 (2010).

45. Porreca, G.J. Genome sequencing on nanoballs. *Nat Biotechnol* **28**, 43-4 (2010).

46. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814-8 (2009).

47. Bowers, J. *et al.* Virtual terminator nucleotides for next-generation DNA sequencing. *Nat Methods* **6**, 593-5 (2009).

48. Ibach, J. & Brakmann, S. Sequencing single DNA molecules in real time. *Angew Chem Int Ed Engl* **48**, 4683-5 (2009).

49. Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198-203 (2011).

50. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29**, 51-7 (2011).

51. Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* **43**, 442-6 (2011).

52. Sloan, J.L. *et al.* Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria. *Nat Genet* **43**, 883-6 (2011).

53. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**, R32 (2009).

54. Ku, C.S., Naidoo, N., Wu, M. & Soong, R. Studying the epigenome using next generation sequencing. *J Med Genet* **48**, 721-30 (2009).

55. Mir, K.U., Qi, H., Salata, O. & Scozzafava, G. Sequencing by Cyclic Ligation and Cleavage (CycLiC) directly on a microarray captured template. *Nucleic Acids Res* **37**, e5 (2009).

56. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809-13 (2009).

57. Vasta, V., Ng, S.B., Turner, E.H., Shendure, J. & Hahn, S.H. Next generation sequence analysis for mitochondrial disorders. *Genome Med* **1**, 100 (2009).

58. Yang, M.Q. *et al.* High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. *BMC Genomics* **10 Suppl 1**, I1 (2009).

59. Summerer, D. *et al.* Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform. *Genomics* **95**, 241-6 (2010).

60. Head, S.R. *et al.* Method for improved Illumina sequencing library preparation using NuGEN Ovation RNA-Seq System. *Biotechniques* **50**, 177-80 (2011).

61. Mihaly, Z. & Gyorffy, B. [Next generation sequencing technologies (NGST) -- development and applications]. *Orv Hetil* **152**, 55-62 (2011).

62. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-4 (2011).

63. Sulonen, A.M. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* **12**, R94 (2011).

64. Turajlic, S. *et al.* Whole genome sequencing of matched primary and metastatic acral melanomas. *Genome Res* **22**, 196-207 (2012).

65. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-6 (2008).

66. Samuel Levy1*, G.S., Pauline C. Ng1, Lars Feuk2, Aaron L. Halpern1, Brian P. Walenz1, Nelson Axelrod1,, Jiaqi Huang1, E.F.K., Gennady Denisov1, Yuan Lin1, Jeffrey R. MacDonald2, Andy Wing Chun Pang2,, Mary Shago2, T.B.S., Alexia Tsiamouri1, Vineet Bafna3, Vikas Bansal3, Saul A. Kravitz1, Dana A. Busam1,, Karen Y. Beeson1, T.C.M., Karin A. Remington1, Josep F. Abril4, John Gill1, Jon Borman1, Yu-Hui Rogers1, & Marvin E. Frazier1, S.W.S., Robert L. Strausberg1, J. Craig Venter1. The Diploid Genome Sequence of an Individual Human. *PLoS One* (2007).

67. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-5 (2008).

68. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 1527-41 (2009).

69. Kim, J.I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011-5 (2009).

70. Abecasis, G.R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).

71. Kitzman, J.O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* **29**, 59-63 (2011).

72. Efficient targeted resequencing of human germline and cancer genomes by oligocleotide-selective sequencing supplemental info 2.

73. Integrated analysis of homozygous deletions, focal amplifications, and sequence alternations in breast and colorectal cancers supporting material.

74. The cancer genome challenge.

75. Association between invasive ovarian cancer susceptibility and 11 breast candidate SNPs from breast cancer genome-wide association study supplemental info.

76. Muhammad Al-Hajj, M.S.W., Adalberto Benito-Hernandez, Sean J. Morrison, Michael F. Clarke. Prospective identification of tumorigenic breast cancer cells.

77. Futreal, P.A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).

78. Van den Eynden, G.G. *et al.* Overexpression of caveolin-1 and -2 in cell lines and in human samples of inflammatory breast cancer. *Breast Cancer Res Treat* **95**, 219-28 (2006).

79. Volik, S. *et al.* Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* **16**, 394-404 (2006).

80. Madrid breast cancer conference changes in the treatment of breast cancer. *Breast Cancer Res* (2007).

81. Wood, L.D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-13 (2007).

82. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-9 (2008).

83. Chen, X. *et al.* Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk. *Hum Mol Genet* **17**, 1336-48 (2008).

84. Chin, L. & Gray, J.W. Translating insights from the cancer genome into clinical practice. *Nature* **452**, 553-63 (2008).

85. Chittenden, T.W. *et al.* Functional classification analysis of somatically mutated genes in human breast and colorectal cancers. *Genomics* **91**, 508-11 (2008).

86. Leary, R.J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci U S A* **105**, 16224-9 (2008).

87. Miller, B.A., Chu, K.C., Hankey, B.F. & Ries, L.A. Cancer incidence and mortality patterns among specific Asian and Pacific Islander populations in the U.S. *Cancer Causes Control* **19**, 227-56 (2008).

88. Schneider, B.P. *et al.* Triple-negative breast cancer: risk factors to potential targets. *Clin Cancer Res* **14**, 8010-8 (2008).

89. Valle, L. *et al.* Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science* **321**, 1361-5 (2008).

90. X, C. Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk. *Hum Mol Genet* **17**, 1336-1348 (2008).

91. Hampton, O.A. *et al.* A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res* **19**, 167-77 (2009).

92. Mardis, E.R. & Wilson, R.K. Cancer genome sequencing: a review. *Hum Mol Genet* **18**, R163-8 (2009).

93. Song, H. *et al.* Association between invasive ovarian cancer susceptibility and 11 best candidate SNPs from breast cancer genome-wide association study. *Hum Mol Genet* **18**, 2297-304 (2009).

94. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-10 (2009).

95. Bell, D.W. Our changing view of the genomic landscape of cancer. *J Pathol* **220**, 231-43 (2010).

96. Bertucci, F., Finetti, P., Birnbaum, D. & Viens, P. Gene expression profiling of inflammatory breast cancer. *Cancer* **116**, 2783-93 (2010).

97. Cavallone, L. *et al.* Comprehensive BRCA1 and BRCA2 mutation analyses and review of French Canadian families with at least three cases of breast cancer. *Fam Cancer* **9**, 507-17 (2010).

98. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005 (2010).

99. Ding, L., Wendl, M.C., Koboldt, D.C. & Mardis, E.R. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum Mol Genet* **19**, R188-96 (2010).

100. Edwards, B.K. *et al.* Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer* **116**, 544-73 (2010).

101. Fletcher, O. & Houlston, R.S. Architecture of inherited susceptibility to common cancer. *Nat Rev Cancer* **10**, 353-61 (2010).

102. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-6 (2010).

103. Robison, K. Application of second-generation sequencing to cancer genomics. *Brief Bioinform* **11**, 524-34 (2010).

104. Ruike, Y., Imanaka, Y., Sato, F., Shimizu, K. & Tsujimoto, G. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics* **11**, 137 (2010).

105. Strausberg, R.L. & Simpson, A.J. Whole-genome cancer analysis as an approach to deeper understanding of tumour biology. *Br J Cancer* **102**, 243-8 (2010).

106. Takata, R. *et al.* Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat Genet* **42**, 751-4 (2010).

107. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).

108. Chang, H. *et al.* Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PLoS One* **6**, e21097 (2011).

109. Dorval, M. *et al.* Breast and ovarian cancer screening of non-carriers from BRCA1/2 mutation-positive families: 2-year follow-up of cohorts from France and Quebec. *Eur J Hum Genet* **19**, 494-9 (2011).

110. Ming-Liang He1, Y.C., Quan Chen1,2, Yaqing He3, Jing Zhao2, Jun Wang2, Huanming Yang2, Hsiang- & Kung1, F. Multiple gene dysfunctions lead to high cancer-susceptibility evidences from a whole-exome sequencing study. *Am J Cancer Res* (2011).

111. Ross, J.S. & Cronin, M. Whole cancer genome sequencing by next-generation methods. *Am J Clin Pathol* **136**, 527-39 (2011).

112. Zang, Z.J. *et al.* Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing. *Cancer Res* **71**, 29-39 (2011).

113. Ghoussaini, M. *et al.* Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet* **44**, 312-318 (2012).

114. Wu, C. *et al.* Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nat Genet* **44**, 62-6 (2012).

115. NimbleGen SeqCap EZ Exome Library LR User's Guide.

116. Unlocking mendelian disease using exome sequencing.

117. Computational and statistical approaches to analyzing variants identified by exome sequencing.

118. Exome sequencing a flash in a pan.

119. Exome sequencing as a tool for mendelian disease gene discovery supplemental info.

120. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers supplemental info.

121. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia supplemental info.

122. Exome sequencing the expert view.

123. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing supporting information.

124. Genetic Variation in an Individual Human Exome.

125. Multiple gene dysfunctions lead to high cancer-susceptibility evidences from a whole-exome sequencing study.

126. NimbleGen SeqCap EZ Exome Library LR User's Guide.

127. No place like exome.

128. The exome factor.

129. The GENCODE exome sequencing the complete human exome supplemental info.

130. Unlocking mendelian disease using exome sequencing.

131. Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways.

132. Editor. Exome sequencing a flash in a pan. *Science*.

133. Pauline C. Ng*, S.L., Jiaqi Huang, Timothy B. Stockwell, Brian P. Walenz, Kelvin Li, Nelson & Axelrod, D.A.B., Robert L. Strausberg, J. Craig Venter. Genetic Variation in an Individual Human Exome. *PLoS One* (2008).

134. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096-101 (2009).

135. Dale Hedges1, D.B., Eric Powell1, Cherylyn Almonte1, Jia Huang1, Stuart Young3, Benjamin, Boese2, M.S., Margaret A. Pericak-Vance1, Eden Martin1, Xinmin Zhang2, Timothy T. Harkins2, & Zuchner, S. Exome Sequencing of a Multigenerational Human Pedigree. *PLoS One* (2009).

136. Maher, B. Exome sequencing takes centre stage in cancer profiling. *Nature* **459**, 146-7 (2009).

137. Martz, L. No place like exome. *Nature* (2009).

138. Murim Choia, U.I.S., Weizhen Jia, Tiewen Liua, Irina R. Tikhonovab, Paul Zumbob, Ahmet Nayirc,, Ays¸in Bakkalog˘ lud, S.O., Sami Sanjade, Carol Nelson-Williamsa, Anita Farhia, Shrikant Maneb, & and Richard P. Liftona. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing supporting information. *Proc Natl Acad Sci U S A* (2009).

139. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-6 (2009).

140. Bainbridge, M.N. *et al.* Whole exome capture in solution with 3 Gbp of data. *Genome Biol* **11**, R62 (2010).

141. Biesecker, L.G. Exome sequencing makes medical genomics a reality. *Nat Genet* **42**, 13-4 (2010).

142. Bilguvar, K. *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**, 207-10 (2010).

143. Bolze, A. *et al.* Whole-exome-sequencing-based discovery of human FADD deficiency. *Am J Hum Genet* **87**, 873-81 (2010).

144. Criqui, A. & Baulande, S. Exome sequencing applied to monogenic disorders. *Med Sci (Paris)* **26**, 452-4 (2010).

145. Gilissen, C. *et al.* Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am J Hum Genet* **87**, 418-23 (2010).

146. Haack, T.B. *et al.* Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency supplemental info. *Nat Genet* **42**, 1131-4 (2010).

147. Haack, T.B. *et al.* Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nat Genet* **42**, 1131-4 (2010).

148. Johnson, J.O., Gibbs, J.R., Van Maldergem, L., Houlden, H. & Singleton, A.B. Exome sequencing in Brown-Vialetto-van Laere syndrome. *Am J Hum Genet* **87**, 567-9; author reply 569-70 (2010).

149. Kim, D.W., Nam, S.H., Kim, R.N., Choi, S.H. & Park, H.S. Whole human exome capture for high-throughput sequencing. *Genome* **53**, 568-74 (2010).

150. Krawitz, P.M. *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* **42**, 827-9 (2010).

151. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42**, 969-72 (2010).

152. Musunuru, K. *et al.* Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N Engl J Med* **363**, 2220-7 (2010).

153. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder supplemental info. *Nat Genet* **42**, 30-5 (2010).

154. Otto, E.A. *et al.* Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. *Nat Genet* **42**, 840-50 (2010).

155. Rosenfeld, J.A., Malhotra, A.K. & Lencz, T. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. *Nucleic Acids Res* **38**, 6102-11 (2010).

156. Teer, J.K. & Mullikin, J.C. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* **19**, R145-51 (2010).

157. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-8 (2010).

158. Zhao, Q. *et al.* Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing. *Genome Biol* **11**, R114 (2010).

159. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154-7 (2011).

160. Alison J. Coffey, F.K., Maria S. Calafato, Carol E. Scott, Priit Palta,, Eleanor Drury, C.J.J., Emily M. LeProust, Jen Harrow, Sarah Hunt, & Anna-Elina Lehesjoki, D.J.T., Tim J. Hubbard, Aarno Palotie. The GENCODE exome sequencing the complete human exome supplemental info. *Eur J Hum Genet* (2011).

161. Asan *et al.* Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* **12**, R95 (2011).

162. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-55 (2011).

163. Clark, M.J. *et al.* Performance comparison of exome DNA sequencing technologies supplemental info. *Nat Biotechnol* **29**, 908-14 (2011).

164. Clark, M.J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* **29**, 908-14 (2011).

165. Coffey, A.J. *et al.* The GENCODE exome: sequencing the complete human exome. *Eur J Hum Genet* **19**, 827-31 (2011).

166. Comino-Mendez, I. *et al.* Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat Genet* **43**, 663-7 (2011).

167. Furukawa, T. *et al.* Whole-exome sequencing uncovers frequent GNAS mutations in intraductal papillary mucinous neoplasms of the pancreas. *Sci Rep* **1**, 161 (2011).

168. Grossmann, V. *et al.* Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* **118**, 6153-63 (2011).

169. Han Chang*, D.G.J., Paul S. Kayne, Petra B. Ross-Macdonald, Rolf-Peter Ryseck, Nathan O. & Siemers. Exome Sequencing Reveals Comprehensive Genomic Alterations across Eight Cancer Cell Lines. *PLoS One* (2011).

170. Jordan, B. Making good use of exome sequencing. *Med Sci (Paris)* **26**, 1111-3 (2011).

171. Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc Natl Acad Sci U S A* **108**, 17087-92 (2011).

172. Leslie G Biesecker, K.V.S.a.J.C.M. Exome sequencing the expert view. *Genome Biol* (2011).

173. Nathan O Stitziel, A.K.a.S.S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol* (2011).

174. Park, D.J. *et al.* FAN1 variants identified in multiple-case early-onset breast cancer families via exome sequencing: no evidence for association with risk for breast cancer. *Breast Cancer Res Treat* **130**, 1043-9 (2011).

175. Parla, J.S. *et al.* A comparative analysis of exome capture. *Genome Biol* **12**, R97 (2011).

176. Stower, H. The exome factor. *Genome Biol* (2011).

177. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539-42 (2011).

178. Wang, K. *et al.* Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet* **43**, 1219-23 (2011).

179. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* **43**, 864-8 (2011).

180. Yan, X.J. *et al.* Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat Genet* **43**, 309-15 (2011).

181. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873-85 (2012).

182. Nikolaev, S.I. *et al.* Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat Genet* **44**, 133-9 (2012).

183. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44**, 47-52 (2012).

184. Stark, M.S. *et al.* Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. *Nat Genet* **44**, 165-9 (2012).

185. Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886-95 (2012).

186. Blow, M.J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-10 (2010).

187. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryo stem cells.

188. Lin, J. Chip-seq The new way to 'seq' genome wide for transcription factor binding patterns. *Biochem* (2010).

189. Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-80 (2009).

190. Ghosh, D. & Qin, Z.S. Statistical Issues in the Analysis of ChIP-Seq and RNA-Seq Data. *Genes* **1**, 317-334 (2010).

191. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-8 (2008).

192. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-82 (2009).

193. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).

194. Zhong Wang, M.G.a.M.S. RNA-Seq a revolutionary tool for transcriptomics. *Nature Reviews/Genetics* (2009).

195. Sengupta, S. *et al.* Highly consistent, fully representative mRNA-Seq libraries from ten nanograms of total RNA. *Biotechniques* **49**, 898-904 (2010).

196. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).

197. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-52 (2011).

198. Iyer, M.K. & Chinnaiyan, A.M. RNA-Seq unleashed. *Nat Biotechnol* **29**, 599-600 (2011).

199. de Magalhaes, J.P., Finch, C.E. & Janssens, G. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res Rev* **9**, 315-23 (2009).

200. Mardis, E.R. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med* **1**, 40 (2009).

201. Francisco M. De La Vega*, J.S., Fiona Hyland*, Wonkuk Kim¶, Stephen J. Finch¶, and Derek Gordon†. Applications of next-generation sequencing in genetic epidemiology. (2010).

202. Roukos, D.H. Next-generation sequencing and epigenome technologies: potential medical applications. *Expert Rev Med Devices* **7**, 723-6 (2010).

203. Tawfik, D.S. & Griffiths, A.D. Man-made cell-like compartments for molecular evolution. *Nat Biotechnol* **16**, 652-6 (1998).

204. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* **100**, 8817-22 (2003).

205. Metzker, M.L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46.

206. Fullwood, M.J., Wei, C.L., Liu, E.T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* **19**, 521-32 (2009).

207. Bayley, H. Sequencing single molecules of DNA. *Curr Opin Chem Biol* **10**, 628-37 (2006).

208. McCarthy, A. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem Biol* **17**, 675-6 (2010).

209. Robertson, J.A. The $1000 genome Ethical and Legal Issues in Whole Genome Sequencing of Individuals. *The American Journal of Bioethics* (2003).

210. Mardis, E.R. The $1000genome, the $100,000 analysis. *Genome Med* (2010).

211. Service, R.F. Gene sequencing The Race for the $1000 genome. *Science* (2006).

212. Flusberg, B.A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461-5 (2010).

213. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26**, 1146-53 (2008).

214. Bala Murali Venkatesan1 & and Rashid Bashir1, 3. Nanopore sensors for nucleic acid analysis. *Nature Biotechnology* (2011).

215. Cherf, G.M. *et al.* Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. *Nat Biotechnol* (2012).

216. Min, S.K., Kim, W.Y., Cho, Y. & Kim, K.S. Fast DNA sequencing with a graphene-based nanochannel device. *Nat Nanotechnol* **6**, 162-5 (2011).

217. Raffan, E. & Semple, R.K. Next generation sequencing--implications for clinical practice. *Br Med Bull* **99**, 53-71 (2011).

218. Magi, A. *et al.* Bioinformatics for Next Generation Sequencing Data. *Genes* **1**, 294-307 (2010).

219. Suk, E.K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res* **21**, 1672-85 (2011).

220. Nagarajan, N. & Pop, M. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *J Comput Biol* **16**, 897-908 (2009).

221. Dassanayake, M. *et al.* The genome of the extremophile crucifer Thellungiella parvula. *Nat Genet* **43**, 913-8 (2011).

222. Hutter, D. *et al.* Labeled nucleoside triphosphates with reversibly terminating aminoalkoxyl groups. *Nucleosides Nucleotides Nucleic Acids* **29**, 879-95 (2010).

223. Church, G.M. Genome for all. *Sci Am* (2005).

224. Xu, M.Y., Aragon, A.D., Mascarenas, M.R., Torrez-Martinez, N. & Edwards, J.S. Dual primer emulsion PCR for next- generation DNA sequencing. *Biotechniques* **48**, 409-12 (2010).

225. Mitra, R.D., Shendure, J., Olejnik, J., Edyta Krzymanska, O. & Church, G.M. Fluorescent in situ sequencing on polymerase colonies. *Anal Biochem* **320**, 55-65 (2003).

226. Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nat Genet* **38**, 382-7 (2006).

227. Rieger, C. *et al.* Polony analysis of gene expression in ES cells and blastocysts. *Nucleic Acids Res* **35**, e151 (2007).

228. Kim, J.B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481-4 (2007).

229. Church, G.M. Polony Sequencing Protocols_Church lab 12Dec07. (2007).

230. Jian Wu*†‡, S.Z., Qinglin Meng*†‡, Huanyan Cao*†, Zengmin Li*†, Xiaoxu Li*†, Shundi & Shi*, D.H.K., Lanrong Bi*†, Nicholas J. Turro†‡¶, Jingyue Ju*†¶. 3'-O-Modified nucleotides as reversible terminators for pyrosequencing supplemental information. *Proc Natl Acad Sci U S A* (2007).

231. Jingyue Ju*†¶, D.H.K., Lanrong Bi*†, Qinglin Meng*†‡, Xiaopeng Bai*†‡, Zengmin, Li*†, X.L., Mong Sano Marma*†, Shundi Shi*, Jian Wu*†‡, John R. Edwards*†, Aireen & Romu*, N.J.T. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators Supplemental information. *Proc Natl Acad Sci U S A* (2006).

232. Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A* **103**, 19635-40 (2006).

233. Ruparel, H. *et al.* Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc Natl Acad Sci U S A* **102**, 5932-7 (2005).

234. Seo, T.S. *et al.* Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc Natl Acad Sci U S A* **102**, 5926-31 (2005).

235. Seo, T.S. *et al.* Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci U S A* **101**, 5488-93 (2004).

236. Wu, J. *et al.* 3'-O-modified nucleotides as reversible terminators for pyrosequencing. *Proc Natl Acad Sci U S A* **104**, 16462-7 (2007).

237. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J. & Schork, N.J. The importance of phase information for human genomics. *Nat Rev Genet* **12**, 215-23 (2011).

238. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).

239. al., M.M.e. The Huntington’s disease candidate region exhibits many different haplotypes. *Nat Genet* **1**, 99–103 (1992).

240. al, Y.C.e. Positional cloning of the Werner’s syndrome gene. *Science* **272**, 258–262 (1996).

241. JC, K. Allele-specific gene expression uncovered. *Trends Genet* **20**, 113-116 (2004).

242. R, P. All elespecific gene expression is widespread across the genome and biological processes. *PLoS One* **4**(2009).

243. Zhang, D.e.a. Genetic control of individual diffrences in gene-specific methylation in human brain. *Am. J.Hum.Genet.* **86**, 411-419 (2010).

244. A, d.l.C. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. *Oncogene* **28**, 3345-3348 (2009).

245. L, V. Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science* **321**, 1361-1365 (2008).

246. Chamberlain, S.J. & Lalande, M. Neurodevelopmental disorders involving genomic imprinting at human chromosome 15q11-q13. *Neurobiol Dis* **39**, 13-20 (2010).

247. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).

248. Allen, A.S. & Satten, G.A. Robust testing of haplotype/disease association. *BMC Genet* **6 Suppl 1**, S69 (2005).

249. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-51 (2001).

250. Metzker, M.L. Sequencing technologies-the next generation poster.

251. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-22 (2010).

252. The international HapMap Project.

253. Consortium, T.I.H. Intergrating ethics and science in the International HapMap Project. *Nature Reviews/Genetics* (2004).

254. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12**, 703-14 (2011).

255. Peters, B.A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190-5 (2012).

256. Kaper, F. *et al.* Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc Natl Acad Sci U S A* **110**, 5552-7 (2013).

257. Pinard, R. *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, 216 (2006).

258. Hutchison, C.A., 3rd, Smith, H.O., Pfannkoch, C. & Venter, J.C. Cell-free cloning using phi29 DNA polymerase. *Proc Natl Acad Sci U S A* **102**, 17332-6 (2005).

259. Miao, K.R. *et al.* The polymorphism and haplotype analysis of HLA-A, -B and -DRB1 genes of population in Jiangsu province of China. *Int J Immunogenet* **34**, 419-24 (2007).

260. Yuan, F. *et al.* [Haplotype and linkage analysis of HLA-I classical genes in Chinese Han population]. *Zhongguo Shi Yan Xue Ye Xue Za Zhi* **15**, 1084-9 (2007).

261. Dasgupta, A., Misri, N. & Bala, S. Population and family studies to demonstrate Ir genes: HLA haplotype in atopic allergy. *Monogr Allergy* **11**, 75-9 (1977).

262. Pollack, M.S. *et al.* HLA linkage and B14, DR1, BfS haplotype association with the genes for late onset and cryptic 21-hydroxylase deficiency. *Am J Hum Genet* **33**, 540-50 (1981).

263. Reitnauer, P.J., DeMars, R. & Sondel, P.M. The proliferative immune response to autologous Epstein-Barr virus-transformed lymphoblastoid cells. I. Studies with HLA haplotype loss variants demonstrate a role for MHC-linked genes. *Hum Immunol* **13**, 177-91 (1985).

264. Rozanski, B.S., Pietrzyk, J.J. & Turowski, G. The maximum likelihood estimation of haplotype frequencies of HLA and HLA-like systems and addition loci of genes with the different modes of inheritance. *Acta Anthropogenet* **9**, 81-90 (1985).

265. Michon, J. *et al.* Highly polymorphic products of both HLA-DR and HLA-DQ genes contribute to the polymorphism of the HLA-DRw13 haplotype. *Immunogenetics* **24**, 177-83 (1986).

266. Wilton, A.N. & Charlton, B. Order of class III genes relative to HLA genes determined by the haplotype method. *Immunogenetics* **24**, 79-83 (1986).

267. Tanigaki, N., Tosi, R., Strominger, J.L. & Cooper, J. Immunochemistry of the HLA class II molecules isolated from a mouse cell transfected with DQ alpha and beta genes from a DR4 haplotype. *Immunogenetics* **26**, 40-7 (1987).

268. Liu, C.P., Bach, F.H. & Wu, S.K. Molecular studies of a rare DR2/LD-5a/DQw3 HLA class II haplotype. Multiple genetic mechanisms in the generation of polymorphic HLA class II genes. *J Immunol* **140**, 3631-9 (1988).

269. Carcassi, C. *et al.* The absence of 21-OH A and C4 B genes is a constant characteristic of the "Sardinian" HLA extended haplotype A30 Cw5 B18 BfF1 DR3 DRw52 DQw2. *Tissue Antigens* **34**, 85-9 (1989).

270. Contu, L., Carcassi, C. & Dausset, J. The "Sardinian" HLA-A30,B18,DR3,DQw2 haplotype constantly lacks the 21-OHA and C4B

genes. Is it an ancestral haplotype without duplication? *Immunogenetics* **30**, 13-7 (1989).

271. Nakagawa, H. *et al.* Study of HLA class I, class II and complement genes (C2, C4A, C4B and BF) in Japanese psoriatics and analysis of a newly-found high-risk haplotype by pulsed field gel electrophoresis. *Arch Dermatol Res* **283**, 281-4 (1991).

272. Hillert, J. & Olerup, O. Multiple sclerosis is associated with genes within or close to the HLA-DR-DQ subregion on a normal DR15,DQ6,Dw2 haplotype. *Neurology* **43**, 163-8 (1993).

273. Louis, P. *et al.* Polymorphism in the regulatory region of HLA-DRB genes correlating with haplotype evolution. *Immunogenetics* **38**, 21-6 (1993).

274. Narko, K., Levo, A. & Partanen, J. A rare neutral polymorphism in 21-hydroxylase genes as HLA haplotype marker. Evidence for strong founder effect in the Finnish population. *Hum Immunol* **43**, 66-71 (1995).

275. Alper, C.A. *et al.* Prospective analysis suggests susceptibility genes for deficiencies of IgA and several other immunoglobulins on the [HLA-B8, SC01, DR3] conserved extended haplotype. *Tissue Antigens* **56**, 207-16 (2000).

276. Selvaraj, P., Sriram, U., Mathan Kurian, S., Reetha, A.M. & Narayanan, P.R. Tumour necrosis factor alpha (-238 and -308) and beta gene polymorphisms in pulmonary tuberculosis: haplotype analysis with HLA-A, B and DR genes. *Tuberculosis (Edinb)* **81**, 335-41 (2001).

277. Vorechovsky, I. *et al.* Short tandem repeat (STR) haplotypes in HLA: an integrated 50-kb STR/linkage disequilibrium/gene map between the RING3 and HLA-B genes and identification of STR haplotype diversification in the class III region. *Eur J Hum Genet* **9**, 590-8 (2001).

278. Pyo, C.W. *et al.* HLA-E, HLA-F, and HLA-G polymorphism: genomic sequence defines haplotype structure and variation spanning the nonclassical class I genes. *Immunogenetics* **58**, 241-51 (2006).

279. Jindra, P., Venigova, P., Lysak, D., Steinerova, K. & Koza, V. Distribution of KIR genes in the population of unrelated individuals homozygous for ancestral haplotype AH8.1 (HLA-A1B8DR3). *Tissue Antigens* **76**, 240-4 (2010).

APPENDICES
APPENDIX A

DPePCR: A novel technique for DNA sequencing

PROTOCOL FOR:

Dual Primer Emulsion PCR for Next-Generation DNA Sequencing

Ming Yan Xu [1], Anthony D. Aragon [1], Norah Torrez-Martinez [1], Jeremy S. Edwards [1, 2, 3]

[1] Molecular Genetics and Microbiology, University of New Mexico Health Sciences Center, Albuquerque, NM 87131

[2] Chemical and Nuclear Engineering, University of New Mexico, Albuquerque, NM 87131

[3] Cancer Research and Treatment Center, University of New Mexico Health Sciences Center, Albuquerque, NM 87131

Note: This papper was published on BioTechniques: 10-12 (Protocol Guide 2011) doi 10.2144/000113542

Abstract and introduction

The Dual Primer Emulsion PCR (DPePCR) we previously described [190] can be used to generate a paired end library for sequencing using a next-generation sequencing platform [9-12]. DPePCR allows for faster and less expensive paired end sequencing when using ePCR–based amplification. The novel approach combines concepts from both emulsion PCR [16,186] and Bridge Amplification [40] (www.promega.com/ geneticidproc/ussymp7proc/0726.html) for the amplification of simple fragment libraries for paired-end sequencing. The DPePCR strategy can amplify ~200 base pair fragments (with the adaptors) and allows users to sequence both ends of the DNA fragment. This effectively shortens library preparation time, and increases library complexity [16] when compared to the construction of a mate paired DNA library. These improvements, based on the current method [43], will definitely contribute to the $1,000 genome goal.

REAGENTS

PAGE (6% Polyacrylamid Gels) (Invitrogen, Carlsbad, CA, USA)

PAGE Elution Buffer (Laboratory prepared)

Phenol: Chloroform: IsoAmylAlcohol (25:24:1) (Sigma-Aldrich，St. Louis, MO USA)

QIAquick PCR Purification Kits (Qiagen, Valencia, CA)

Fragmentase (England BioLabs, Ipswich, MA, New USA)

End-Repair Mix (Enzymatics, Beverly, MA, USA)

Taq DNA polymerase (England BioLabs, Ipswich, MA, New USA)

T4 DNA ligase (England BioLabs, Ipswich, MA, New USA)

T3 DNA ligase (England BioLabs, Ipswich, MA, New USA)

JumpStar Taq (England BioLabs, Ipswich, MA, New USA)

AcuI (England BioLabs, Ipswich, MA, New USA)

BceAI (England BioLabs, Ipswich, MA, New USA)

Antarctic Phosphatase (England BioLabs, Ipswich, MA, New USA)S-adenosylmethionine (SAM) (England BioLabs, Ipswich, MA, New USA)

Bovine serum albumin (BSA) (England BioLabs, Ipswich, MA, New USA)NaOH (Sigma-Aldrich，St. Louis, MO USA)

Isopropanol (Sigma-Aldrich，St. Louis, MO USA)

Ethanol (Sigma-Aldrich，St. Louis, MO USA)

Saline-sodium phosphate-EDTA solution (SSPE) (Fisher Scientific, Fairlawn, NJ, USA)

Tris·Cl (Fisher Scientific, Fairlawn, NJ, USA)

KCl (USB Corporation, Cleveland, OH, USA)

EDTA (Fisher Scientific, Fairlawn, NJ, USA)

NaCl (Sigma-Aldrich，St. Louis, MO USA)

Triton X-100 (Dow Chemicals, Midland, MI, USA)

MyOne Streptavidin beads (Invitrogen, Carlsbad, CA, USA)

Silicone oil AR20 (Sigma-Aldrich，St. Louis, MO USA)

749 FLUID (INCI name: cyclopentasiloxane and trimethylsiloxysilicate; Dow Chemicals, Midland, MI, USA)

5225C Formulation Aid (INCI name: cyclopentasiloxane and PEG/PPG-18/18 dimethicone; Dow Chemicals, Midland, MI, USA)

Oligos (all purchased from Integrated DNA Technologies, Inc., IA, USA)

	Primers	**Sequence of the primers**
Library Preparation Primers	FDV2-BceAI-Primer-F	5'-AAC CAC TAC GCC TCC GCT TTC CTC TCT ATG TCT ACT AGT CAA CGG CT -3'
	FDV2-BceAI-Primer-R	5' -/5Phos/ GCC GTT GAC TAG TAG ACA TAG AGA GGA AAG CGG AGG CGT AGT GGT T/3AmMO/-3'
	RDV2-AcuI-Primer-F	5'-ATA TGT CAA CTG CCC CGG GTT CCT CAT TCA CTC ATG ACT ACT GAA GT-3'
	RDV2-AcuI-Primer-R	5'-/5Phos/CTT CAG TAG TCA TGA GTG AAT GAG GAA CCC GGG GCA GTT GAC ATA T/3AmMO/-3'
	FDV2-BceAI-PCR	5'-AAC CAC TAC GCC TCC GCT TTC-3'
	RDV2-AcuI-PCR	5'-CTG CCC CGG GTT CCT CA -3'
Emulsion Primers	FDV2-dualbiotin	5'-/52-bio/TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT AAC CAC TAC GCC TCC GCT TTC CTC TCT ATG-3'
	RDV2-dualbiotin	5'-/52-Bio/ TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT

		TTT TTT TTT TTT TTT TTA TAT GTC AAC TGC CCC GGG TTC CTC ATT C-3’
	FDV2-short-unmodified-primer	5’-AAC CAC TAC GCC TCC GCT TTC CTC-3’
	RDV2-short-unmodified-primer	5’-A TAT GTC AAC TGC CCC GGG TTC CT-3’
Capping Oligos	BceAI-PlusPrimer-3’over-F	5’-/5Phos/NNT ATC ACT CAG ACT ATA CTC TAA ACC A-3’
	BceAI-PlusPrimer-3’over-R	5’-TAG AGT ATA GTC TGA GTG ATA/3AmMO/-3’
	AcuI-PlusPrimer-3’over-F	5’-/5Phos/TCT ACA TAT TCT ACC ATC TAC TTA TCA CAC A-3’
	AcuI-PlusPrimer-3’over-R	5’-GAT AAG TAG ATG GTA GAA TAT GTA GAN N/3Phos/-3’
Anchor Primers	FDV2-PM	5’-/5Phos/AGC CGT TGA CTA GTA GAC-3’
	Cy5-labeled FDV2-PM	5’-/Cy5/AGC CGT TGA CTA GTA GAC-3’
	RDV2-PM	5’-/5Phos/ACT TCA GTA GTC ATG AGT-3’
	BceAI-Plus-AchorPrimer	5’-TAG AGT ATA GTC TGA GTG ATA-3’

	AcuI-Plus-AchorPrimer	5’-GAT AAG TAG ATG GTA GAA TAT GTA GA-3’

-1 sequencing nonamers

5' Cy5-NNNNNNNNT

5' Cy3-NNNNNNNNA

5' TxRed-NNNNNNNNC

5' FAM-NNNNNNNNG

+1 sequencing nonamers

5’-/5Phos/TNN NNN NNN/3Cy5/-3’

5’-/5Phos/ANN NNN NNN/3Cy3/-3’

5’-/5Phos/CNN NNN NNN/3TxRed/-3’

5’-/5Phos/GNN NNN NNN/3FAM/-3’

All other nonamers (-2, -3, -4, -5, -6, -7, +2, +3, +4, +5, +6) are generated by moving the fixed base. Nonamers were purified via high performance liquid chromatography (HPLC) and degenerated bases were hand mixed to the 25:25:25:25 ratio.

Procedure

1 Prepare the DPePCR library

1.1. Fragment the *Streptococcus pyrogenes* genomic DNA by Fragmentase for 45-60 min at 37°C to achieve a size distribution of 75-150 bp.

1.2. The ~100-bp genome fragments are isolated by using PAGE.

1.3. DNA is extracted from the gel by placing gel fragments of the appropriate size in PAGE Elution Buffer and incubating at 50°C overnight followed by a Phenol: Chloroform: IsoAmylAlcohol (25:24:1) clean up.

1.4. The DNA was blunt-ended at 25°C for 30 min using End-Repair Mix, then the enzyme is heat inactivated at 70°C for 10 min then cleaned using Phenol: Chloroform: IsoAmylAlcohol (25:24:1).

1.5. The DNA is then A-tailed at 70°C for 30 min (Taq DNA polymerase) then cleaned using Qiagen column.

1.6. Ligate FDV2-BceAIPrimer-F/R and RDV2-AcuI-Primer-F/R at 25°C for 10 min (see “Preparation FDV2-BceAI-Primer-F/R (50 μM), RDV2-AcuI-Primer-F/R (50 μM), AcuI-PlusPrimer-3′over-F/R (50 μM), and BceAI-PlusPrimer-3′over-F/R (50 μM)”section) to the DNA fragments by T4 DNA polymerase.

1.7. PCR-amplify the products from step 4 for 13 cycles by using the following cycle parameters: 94°C for 2 min; 94°C for 30 s; 58°C for 30 s; 72°C for 90 s; Go to step 2; repeat 12 times; 72°C for 5 min; 4°C hold

1.8. The PCR products were then PAGE purified as described in steps 2 and 3 to isolate the “final” DPePCR library.

REST: the library can be stored at -20℃ for at least 12 months.

Bind forward and reverse PCR primers to microbeads prior to DPePCR

2 To bind the PCR primers to microbeads, in a 1.5-mL tube add 240 μL B&W buffer to 240 μL MyOne streptavidin beads.

3 Mix sample by inversion. Place on MPC to remove liquid by pipetting.

4 Wash beads two times with 480 μL B&W buffer.

5 Resuspend beads in 480 μL B&W buffer, then add 4.8 μL 1 mM FDV2-dual biotin and RDV2-dual biotin (note that the dual biotin primers are premixed before adding to the beads), vortex, and incubate at room temperature for 20 min with mixing.

****IMPORTANT: mix the two dual biotin primers first before adding them to the beads to ensure that equal numbers of both primers bind the beads.***

6 Wash beads three times with 600 μL B&W buffer. Resuspend beads in 640 μL TE.

REST: the DPePCR beads can be stored at 4℃ for at least 6 months.

Amplify the DPePCR library for sequencing

7 2 μL aliquots of fragment library (at the appropriate concentration [see Troubleshooting Problem 1] to allow a single molecule per drop) was mixed with 265 μL DPePCR master mix (see "Dual Primer ePCR Master Mix" table), 13 μL JumpStar Taq (20,000U/mL), 0.1 μM FDV2-Short-Unmodified-Primer and 0.1 μM RDV2-Short-Unmodified-Primer and 20 μL Dual Primer beads (see "Bind forward and reverse PCR primers to microbeads prior to DPePCR" section).

**HINT: the optimized library concentration is critical for single molecule amplification.*

Troubleshooting

8 The mixture was added to 400 μL silicone oil phase in a 2-mL tube and put on a TissueLyser II to emulsify (19 Hz for 90 s).

9 The DPePCR emulsion mixture was aliquoted to three 650-μL tubes and

thermocycled by using the following DPePCR cycle parameters:

9.1 94°C for 2 min

9.2 90°C for 15 s

9.3 56°C for 2 min

9.4 72°C for 45 s

9.5 Go to Step 2; repeat 99 times

9.6 90°C for 15 s

9.7 61°C for 5 min

9.8 Go to 6; repeat 19 times

9.9 72°C for 5 min

9.10 4°C hold

DPePCR emulsion breaking procedure

10 Add 400 μL of silicon oil AR20 to original PCR tubes, vortex, and pool into 1.5-mL tubes.

**HINT: We have observed variability in different lots of the surfactants in the oil phase. Every time a new batch of surfactant is used the ratio of surfactant to silicone oil must be optimized. The volume of aqueous is also important in determining the stability of the emulsion as well as the droplet size. The parameters used in the TissueLyser to get a 10 micron droplet may need to be adjusted with the use of the new Silicone Oil Phase. Once an appropriate oil phase is determined a mock PCR run should be completed to ensure that the emulsion does not break after 120 cycles.*

Troubleshooting

11 Spin 1 min at top speed and remove aqueous using pipette.

12 Wash two more times with 500 μL silicon oil AR20.

****HINT: use silicone oil AR20 to wash until the most emulsions broke, which may require more than 3 washes with silicon oil.***

13 Wash the beads three times with 300 μL 100% isopropanol (place tubes on MPC allow beads to collect beads and then remove aqueous using pipette for each wash).

14 Wash the beads three times with 300 μL 80% EtOH/annealing buffer.

15 Wash the beads three times with 300 μL 1× NXS (pool all samples into new tubes at first wash).

16 Wash the beads three times with 300 μL 1× TE/0.01% Triton.

17 After final wash, resuspend beads in 100 μL 1× TE/0.01% Triton.

**HINT: Beads MUST be clean before proceeding to the next step.*

REST: the DPePCR product beads can be stored at 4℃ overnight.

Restriction digestion to release the DNA to be sequenced

18 Digest the DPePCR product beads with 5 μL AcuI (5000 U/mL), 0.8 μL S-adenosylmethionine (SAM) (32 mM) and 394.5 μL 1× NEBuffer4 by incubating at 37°C for 60 min.

19 Wash the beads three times with NXS and three times with TE with 0.01% Triton.

20 After washing the DNA must be treated with a phosphatase to remove the 5' phosphoryl groups. To do this the DNA is treated with 2 μL Antarctic Phosphatase (5000U/mL) in 198 μL 1× Antarctic Phosphatase Reaction Buffer at 37°C for 60 min.

21 Wash the beads three times with NXS and three times with TE with 0.01%

Triton.

22 Digest the DNA with 4 μL BceAI (1000U/mL), in 392 μL1× NEBuffer 3 plus 4 μL BSA (10mg/mL) at 37°C for 60 min. After BceAI digestion, wash the beads three times with NXS and three times with TE with 0.01% Triton. ****HINT: Beads MUST be clean before proceeding to the next step.***

23 Before ligating capping adaptors, the DNA must be treated with a phosphotase to remove 5′ phosphoryl groups from nucleic acids. 2 μL of Antarctic Phosphatase (5000U/mL) in 198 μL of 1× Antarctic Phosphatase Reaction Buffer at 37°C for 60 min.

24 After dephosphorylation, the beads were washed three times with NXS and three times TE with 0.01% Triton.

****HINT: Beads MUST be clean before proceeding to the next step.***

25 For capping adaptors ligation, the beads were incubated with 2.5 μL of AcuI-PlusPrimer-3′over-F/R (50 μM), 2.5 μL of BceAI-PlusPrimer-3′over-F/R (50 μM), 2 μL of T4 DNA ligase (600,000 U/mL) and 193 μL of 1× quick ligation buffer at room temperature for 60 min.

26 The beads were treated with 0.1 M NaOH and washed three times with 1× saline-sodium phosphate-EDTA solution (SSPE)

****HINT: Beads MUST be clean before proceeding to the next step.***

Capping with dideoxynucleotide reaction

27 Mix the beads with 15 µL of $CoCl_2$ (2.5mM), 114µL of ddH_2O, 15µL of 10× Tailing buffer, 4µL of 1.25 mM (each) dideoxynucleotide mix, and 2µL of terminal transferase (20000U/mL) incubate at 37°C for 60min.

28 Wash 3 times with NXS.

29 Wash 3 times with 1× TE with 0.01% Triton.

30 Wash 3 times with PBS solution.

****HINT: Beads MUST be clean before proceeding to the next step.***

Sequence the DPePCR beads

31 The DPePCR beads were sequenced from the FDV2 and RDV2 strand respectively using the same strategy from both the 3′→5′ and 5′→3′ directions. For test sequencing of the FDV2 strand in 3′→5′direction, the beads were resuspended in135 µL of 1× SSPE with 15 µL of anchor primer

FDV2-PM. (100 μM), incubated at 70°C for 5 min, 50°C for 5 min and 25°C for 5 min, and washed three times with Wash 1E wash buffer

32 Then 1 μL of T3 DNA ligase and 200 μL of ligation buffer [2 μL of each nonamer (300 μM) and 200 μL of 1× DNA quick ligase buffer] were added to the beads and incubated at room temperature for 30 min.

33 The beads were then washed with 1E buffer and resuspended in 50 μL of 1E for observation on fluorescent microscope.

34 For test sequencing of the RDV2 strand in 3′→5′ direction, the beads were treated with 0.1M NaOH and then hybridized anchor primer RDV2-PM and ligated nanomers using the same protocol as FDV2 strand.

35 For 5′→3′ direction sequencing, the beads need to be capped with dideoxynucleotide (see "Capping with dideoxynucleotide reaction" section) before hybridization and ligation.

Optimize the DPePCR library concentration by hybridizing Cy5-labeled FDV-PM

36 1 μL of Cy5-Labeled FDV-PM (100 μM) and 99 μL of 6xSSPE were add to the denatured DPePCR product beads (from step 30). Mix them and vortex

the tube.

37 Incubate the beads at 50°C for 5 min

38 Incubate the beads at 37°C for 5 min

39 Incubate the beads at 25°C for 5 min

40 Wash the beads three times with 1E wash.

41 Resuspend the beads into 50 μL for fluorescent microscope observation.

REST: the beads can be stored at 4℃ for at least one week.

Anneal Primers

46. Mix 20 μL of FDV2-BceAIPrimer-F (100 μM) and 20 μL of FDV2-BceAI-Primer-R (100 μM), 20 μL of RDV2-AcuI-Primer-F (100 μM) and 20 μL of RDV2-AcuI-Primer-R (100 μM), 20μL of BceAI-PlusPrimer-3′over-F (100 μM) and 20 μL of BceAI-PlusPrimer-3′over-R (100 μM), and 20 μL of AcuI-PlusPrimer-3′over-F (100 μM) and 20 μL of AcuI-PlusPrimer-3′over-R (100 μM) respectively, and then follow the parameters:

- 80°C for 2 min
- 70°C for 2 min

- 60°C for 5 min
- 50°C for 5 min
- 40°C for 5 min
- 30°C for 5 min
- 20°C for 5 min
- 4°C hold

REST: the annealed primers can be stored at -20℃ for at least 6 months.

RECIPES

1. Bind and wash buffer (B & W buffer) (1L)

5 mM Tris·Cl, pH 7.5

0.5 mM EDTA, pH 8.0

1.0 M NaCl

2. NXS buffer (1L)

10 mM Tris·Cl, pH 7.5

1 mM EDTA, pH 8.0

100 mM NaCl

1% Triton X-100

3. Wash 1E wash buffer (5L)

10 mM Tris·Cl, pH 7.5

50 mM KCl

2 mM EDTA, pH 8.0

0.01% (v/v) Triton X-100

4. 80% Ethanol/annealing buffer (1L)

800mL EtOH

20mL Tris·Cl, pH 7.5 (1M)

5mL Magnesium Acetate (1M)

1mL Tween 80

174mL dH_2O

5. Silicone oil phase

The silicone oil phase is prepared by mixing:

10 g silicone oil AR20

7.5 g 749 FLUID

7.5 g 5225C Formulation Aid

The oil phase is vortexed and immediately aliquoted into single-use tubes and stored at room temperature.

****HINT: to weight the three components as accurate as possible and to make sure the silicone oil phase is mixing completely before aliquot to single tubes are critical to succeed.***

Troubleshooting

6. PAGE Elution Buffer (500mL)

5mL 1M Tris·Cl, pH 7.5

25mL 1M NaCl

1mL .5M EDTA

469mL ddH$_2$O

7. Library PCR Master Mix

Reagent	Final Concentration
10x Jumpstart Buffer	1x
dNTPs	0.2 mM
FDV2-BceAI-PCR	0.5 µM
RDV2-AcuI-PCR	0.5 µM
Jumpstart Taq	0.01 U/µl

M B Water	-

8. Dual Primer ePCR Master Mix:

Reagent	Final Concentration
10x EXT	1.2X
MgCl2 (50 mM)	7.0 mM
dNTPs (10 mM)	1.0 mM
Triton (10%)	0.1%
BSA (10 mg/ml)	1 mg/ml
FDV2 short unmodified primer	0.1 µM
RDV2 short unmodified primer	0.1 µM
Glycerol (50%)	5%
JumpStart Taq	0.1 U/µl
M B Water	-

TROUBLESHOOTING

PROBLEM 1. Multiple-amplification or too many negative beads

The DPePCR library concentration is the main factor which causes multiple-amplification or too many blank beads. The solution to this problem is to optimize the library concentration to one DNA library molecule to one emulsion drop. To optimize the library concentration we aim for 25% of the beads to be positive when hybridized with an anchor primer that contains Cy5. Because Cy5-labeled FDV-PM can only be specifically hybridized to the FDV strand of any beads which have been DPePCR amplified, the Cy5 positive beads indicate all the amplified beads and only these beads will be sequenced and show up different fluorescent colors.

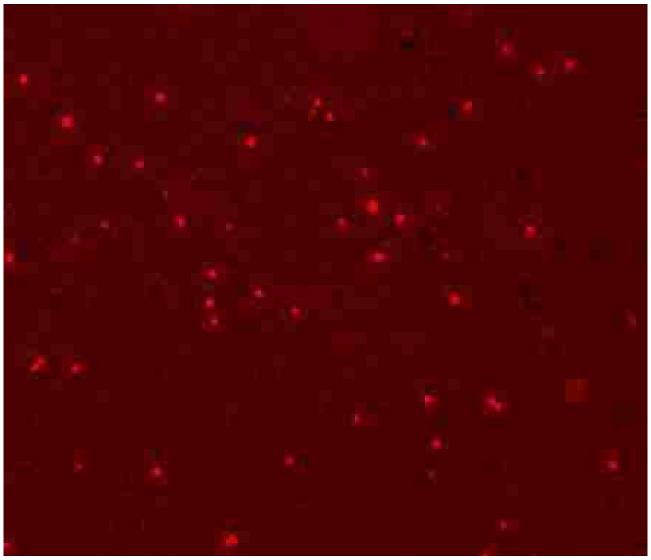


Figure 1. Optimize the DPePCR library by hybridizing Cy5-Labeled FDV-FM. The positive beads show up as red while the negative beads keep black under fluorescent microscope. The optimized ratio of the positive beads for acquiring best data is around 25%.

PROBLEM 2. Emulsion breaking after 120 cycles.

The first possible cause: when making the silicone oil phase, the mixture was not mixed properly. Ensure completely mixing vortex the solution for at least 10 min.

The second possible cause: Too much silicone oil AR20 or other surfactants may cause emulsion broken. When the three ingredients are weighed out, make sure the amount of each component is accurate. Different lots of these surfactants may contain different concentrations of the chemicals and still be qualified to sale, which may affect the surfactant property when making emulsion. Therefore it may be necessary to adjust the ratios of the three components of the silicone oil phase when using different lots of surfactants.

PROBLEM 3. Beads clumps

The possible cause: after restriction enzymes digestion, the enzymes may still bind to the DNA and may cause interaction between beads and beads, which may cause the beads clumping. If this happens, trying to wash the clumping beads three times with NXS buffer and three times with 1xTE with Triton can solve the problem.

EQUIPMENT

Magnetic Particle Concentrator (MPC) (Invitrogen, Carlsbad, CA, USA)

Pipet-lites multichannel pipette with LTS (Rainin Instrument, Oakland, CA, USA)

TissueLyser II (Qiagen, Valencia, CA, USA)

MJ PTC-200 Peltier Thermal Cycler (MJ Research, Inc., MA, USA)

Epi-fluorescence Microscopy (Nikon Inc., Melville, NY, USA)

High-speed mini-spin (Eppendorf, Hauppauge, NY, USA)

Mini centrifuge (Fisher Scientific, Fairlawn, NJ, USA)

Vortex-Genie 2 (Scientific Industries, Bohemia, NY, USA)

APPENDIX B

DNA Gel Purification protocol

Digestion DNA purification by Phenol: Chloroform: IsoAmyl Alcohol (25:24:1) (PCP)

1. pool 5 100ul tubes of amplified DNA into 2- 1.5ml tubes (500ul/tube)
2. Add 500ul PCP to each of the 1.5ml tube
3. Vortex and spin sample at room temperature 10min at top speed
4. Collect top layer and transfer to a fresh 1.5ml tube
5. Add 900ul Cold 100% Ethanol (Stored at -20℃)
6. Add 50ul 3M NaOAc (pH5.2)
7. Add 2ul Glycogen (20ug/ul)
8. Mix by vortexing
9. Store at -20℃ for 20min
10. Spin at max speed in microcentrifuge at 4℃ for 15min
11. Decant supernatant
12. Add 1ml 70% Cold Ethanol (Stored at -20℃) to DNA pellet
13. Vortex and spin at max speed in microcentrifuge at room temperature for 10min
14. Decant supernatant, blot samples on clean paper towel
15. Air dry DNA

16. Resuspend DNA 40ul 1x TE (Final Volume = 80ul)

17. Add 40ul 5x PAGE loading buffer to 80ul DNA sample

PAGE purification of tag-T30-tag

1. Run the full amount of the digested DNA on a pre-cast 6% Polyacryamide gel (PAGE)
2. Load 20ul/lane
3. Run a 25 bp ladder
4. Run gel approximately 20min. DO NOT ALLOW LOWER MARKER DYE BAND TO RUN OFF OF GEL OR 70bp PRODUCT WILL BE LOST!
5. Stain gel with 50ml 1X TBE and 5ul SYBR-GREEN for 10 min while shaking
6. Cut the bands between 150-250bp
7. Add 600ul 1X PAGE Elution Buffer to a 1.5ml tube. Place excised gel bands in elution buffer
8. Vortex tube to make sure gel is immersed in elution buffer
9. Incubate excised gel and elution buffer at 50℃ overnight

Gel Eluted DNA purified by Phenol: Chloroform: IsoAmyl Alcohol (25:24;1) (PCP)

1. Spin samples 5 min at room temperature max speed

2. Label a 1.5ml tube for each sample and add 600ul PCP to the 1.5ml tube
3. Remove aqueous away from gel (should recover ~600ul) and add to 1.5ml tube with PCP
4. Vortex and spin sample at room temperature 10min at top speed
5. Remove top layer and transfer to a fresh 1.5ml tube
6. Add 900ul Cold 100% Ethanol (Stored at -20℃)
7. Add 60ul 3M NaOAc (pH5.2)
8. Add 2ul Glycogen (20ug/ul)
9. Mix by vortexing
10. Store at -20℃ for 20min
11. Spin at max speed in microcentrifuge at 4℃ for 15min
12. Decant supernatant
13. Add 1ml 70% Cold Ethanol (Stored at -20℃) to DNA pellet
14. Vortex and spin at max speed in microcentrifuge at room temperature for 10min
15. Decant supernatant, blot samples on clean paper towel
16. Air dry DNA
17. Resuspend DNA in 20ul 1x EB

DNA concentration by Nanodrop

- Use 1x EB buffer as blank for Nanodrop

- Measure 1ul purified DNA

Sample	260/280	ng/ul	Total ng	100ng DNA(ul)

APPENDIX C

Lambda DNA fragment Rolony library preparation

Jul. 28-Aug.2nd 2011

1. Fragmentation lambda DNA

Reaction Components	Starting amount
Lambda DNA(0.5ug/uL)	10uL
10x Fragmentase reaction buffer	20uL
100x BSA	2uL
MB Water	148uL
Final Volume	200uL

a. Vortex thoroughly.

b. Incubate on ice for 5min

c. Vortex NEBNext dsDNA fragmentase and add 20uL to the reaction

d. Vortex thoroughly

e. Incubate at 37℃ for 1hour.

f. Add 10uL of 0.5M EDTA to stop the reaction.

g. Mini Elute PCR purification kit (50) (Qiagen) clean.

1. Mix the reaction with 1mL PB, load on the column and spin for 1min. Toss the solution.

2. Add 750uL PE into the column, spin 1min. Toss the solution, spin 2min.

3. Change a tube and add 30uL EB to the center of the column, RT for 5min, spin 2min.

2. Gel purification and selection

a. Add 10uL DNA loading buffer into 30uL DNA

b. Load 5ul of 25bp DNA ladder on a 6% PAGE

c. Load 20uL of DNA on the gel (2 lanes)

d. Run 200mV for 25min

e. Stop gel running.

f. Place the gel into a box containing 5uL Cyber Green in 50uL 1x TBE, shake and incubate at RT for 10min.

g. Place the gel in the gel image machine.

h. Turn on the UV, manually exposure for 1.5 seconds

i. Freeze and print the image. Turn off the UV.

j. Under UV, cut out the bands (smear) as needed(75-200, 200-400bp)

k. Cut small pieces and put bands (smear) into 500uL PAGE Elution Buffer, incubate at 55℃ overnight.

3. Phenol purification

a. Spin down the gel, and add the supernatant solution into 500uL phenol mix (phenol: chloroform: isoamylalcohol = 25: 24: 1). Vortex and spin at top speed for 10min.

b. Collect the supernatant (~500uL), add 900uL 100% EtOH, 2uL Glycogen (20ug/uL) and 60uL 3M NaOAc (pH=5.2)

c. Vortex and store at -20℃ for 20min.

d. Spin for 15min at 4℃. Toss the solution, keep the DNA pellet.

e. Add 1mL 70% EtOH to the pellet. Vortex and spin down for 10min. Toss the solution, air dry the pellet.

f. Add 50uL EB to DNA pellet. (Ready for blunt-end repair)

4. Blunt-end repair (100ul total volume)

10uL 10x blunt-end buffer

50uL DNA in EB

1uL dNTPs (1mM)

2uL Enzyme

37uL H_2O

Incubate at 25℃ for 30min

Qiagen clean, suspend into 30uL EB

5. A-tailed (100ul total volume)

30uL DNA in EB

10uL 10x Standard buffer

1uL dATP (100mM)

2uL Tag DNA polymerase

57uL H2O

Incubate at 70℃ for 30min

Qiagen clean, and suspend into 30uL EB. NanoDrop

6. **Adaptor ligation (100ul total volume)**

30uL DNA in EB

~ Adaptors (50uM) (Rolony_F-adaptor 1/2 and Rolony_R-adaptor 1/2) (DNA: Adaptor = 1:10)

50uL 2x QLB

4uL T4 DNA ligase

~ H2O

Incubate at 25℃ for 30min

Qiagen clean, suspend into 50uL EB.

7. **PCR_1 Application**

50uL DNA

100uL 10x Jumpstar Buffer

20uL dNTPs (10mM)

5uL 100uM 5P_PCR_Primer

5uL 100uM 5B_PCR_Primer

815uL H2O

5uL JumpStar TAQ

94℃ 2min

94℃ 30sec

46℃ 30sec

72℃ 30sec

Go to 2nd step 29 cycles

72℃ 5min

4℃ forever

Phenol clean, suspend into 50uL EB.

Nanodrop result:

F1(125-200bp): 484.3ng/ul, 319.5ng/ul 260/280=1.70   ~400pmole

F2(200-400bp): 278.8ng/ul, 355ng/ul; 260/280=1.80,   ~160pmole

8. **Beads selection**

   a. Take 100uL Myone beads; wash with 1xB&W 2 times.

   b. Add 50uL DNA to the beads, add 100uL 2x B&W and 50uL H_2O.

   c. Incubate at RT for 1hour with rotisserie.

d. Wash with B&W 2times. Then add 100uL 0.1N NaOH, and incubate at RT for 10min.

e. Collect the NaOH solution, add 150uL NaOH to wash the beads and collect NaOH solution, add 250uL Tris-HCl (1M, pH=7.5) to NaOH.

f. Add 900uL 100% EtOH, 2uL Glycogen (20ug/uL) and 60uL 3M NaOAc (pH=5.2). Vortex and store at -20℃ for 20min.

g. Spin for 15min at 4℃. Toss the solution, keep the DNA pellet.

h. Add 1mL 70% EtOH to the pellet. Vortex and spin down for 10min. Toss the solution, air dry the pellet.

i. Add 20uL EB to DNA pellet.

j. (Add 2500uL PB into 500uL, mix and then add them to 4 Qiagen columns. Spin 1min. Toss the solution, and add 750uL PE to wash the columns. Spin 1min, toss the solution, spin empty columns for 2min. Then toss the column tube and remove the column to a clean tube. Add 10uL EB to the central of the column, incubate at RT for 5min and spin 2min to collect the DNA samples.)

For circligase strategy (Qubit result):

F1(125-200bp): 71ng/ul ~13.6pmole

F2(200-400bp): 85.6ng/ul; ~8.79pmole

9. Circularization of library using Circligase

1.1 Assemble reaction as follow:

	Volume(µL)	Final Concentration
H_2O	5	
10x reaction buffer	2	1x
5’P_ssDNA library(linear DNA from step 10)	10	~1-10pmol
50mM $MnCl_2$	1	2.5mM
1mM ATP	1	50mM
CircLigase	1	5U/µL (Epicenter CL9021K)

Incubate at 60℃ for 60min

Heat inactivate at 80℃ for 10min

1.2 Digest residual linear ssDNA

Add directly to the Circligase reaction solution:

➢ ExonucleaseI (20U/µL)	1µL
➢ ExonucleaseIII (100U/µL)	1µL

Incubate at 37℃ for 45min

Heat inactivate at 80℃ for 15min

1.3 Purify circularized template using Qiagen Mini Elute Purification PCR kit

- Add 5-fold PB to reaction solution, load it into Qiagen column. Spin 1min at top speed, and then toss the liquid.
- Add 750µL PE to the column; spin 2min at top speed. Dump the liquid, empty spin 2min at top speed.
- Use a new tube to collect the COT product. Add 20µL EB to the column; Spin 2min at top speed.
- Take 1µL to do Nanodrop, using EB to blank.

Qubit result:

F1 (125-200bp): 26ng/ul;   ~7.51pmole

F2 (200-400bp): 22ng/ul;   ~3.39pmole

10. Rolling Circle Amplification

Assemble RCA reaction as follow:

	Volume (µL)	Final concentration
H_2O	74	
10X Reaction buffer	10	1x
Circularized DNA library (~4ng/µL)	5	~1pmole

1µM Right Hairpin RCA-Primer	1	1pmole
1mM AA-dUTP	1	10uM
100mM DTT	4	4mM
25mM dNTPs	4	1mM each

Incubate at 95℃ for 2min, then 60℃ for 5min, 4℃ pause

Add following on ice:

- ✧ Phi29 DNA Polymerase (10U/µL)(Enzymatics) 1µL 10U

Incubate at 30℃ for 6hours, then 4℃ forever.

Add 300µL of PBS, and then store at 4℃ until needed.

Lamda DNA fragment Library oligos

Final library				
Circligase	/5PHOS/ GTT CCT CAT TCT CTG AAG A TNN NNN NNN NNN ... NNN NNN NNN NNN N AC TTC AGC TGC CCC GG			
T4 Ligase	/5PHOS/ACG CGT GTT CCT CAT TCT CTG AAG A TNN NNN NNN NNN ... NNN NNN NNN NNN N AC TTC AGC TGC CCC GG			
TM	**5P_PCR_Primer**		**5P_PCR2_Primer**	
47.68	/5PHOS/GTT CCT CAT TCT CTG AAG AT	57.9615	/5Phos/ACG CGT GTT CCT CAT TCT CTG AAG AT	
	5B_PCR_Primer		**5B_PCR2_Primer**	
54.29	/5Biosg/CC GGG GCA GCT GAA GT	64.187	/5Biosg/ACG CGT CC GGG GCA GCT GAA GT	
	dsDNA Fragment library			
74.88	5-GTT CCT CAT TCT CTG AAG AT NN NNN NNN NNN ... NNN NNN NNN NNN NA C TTC AGC TGC CCC GG			
74.88	3-CAA GGA GTA AGA GAC TTC T ANN NNN NNN NNN ... NNN NNN NNN NNN N TG AAG TCG ACG GGG CC			
	Rolony_F-adaptor 1/2		**Rolony_R-adaptor 1/2**	
47.68	5-GTT CCT CAT TCT CTG AAG AT		5'/5Phos/C TTC AGC TGC CCC GG-3'	50.14
46.77	3-CAA GGA GTA AGA GAC TTC T/5Phos/		3-TG AAG TCG ACG GGG CC-5'	51.06

		Minus Anchor			
	RCA-Primer 2	**Primer 2**			
57.9	AAT GAG GAA CCC GGG	/5Phos/ACT TCA GCT			47.4
3	GCA GC	GCC CCG			1
				RCA-Primer 2	
				LamdaDNA_T4	
	Plus anchor primer "-" 1			**GAA CAC GCG T**	
	= AP2		56.7	**CC GGG GC**	
60.9	5'- CCC CGG GTT CCT CAT TCT CTG			**Plus AP2 N-1**	
6	AAG A -3'			**LamdaDNA_T4**	
			67.0	CCC CGG ACG CGT GTT CCT CAT	
			161	TCT CTG AAG A	
	Ligate circle template				
	probe				
67.9	5'-TCT TCA GAG AAT GAG GAA C CCG				
5	GGG CAG CTG AAG T				

APPENDIX D

Fragment DNA library and Rolony preparation

1. Fragmentation of genomic DNA

Reaction Components	Starting amount
Genomic DNA(0.5ug/uL)	10uL
10x Fragmentase reaction buffer	20uL
100x BSA	2uL
Molecular Biology Water	148uL
Final Volume	200uL

h. Vortex thoroughly.

i. Incubate on ice for 5min

j. Vortex NEBNext dsDNA fragmentase and add 20uL to the reaction

k. Vortex thoroughly

l. Incubate at 37℃ for 1hour.

m. Add 10uL of 0.5M EDTA to stop the reaction.

n. Mini Elute PCR purification kit (50) (Qiagen) clean (for all the spin, the speed is 1300rpm in this protocol).

   1. Mix the reaction with 5-fold volume of PB, load on the column and spin for 1min. Toss the solution.

   2. Add 750uL PE into the column, spin 1min. Toss the solution, spin 2min.

   3. Change a tube and add 30uL EB to the center of the column, RT for 5min, spin 2min.

2. **Gel purification and selection**

   a. Add 10uL DNA loading buffer into 30uL DNA

   b. Load 5ul of 25bp DNA ladder on a 6% PAGE

   c. Load 20uL of DNA on the gel (2 lanes)

   d. Run 200mV for 25min

   e. Stop gel running.

   f. Place the gel into a box containing 5uL Cyber Green in 50uL 1x TBE, shake and incubate at RT for 10min.

   g. Place the gel in the gel image machine.

   h. Turn on the UV, manually exposure for 1.5 seconds

   i. Freeze and print the image. Turn off the UV.

   j. Under UV, cut out the bands (smear) as needed(i.e. 100-300bp)

   k. Cut small pieces and put bands (smear) into 500uL PAGE Elution Buffer, incubate at 55℃ overnight.

Can stop here.

3. **Phenol purification**

   a. Spin down the gel, and add the supernatant solution into 500uL phenol mix (phenol: chloroform: isoamylalcohol = 25: 24: 1). Vortex and spin at top speed for 10min.

   b. Collect the supernatant (~500uL), add 900uL 100% EtOH, 2uL Glycogen (20ug/uL) and 60uL 3M NaOAc (pH=5.2)

   c. Vortex and store at -20℃ for 20min.

d. Spin for 15min at 4℃. Toss the solution, keep the DNA pellet.

e. Add 1mL 70% EtOH to the pellet. Vortex and spin down for 10min. Toss the solution, air dry the pellet.

f. Add 50uL EB to DNA pellet. (Ready for blunt-end repair)

Can stop here.

4. Blunt-end repair (100ul total volume)

10uL 10x blunt-end buffer

50uL DNA in EB

1uL dNTPs (1mM)

2uL Enzyme

37uL H_2O

Incubate at 25℃ for 30min

Qiagen clean, suspend into 30uL EB.

5. A-tail (100ul total volume)

30uL DNA in EB

10uL 10x Standard buffer

1uL dATP (100mM)

2uL Tag DNA polymerase

57uL H2O

Incubate at 70℃ for 30min

Qiagen clean, and suspend into 30uL EB. Qubit measures DNA concentration.

6. Adaptor ligation (100ul total volume)

30uL DNA in EB

~ Adapters (50uM stock) (Rolony_F-adapter 1/2 and Rolony_R-adapter 1/2) (DNA: Adapter = 1:50, Qubit reads) (The adapters F 1 and F2, R1 and R2 should be annealed respectively before adding to the ligation solution.)

50uL 2x QLB

4uL T4 DNA ligase

H_2O (add Molecular Biology water to adjust total volume to 100uL)

Incubate at 25℃ for 30min

Qiagen clean, suspend into 50uL EB.

Can stop here.

7. PCR Amplification

50uL DNA

100uL 10x EXT Buffer

20uL dNTPs(10mM)

5uL 100uM 5P_PCR_Primer

5uL 100uM 5B_PCR_Primer

815uL H2O

5uL Taq-B

The total volume is 1000uL, aliquot 125ul each to 8 PCR tubes.

The PCR parameters:

94℃ 2min

94℃ 30sec

46℃ 30sec

68℃ 30sec

Go to 2nd step 12 cycles

68℃ 5min

4℃ forever

Pool 4 PCR tube sample to one Phenol tube (contains 500uL phenol mix). Do phenol purification as step 3. And then, re-suspend sample into 50uL EB (100uL sample for 2 tubes). Do Qubit test to measure the DNA concentration. Should get ~100pmol dsDNA sample.

Can stop here.

8. **Beads selection (for 20ug DNA scale)**
    a. Take 100uL Myone strepavidin beads; wash with 1xB&W 2 times.
    b. Add 50uL DNA (=< 20ug) to the beads, add 100uL 2xB&W, and 50uL H_2O.
    c. Incubate the beads at RT for 30min with rotisserie.
    d. Wash with 1xB&W 2times.
    e. Add 100uL 0.125N NaOH, and incubate at RT for 10min.
    f. Collect the NaOH solution, add 115uL 0.125N NaOH to wash the

beads and collect NaOH solution, add 20uL Tris-HCl (1M, pH=7.5) and 15uL 3M NaOAc (Sodium acetate) (pH5.2) to NaOH to buffer the pH to 7.5.

g. Add 1250uL PB into 250uL sample, mix well. If the pH indicator changed to non-yellow, add more NaOAc to adjust pH to 7.5 (pH indicator should show yellow). Then add them equally to 4 Qiagen Minielute PCR purification columns. Do Qiagen clean and re-suspend DNA into 20uL EB.

h. Use Qubit to measure the DNA concentration. Should get ~30% sample back (theoretically can get back 50%). Take 10pmol to process step 9, and keep the rest sample in -20℃ freezer, it will be good at least 3 months

Can stop here.

9. Circularization of library using Circligase

1.4 Assemble reaction as follow:

	Volume(μL)	Final Concentration
H_2O	5	
10x reaction buffer	2	1x
5'P_ssDNA library(linear DNA from step 8)	10 (~10pmol)	
50mM $MnCl_2$	1	2.5mM
1mM ATP	1	50mM

CircLigase	1	5U/μL (Epicenter CL9021K)

Incubate at 60℃ for 60min

Heat inactivate at 80℃ for 10min

1.5 Digest residual linear ssDNA

Add directly to the Circligase reaction solution:

- ExonucleaseI (20U/μL) 1μL
- ExonucleaseIII (100U/μL) 1μL

Incubate at 37℃ for 45min; Heat inactivate at 80℃ for 15min

1.3 Purify circularized template using Qiagen Mini Elute Purification PCR kit, and re-suspend into 20μL EB. Take 1μL to do Qubit.

Take 1pmol to process step 10, and keep the rest sample in -20℃ freezer, it will be good for at least 3 months

Can stop here.

10. Rolling Circle Amplification

Assemble RCA reaction as follow:

	Volume (μL)	Final concentration
H_2O	77	
10X Reaction buffer	10	1x

Circularized DNA library	5 (~1pmol)	
1μM Right Hairpin RCA-Primer	1 (1pmol)	10nM
1mM AA-dUTP	1	10uM
25mM dNTPs	4	1mM each

Incubate at 95℃ for 1min, then 60℃ for 1min, 40℃ for 1min, 4℃ pause

Add following on ice:

- Phi29 DNA polymerase (10U/μL) 1μL 10U

Incubate at 30℃ for 6hours, then 4℃ forever.

Add 400μL of 1xPBS, and then store at 4℃ until needed. The Rolony can be stable and good for at least 3 months.

Before loading Rolony onto the surface, make additional 1:10 dilution using 1xPBS.

APPENDIX E

Paired-end Rolony library preparation

1. Fragmentation lamda DNA

Reaction Components	Starting amount
Lamda DNA(0.5ug/uL)	10uL
10x Fragmentase reaction buffer	20uL
100x BSA	2uL
MB Water	148uL
Final Volume	200uL

a. Vortex thoroughly.

b. Incubate on ice for 5min

c. Vortex NEBNext dsDNA fragmentase and add 20uL to the reaction

d. Vortex thoroughly

e. Incubate at 37℃ for 1hour.

f. Add 10uL of 0.5M EDTA to stop the reaction.

g. Mini Elute PCR purification kit (50) (Qiagen) clean.

1. Mix the reaction with 1mL PB, load on the column and spin for 1min. Toss the solution.

2. Add 750uL PE into the column, spin 1min. Toss the solution, spin 2min.

3. Change a tube and add 30uL EB to the center of the column, RT for 5min, spin 2min.

2. Gel purification and selection

a. Add 10uL DNA loading buffer into 30uL DNA

b. Load 5ul of 25bp DNA ladder on a 6% PAGE

c. Load 20uL of DNA on the gel (2 lanes)

d. Run 200mV for 25min

e. Stop gel running.

f. Place the gel into a box containing 5uL Cyber Green in 50uL 1x TBE, shake and incubate at RT for 10min.

g. Place the gel in the gel image machine.

h. Turn on the UV, manually exposure for 1.5 seconds

i. Freeze and print the image. Turn off the UV.

j. Under UV, cut out the bands (smear) as needed

k. Cut small pieces and put bands (smear) into 500uL PAGE Elution Buffer, incubate at 55℃ overnight.

3. Phenol purification

a. Spin down the gel, and add the supernatant solution into 500uL phenol mix (phenol: chloroform: isoamylalcohol = 25: 24: 1). Vortex and spin at top speed for 10min.

b. Collect the supernatant (~500uL), add 900uL 100% EtOH, 2uL Glycogen (20ug/uL) and 60uL 3M NaOAc (pH=5.2)

c. Vortex and store at -20℃ for 20min.

d. Spin for 15min at 4℃. Toss the solution, keep the DNA pellet.

e. Add 1mL 70% EtOH to the pellet. Vortex and spin down for 10min. Toss the solution, air dry the pellet.

f. Add 50uL EB to DNA pellet. (Ready for blunt-end repair)

4. Blunt-end repair (100ul total volume)

10uL 10x blunt-end buffer

50uL DNA in EB

1uL dNTPs (1mM)

2uL Enzyme

37uL H_2O

Incubate at 25℃ for 30min

Qiagen clean, suspend into 30uL EB.

5. A-tailed (100ul total volume)

30uL DNA in EB

10uL 10x Standard buffer

1uL dATP (100mM)

2uL Tag DNA polymerase

57uL H2O

Incubate at 70℃ for 30min

Qiagen clean, and suspend into 30uL EB. NanoDrop

6. Adaptor ligation (100ul total volume)

30uL DNA in EB

~ Adaptors (100uM) (Left Hairpin and Right Hairpin) (DNA: Adaptor = 1:10)

50uL 2x QLB

4uL T4 DNA ligase

~ H2O

Incubate at 25℃ for 30min

Qiagen clean, suspend into 50uL EB.

7. **Beads selection**

    a. Add 10uM Left Biosg Select Primer (in 50uL 6xSSPE) in 50uL DNA (in EB)

    b. Incubate at 70℃ for 5min; 50℃ for 5min; 25℃ for 5min.

    c. Take 200uL Myone beads; wash with B&W 2 times.

    d. Add 100uL DNA to the beads, add 400uL B&W.

    e. Incubate at RT for 30min with rotisserie.

    f. Wash with B&W 2times.

    g. Add 100uL 0.1N NaOH, and incubate at RT for 10min.

    h. Collect the NaOH solution, add 100uL Tris-HCl (1M, pH=7.5)

    i. Add 900uL 100% EtOH, 2uL Glycogen (20ug/uL) and 60uL 3M NaOAc (pH=5.2)

    j. Vortex and store at -20℃ for 20min.

    k. Spin for 15min at 4℃. Toss the solution, keep the DNA pellet.

    l. Add 1mL 70% EtOH to the pellet. Vortex and spin down for 10min. Toss the solution, air dry the pellet.

    m. Add 20uL EB to DNA pellet. Nanodrop. Name as Hairpin DNA

Library(HDL)

8. Rolling Circle Amplification

Assemble RCA reaction as follow:

	Volume (µL)	Final concentration
H_2O	81	
10X Reaction buffer	10	1x
HDL(~4ng/µL)	1	1pmole
1µM Right Hairpin RCA-Primer	1	1pmole
1mM AA-dUTP	1	10uM
25mM dNTPs	4	1mM each

Incubate at 95℃ for 1min, then 60℃ for 1min, 50℃ for 1min, 40℃ for 1min, 4℃ pause

Add following on ice:

✧ Phi29 DNA polymerase (10U/µL) 2µL 20U

Incubate at 30℃ for 6hours, then 4℃ forever.

Add 300µL of PBS, and then store at 4℃ until needed.

APPENDIX F

Rolony sequencing protocol on Polonator

2. Amino-silane coding flowcell

   2.1 Thoroughly wash and dry the flowcell to be coated

   - Wash flowcell with dH_2O
   - Wash flowcell with 100% isopropanol
   - Wash flowcell with dH_2O
   - Wash flowcell with acetone
   - Air dries the flowcell on a vacuum.

   2.2 Coat flowcell with Amino-silane (Perform acetone steps in a fume hood)

   - Prepare 1mL 2% solution of 3-Aminopropyltriethoxysilane in dH_2O
   - Use syringe and tube to suck 3-Aminopropyltriethoxysilane solution into flowcell, incubate at RT for 5min
   - Suck 2mL dH_2O through flowcell to wash
   - Suck 2mL acetone through flowcell to wash
   - Air dries the flowcell on a vacuum.

3. **Bind Rolony on flowcell**

   3.1 Load 35µL Rolony (in 1xPBS solution) into each lane of the flow cell. Incubate at room temperature for 2hr.

   3.2 Wash flowcell with 5mL 1xPBS

3.3 Load 1/7000 BS3 (1 tube no weight BS^3 (2mg) in 700uL 1xPBS, aliquot 100uL to 7 tubes, -20centdegree store. Every time, take 1 tube, add 900 1xPBS just before use), RT incubate for 1hour.

- Wash with 5mL 1xPBS.
- Wash with 5mL 5xSSC+0.01%Tween 20(Wash 2).

Load the flowcell on the Polonator, wash with Wash 2.

4. Hybridization

- Prepare and fill flow cell with the primer solution:

5xSSC+0.01%Tween20	497.5µL
Anchor primer (1mM)	2.5µL

Incubate at 50℃ for 5min, then room temperature for 5min.

- Wash 2 times with Wash 2
- The flowcell is ready for sequencing (SBL or SBS)

APPENDIX G

Rolony Sequencing Protocol

5. Oligonucleotides

Rolony_test1

5- /5Phos/ATG ACA TGA CTA CAG CCG TAN NNN NNN NNN NNN NNN NNN NGA TCA TAG GCATGA GTC AGT -3’

RCA-Primer1

5’-CGG CTG TAG TCA TGT-3’

Rolony-Minus anchor primer1

5- /5Phos/GAT CAT AGG CAT GAG T-3’

Rolony-Plus anchor primer1

5’-ATG ACA TGA CTA CAG CCG TA-3

Rolony-Minus anchor primer1-2bp

5- /5Phos/TCA TAG GCA TGA GTC AGT -3’

7-Mer-GA1

5’-NNN NNG A-3’

6. Circularization of template(COT1)

6.1 Assemble reaction as follow:

	Volume(µL)	Final Concentration
H_2O	11	
10x reaction buffer	2	1x

10µM Rolony_test1 (linear DNA)	1	10pmole
50mM $MnCl_2$	1	2.5mM
5M Betaine	4	1M
CircLigaseII	1	5U/µL (Epicenter CL9021K)

Incubate at 60℃ for 60min

Heat inactivate at 80℃ for 10min

6.2 Digest residual linear ssDNA

Add directly to the Circligase reaction solution:

- ExonucleaseI (20U/µL) 1µL
- ExonucleaseIII (100U/µL) 1µL

Incubate at 37℃ for 45min

Heat inactivate at 80℃ for 15min

6.3 Purify circularized template using Qiagen Mini Elute Purification PCR kit

- Add 5-fold PB to reaction solution, load it into Qiagen column. Spin 1min at top speed, and then toss the liquid.
- Add 750µL PE to the column; spin 2min at top speed. Dump the liquid, empty spin 2min at top speed.
- Use a new tube to collect the COT product. Add 30µL EB to the column; Spin 2min at top speed.
- Take 1µL to do Nanodrop, using EB to blank.

7. Rolling circle amplification (RCA)

7.1 Assemble RCA reaction as follow:

	Volume (µL)	Final concentration
H_2O	82	
10X Reaction buffer	10	1x
COT(3.5ng/µL)	1	1pmole
1µM RCA-Primer	1	1pmole
1mM AA-dUTP	1	10uM
25mM dNTPs	4	1mM each

Incubate at 95℃ for 1min, then 60℃ for 1min, 50℃ for 1min, 40℃ for 1min, 4℃ pause

Add following on ice:

- ✧ Phi29 DNA polymerase (10U/µL) 1µL 10U

Incubate at 30℃ for 6hours, then 4℃ forever.

7.2 Add 300µL of PBS, and then store at 4℃ until needed.

8. Amino-silane coding slide

8.1 Thoroughly wash and dry the glass surface to be coated

- ➢ Rinse microscope slide with dH_2O
- ➢ Rinse microscope slide with isopropanol
- ➢ Rinse microscope slide with dH_2O

- Rinse microscope slide with acetone
- Air dries the slide.

8.2 Coat slide with Amino-silane (Perform acetone steps in a fume hood)

- Prepare a 2% solution of 3-Aminopropyltriethoxysilane in dH_2O
- Immerse surface in the diluted reagent for 5min
- Rinse slide with dH_2O
- Rinse slide with acetone
- Air dries the slide.

9. Bind Rolony to amino glass slide

9.1 Make a flow cell using HybriWell^{TM} (GRACE Bio-Labs, Inc.)

9.2 Load 100µL Rolony on the flow cell. Incubate at room temperature for 2hr.

9.3 Wash channel with Wash 2 as follow:

Fill flow cell with Wash 2 (5x SSC+0.01%Tween20) by capillary action:

- Tile the slide at 45 degree
- Add a drop liquid to the inlet of the chamber, the capillary action should displace the drop and accumulation of wash should be seen at the outlet of the chamber, avoid bubble formation.
- Repeat wash 3 times.
- Wash with 1xPBS 3 times.
- Load 1/7000 BS3 (1 tube no weight BS in 7000uL 1xPBS), RT incubate for 1hour.
- Wash with PBS 3 times.

- Wash with Wash2 3 times.

10. Sequencing by Ligation (minus 1-5 base)

10.1 Hybridization

- Prepare and fill flow cell with the primer solution:

Wash2	200µL
Anchor primer (1mM)	1µL

Incubate at 50℃ for 5min, then room temperature for 5min.

- Wash chamber 3 times with Wash 2

10.2 Ligation

- Add the following ligation mixture to the flow cell(200µL):

◆ H_2O	94µL
◆ 2x Quick ligation buffer	100µL
◆ Fluorescent nonamer mix	2µL
◆ T4 ligase (600U/µL, Enzymatics)	4µL

- Incubate at 37℃ for 5min, then room temperature for 5min.
- Wash chamber 3 times with Wash 2
- Wash chamber with 1E (1E is better than TE for FAM signal)

10.3 Stripping

- Add 200µL of 65% Formamide (in TE) into the chamber, incubate 5min.

- ➢ Wash with 65% Formamide 3 times
- ➢ Wash 3 times with Wash 2

11. Sequencing minus 6-10 base

11.1 Hybridizing Rolony-minus Anchor Primer-2 (same as 6.1)

11.2 Ligate 7-mer-GA

- ➢ Add 6-fold 7-mer-GA(30µM) in 200µL of 1x quick ligase buffer
- ➢ Incubate at 37℃ for 7min, then room temperature at 7min, and then repeat 37℃ and RT for 7min.
- ➢ Wash the chamber with 1E.

11.3 5'Phosphorylation

- ➢ Add 200µL phosphorylation mixture (4µL T4 polynucletide kinase(PNK) in 196µL 1x PNK buffer)
- ➢ Incubate at 37℃ for 30min
- ➢ Wash 3 times with 1E

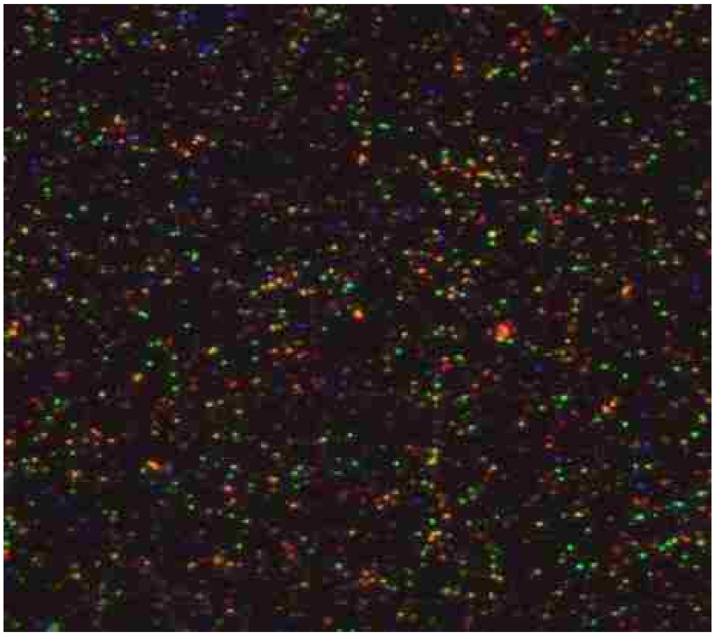
11.4 Ligation (Same as 6.2)

8. Sequencing by Synthesis (Plus direction) (FfAME reagents: all reagent treated with 0.5%NH_2OH, except deprotect and cleavage buffer)

- ➢ Hybridizing plus anchor primer (same as 6.1)

- Add 200µL SBS fluorescent master mix and 4µL Therminator I DNA polymerase into chamber
- Incubate at 50℃ for 10min
- Wash 3 times with 1E.

9. **Image with Microscope**



APPENDIX H

Hairpin strand-displacement test

1. Mix 5uL Hairpin strand-displacement test (10uM);

   5uL Hairpin strand-displacement AP2 (10uM);

   20uL 1xTE buffer.

   Mix 5uL Hairpin strand-displacement test (100uM);

   5uL Hairpin strand-displacement AP2 (100uM);

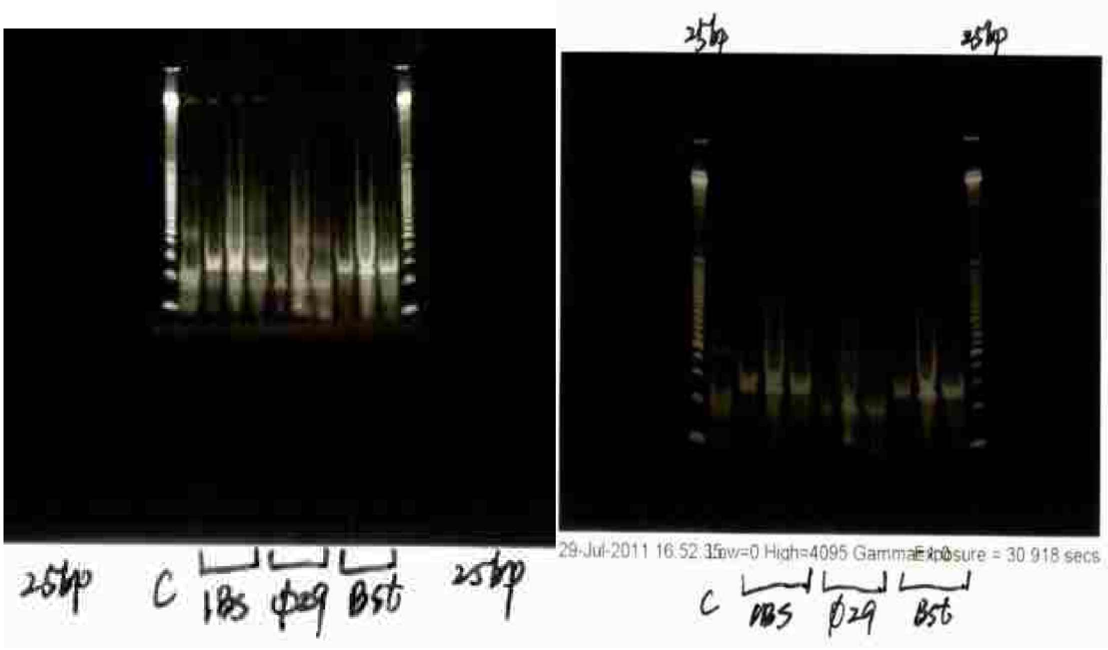
   30uL 1xTE buffer.

   70℃ for 5min; 50℃ for 5min; 25℃ for 5min;

2.

Control(100uM)	Bst	Phi 29	IBS N-9℃
10uL Mix(step 1)	10uL Mix(step 1)	10uL Mix(step 1)	10uL Mix(step 1)
2uL dNTPs(25mM)	2uL dNTPs(25mM)	2uL dNTPs(25mM)	2uL dNTPs(25mM)
0	2uL Bst	1uL Phi29	2uL IBS N-9℃
86uL 1x Thermo Buffer	86uL 1x Thermo Buffer	87uL 1x Phi29 Buffer	86uL A1 IBS bffuer
55℃ for 30min	55℃ for 30min	30℃ for 30min	55℃ for 30min

3. Phenol clean, then run **control, 10uM, 100uM (1/2, 1/4)** on a 6%PAGE gel.


25bp
C
IBS
ϕ29
Bst
25bp
25bp
25bp
29-Jul-2011 16:52:15 Low=0 High=4095 Gamma=1.00 Exposure = 30.918 secs
C
IBS
ϕ29
Bst

25ladder	control	IBS_10	IBS_100/2	IBS_100/4	P29_10	P29_100/2	P29_100/4	Bst_10	Bst_100/2	Bst_100/4	25ladder

APPENDIX I

Ion TargetSeq Exome Enrichment for the Ion Proton System

1. **Prepare adapter-compatible DNA**

i) Purifiy the gDNA with AMPure XP reagent

(1) Diluted 1ug of concentrated gDNA to a final volume of 50ul (20ng/ul) with nuclease-free H2O.

(2) Add 90ul of AMPure XP reagent, pipet up and down 5 times, quick spin and incubate at RM for 5min

(3) Place the tube on a MPC for 3min, carefully remove and discard the supernatant without disturbing the bead pellet.

(4) Without removing the tube from MPC, add 500ul of freshly prepared 70% EtOH, incubate 30sec, turning the tube around twice in the magnet to move the beads around. After the solution clears, remove and discard the supernatant without disturbing the pellet.

(5) Repeat step (4) for a second wash

(6) Use a 20-ul pipettor to remove residual ethanol

(7) Keep the tube on MPC, air-dry the beads at RT for less than 5min.

(8) Remove tube from MPC, add 35ul of nuclease-free water to the beads. Pipet up and down 5 times, then vortex for 10sec.

(9) Quick spin and proceed directly to **Fragment the DNA and purification**

ii) Fragment the DNA and purify

(1) Purified gDNA with beads	35ul
(2) Ion Shear Plus 10X reaction buffer	5ul
(3) Ion Shear Plus Enzyme Mix II	10ul

(4) Using a p100-P200 pipettor set at a 40-ul volume, mix the raction by rapidly pipetting up and down 8-10 times. Do not mix by vortexing.

(5) Incubate the reaction on a thermo cycler at 37 cent-degree for 40min.

(6) Add 5ul of Ion Shear Stop buffer after incubation and vortex for 5sec.

(7) Purify the fragmented DNA

- 99ul of AMPure XP reagent
- 500ul of 70% fresh EtOH wash 2 times
- Air-dry for 5min, add 25ul low TE and collect DNA for ligation. (the DNA should be 70-400bp)

2. Ligate adapters, nick repaire, and purify the ligated DNA

i)	DNA	~25 ul
ii)	10x Ligase buffer	10ul
iii)	Adapters	10ul
iv)	dNTP Mix	2ul
v)	Nuclease-free Water	41ul
vi)	DNA Ligase	4ul
vii)	Nick Repair Polymerase	8ul
viii)	Total	100ul

ix) 25 cent degree for 15min

x) 72 cent degree for 20min

xi) 4 cent degree hold

xii) Purify the adapter-ligated and nick-translated DNA

(a) 180ul of AMPure XP reagent

(b) 500ul of fresh 70% EtOH wash 2 times

(c) Air-dry and add 20ul low TE and collect DNA.

3. **Size-select the unamplified library**

    a. E-Gel SizeSelect 2% Agarose Gels

    b. Target size ~220bp.

    c. Load DNA in 2 lanes

    d. Collect 60ul DNA product.

    e. Purify the size-selected DNA

    - Add 108ul AMPure reagent
    - 2X 500ul of fresh 70% EtOH wash
    - Air-dry for 4min, and add 30ul Low TE
    - Transfer DNA supernatant to amplification step.

4. Amplify and purify the size-selected fragment library

i)	Platinum PCR SuperMix High Fidelity	200ul
	Library Amplification Primer Mix	20ul
	Purified, size-selected library DNA	30ul

ii) Denature at 95 degree for 5min

8-10 cycles of 95---15sec, 58---15sec, and 70---1min

Hold at 4 degree.

iii) 375ul of AMPure reagent

2x 500ul of fresh 70% EtOH wash

Air-dry less than 5min, add 50ul of Low TE

iv) Transfer DNA supernatant to a new 1.5ml tube. **STOP Point.**

Aliquot 500ng of DNA to a new tube for Exome Enrichment step.

5. Quantitate and qualify the amplified library

   Qubit and Bioanalyzer.

6. Exome Enrichment

   See the manufacture’s instruction.

7. Prepare the exome-enriched library for template preparation

APPENDIX J

Illumina Nextera DNA sample prep kit

1. **Tagmentation.**

20ul of WGA DNA at 2.5ng/ul (50ng total);

25ul of TD buffer

5ul of TDE1

Pipette up and down 10times to mix;

Quick pin 280xg at 20centdegree for 1min;

55 cent degree for 5min;

Hold at 10degree.

2. **Clean-up of tagmented DNA**

Zymo DNA clean-up	Qiagen Mini elute PCR column
180ul of Zymo DNA binding buffer	250ul of PB
1,300xg at RT for 2min	13.7x13,700 rpm at RT for 1min
300ul of Zymo wash buffer, 1,300xg at	750ul of PE wash, 13.7x13,700 rpm at

RT for 2min	RT for 1min
300ul of Zymo wash buffer, 1,300xg at RT for 2min	
1,300xg at RT for 2min	13.7x13,700 rpm at RT for 2min
25ul of RSB, 2min at RT	25ul of RSB, 2min at RT
1,300xg at RT for 2min	13.7x13,700 rpm at RT for 2min

Load 1ul of undiluted Zymo eluate on a HS Bioanalyzer chip, and this should produce a broad distrubtion of DNA fragments with a size range from ~150bp --- <1kb.

3. **PCR to insert index 1 (i7) and index 2 (i5), and P5 and P7 for cluster generation and sequencing.**

5ul of index 2 primers (white caps)

5ul of index 1 primers (orange caps)

15ul of NPM (Nextera PCR Master Mix)

5ul of PPC (PCR Primer Cocktail)

20ul of purified tagmented DNA

Pipette up and down 3-5times

Quick spin at 280xg at RT for 1min

Perform PCR using the following program on a thermal cycler:

- 72 degree for 3min;
- 98 degree for 30sec;
- 5cycles of
    - 98 degree for 10sec;
    - 63 degree for 30sec;
    - 72 degree for 3min;
- Hold at 10 degree

The PCR tubes can be remained on the thermal cycler overnight, or be stored at 4 degree up to 2 days.

4. PCR clean-up

- Transfer 50ul of post-PCR mix to a 1.5ml non-stick tube;
- Add 30ul of AMPure XP beads (for 2x250 runs on the MiSeq, add 25ul of AMPure XP beads), mix up and down 10times.
- Incubate at RT without shaking for 5min;
- Place the tubes on a MPC for 2min or until the supernatant has cleared.
- Remove and discard the supernatant carefully without touching the beads.
- With the tube on MPC, wash beads with freshly prepared 80% Ethanol as follows:
    - Add 200ul of fresh 80% Ethanol;

    - Incubate the tube on MPC for 30sec or until the supernatant appears clear;
    - Carefully remove and discard the supernatant.

- With the tube on MPC, perform a second ethanol wash was follows:
    - Add 200ul of fresh 80% Ethanol;
    - Incubate the tube on MPC for 30sec or until the supernatant appears clear;
    - Carefully remove and discard the supernatant;
    - Use a P20 pipette with fine pipette tips to remove excess ethanol

- With the tube on MPC, allow the beads to air-dry for 15min;
- Remove the tube from MPC, add 32.5ul of RSB to each tube;
- Gently pipette mix up and down 10 times, and incubate at RT for 2min;
- Place the tube on MPC for 2min or until the supernatant appears clear.
- Carefully transfer 30ul of the supernatant from the tube to library collecting tubes.

5. Validate library

- Quantify libraries using Qubit.

  Based on an average size of 500bp, 1ng/ul =3nM

- Quality control

  Run 1ul of 1:3 diluted library on an Agilent Technologies 2100 Bioanalyzer using a High Sensitivity DNA chip. ~250-1000bp

    - For modified PCR clean-up step for 2x250 runs on MiSeq, 1ng/ul = 1.5nM.

6. **Pooling libraries**

    - Normalizing 10ul of sample library to 2nM final concentration using Tris-HCl 10mM, pH 8.5 with 0.1% tween 20.
    - Transfer 5ul of each sample to a 1.5ml non-stick tube, mix well, and store at -20 degree.

APPENDIX K

Truseq DNA sample preparation protocol

Fragment DNA

JSE genome DNA (Qubit read, 50.8 ng/uL)

Three reactions of following mix

JSE Genome DNA	20 uL
Ion Shear Plus 10X Reaction Buffer	5 uL
Nucleate free water	15 uL
	40 uL

Add 10 uL of enzyme (Ion Shear Plus Enzyme Mix II) and start reaction at 37C for 15min on PCR machine

Stop reactions by adding 5 uL of Ion Shear Stop Buffer.

Purified with AmPure XP II beads

Add 100ul AMPure beads, mix well and incubate at RT for 15min

200ul of 70% EtOH wash 2x.

RT air-dry for 15min

Eluted with 17.5 uL of RSB

15 uL of elution was transferred into new PCR tubes.

Adenylate 3' Ends (A tailing)

End repaired DNA	15 uL
RSB	2.5 uL
A-Tailing Mix	12.5 uL
	30 uL

Incubated at 37 C for 30 min

Ligate Adapter

A-tailed DNA	30 uL
RSB	2.5 uL
Ligation Mix	2.5 uL
Adapter Index 1	2.5 uL
	37.5 uL

Sequence of Adapter Index1 (AD001) = ATCACG(A)

Incubated at 30 C for 10 min, add 5ul of Stop Ligation Mix

Purified with AmPure XP II beads (two times)

42.5ul of AMPure beads, mix well and incubate at RT for 15min

200ul of 80% EtOH wash 2x.

RT air-dry for 15min

Add 52.5ul of RSB, transfer 50ul of supernatant. Re-do AMPure beads cleanup.

Eluted with 22.5 uL of RSB

20 uL of elution was transferred into new PCR tubes.

Purification by DNA E-gel

Ran DNA gel

2% size select Agarose gel on E-Gel system.

Cut out 800-900bp product

Eluted with 25 uL of EB buffer.

20 uL of elution was transferred into new PCR tubes.

PCR

gel purified DNA	20 uL
PCR Primer Mix	5 uL
PCR Master Mix	25 uL
	50 uL

98C, 30 sec

98C, 10 sec

60C, 30 sec X 6 cycles

72C, 30 sec

72C, 5min

10C, Hold

Purified with AmPure XP II beads

Add 50ul AMPure beads, mix well and incubate at RT for 15min

200ul of 80% EtOH wash 2x.

RT air-dry for 15min

Eluted with 32.5 uL of RSB

20 uL of elution was transferred into new PCR tubes.

Qubit reads:

700-800bp: ______ng/ul (20ul)

Run 30ng of product on E-Gel to check the DNA purity and size.

700-800bp: one smear ________bp

APPENDIX L

Nextera DNA sample prep kit Aliquot protocol 03/20/2013

1. N701: Aliquot 42.5ul of N701 in 1 strip PCR tube (8.5x reactions) X 2tube
2. N702: Aliquot 42.5ul of N702 in 1 strip PCR tube (8.5x reactions) X 2tube
3. N703: Aliquot 42.5ul of N703 in 1 strip PCR tube (8.5x reactions) X 2tube
4. N704: Aliquot 42.5ul of N704 in 1 strip PCR tube (8.5x reactions) X 1tube
5. N705: Aliquot 42.5ul of N705 in 1 strip PCR tube (8.5x reactions) X 1tube
6. N706: Aliquot 42.5ul of N706 in 1 strip PCR tube (8.5x reactions) X 1tube
7. N707: Aliquot 42.5ul of N707 in 1 strip PCR tube (8.5x reactions) X 1tube
8. N708: Aliquot 42.5ul of N708 in 1 strip PCR tube (8.5x reactions) X 1tube
9. N709: Aliquot 42.5ul of N709 in 1 strip PCR tube (8.5x reactions) X 1tube
10. N710: Aliquot 42.5ul of N710in 1 strip PCR tube (8.5x reactions) X 1tube
11. N711: Aliquot 42.5ul of N711 in 1 strip PCR tube (8.5x reactions) X 1tube
12. N712: Aliquot 42.5ul of N712 in 1 strip PCR tube (8.5x reactions) X 1tube

13. N501: Aliquot 12.5ul of N501 in 1 strip PCR tube (2.5x reactions) X 8tube
14. N502: Aliquot 12.5ul of N502 in 1 strip PCR tube (2.5x reactions) X 8tube
15. N503: Aliquot 12.5ul of N503 in 1 strip PCR tube (2.5x reactions) X 8tube
16. N504: Aliquot 12.5ul of N504 in 1 strip PCR tube (2.5x reactions) X 8tube
17. N505: Aliquot 12.5ul of N505 in 1 strip PCR tube (2.5x reactions) X 8tube
18. N506: Aliquot 12.5ul of N506in 1 strip PCR tube (2.5x reactions) X 8tube
19. N507: Aliquot 12.5ul of N507in 1 strip PCR tube (2.5x reactions) X8tube
20. N508: Aliquot 12.5ul of N508 in 1 strip PCR tube (2.5x reactions) X 8tube

21. TD buffer: Aliquot 412.5ul of TD in 1.5ml tube (16.5x reactions) X 8tube

22. TDE1: Aliquot 82.5ul of TDE1 in 1 strip PCR tube (16.5x reactions)X 8tube

23. RSB: Aliquot 1000ul of RSB in 1.5ml tube (16.5x reactions) X 8tube

24. NPM: Aliquot 247.5ul of NPM in 1.5ml tube (16.5x reactions) X 8tube

25. PPC: Aliquot 82.5ul of PPC in 1 strip PCR tube (16.5x reactions) X 8tube

This aliquot is enough for 128 libraries. Store the aliquots at -20degree, and keep the rest of reagents in the original kit box, store at -20 degree.