Spring 5-1-2015

Novel Bioinformatic Approaches for Analyzing Next-Generation Sequencing Data

Yan Peng
University of Southern Mississippi

The University of Southern Mississippi

NOVEL BIOINFORMATIC APPROACHES FOR ANALYZING NEXT-GENERATION SEQUENCING DATA

by

Yan Peng

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

May 2015

ABSTRACT

NOVEL BIOINFORMATIC APPROACHES FOR ANALYZING

NEXT-GENERATION SEQUENCING DATA

by Yan Peng

May 2015

In general, DNA reconstruction is deemed as the key of molecular biology since it makes people realize how genotype affects phenotypes. The DNA sequencing technology emerged exactly towards this and has greatly promoted molecular biology's development. The traditional method, "Sanger," is effective but extremely expensive on a cost-per-base basis. This shortcoming of Sanger method leads to the rapid development of next-generation sequencing technologies. The NGS technologies are widely used by virtue of their low-cost, high-throughput, and fast nature. However, they still face major drawbacks such as huge amounts of data as well as relatively short read length compared with traditional methods. The scope of the research mainly focuses upon a quick preliminary analysis of NGS data, identification of genome-wide structural variations (SVs), and microRNA prediction. In terms of preliminary NGS data analysis, the author developed a toolkit named "SeqAssist" to evaluate genomic library coverage and estimate the redundancy between different sequencing runs. Regarding the genome-wide SV detection, a one-stop pipeline was proposed to identify SVs, which integrates the components of preprocessing, alignment, SV detection, breakpoints revision, and annotation. This pipeline not only detects SVs at the individual sample level, but also identifies consensus SVs

at the population and cross-population levels. At last, miRDisc, a pipeline for microRNA discovery, was developed for the identification of three categories of miRNAs, i.e., known, conserved, and novel microRNAs.

The University of Southern Mississippi

NOVEL BIOINFORMATIC APPROACHES FOR ANALYZING NEXT-GENERATION SEQUENCING DATA

by

Yan Peng

A Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:

_Dr. Nan Wang_____
Committee Chair

_Dr. Chaoyang Zhang_____

_Dr. Ping Gong_____

_Dr. Chenhua Zhang_____

_Dr. Zheng Wang_____

_Dr. Karen S. Coats_____
Dean of the Graduate School

May 2015

ACKNOWLEDGMENTS

This is a great opportunity to express the immense gratitude to Dr. Nan Wang, Dr. Ping Gong, and all the other committee including, Dr. Chaoyang Zhang, Dr. Chenhua Zhang, and Dr. Zheng Wang, for their invaluable advice on my research as well as the dissertation. I gratefully acknowledge my co-advisors, Drs. Nan Wang and Ping Gong, for their generous help and continuous support in the past four years. Throughout my Ph.D program, they have been kindly, offering me her guidance, suggestions, and support.

Special thanks that words cannot carry to my parents for giving birth to me in the first place and for their unconditional support throughout my life. Also, I would like to express much appreciation to my beloved husband, Dr. Jianan Wang, who has always been my support through the good times and bad.

Thanks also go to all my co-workers at the Computational Biology and Bioinformatics Lab (CBBL), especially Lijuan Yang and Andrew Maxwell, for their help throughout my four year study.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

CHAPTER I

INTRODUCTION

DNA Sequencing Technologies

Deoxyribonucleic acid (DNA) is the carrier of genetic materials for all living organisms and many viruses. It is the most essential component of chromosomes and plays an important role in developing and functioning organisms. It consists of four kinds of nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The important role of DNA leads people to explore and research on DNA. This gives rise to the rapid development of DNA sequencing technology. DNA sequencing is a process for determining the exact type and order of nucleotides for a fragment of genome or the whole genome. Evolving from the traditional sequencing technology, sanger method (Sanger & Coulson, 1975) to the currently widely used next-generation technologies (NGS) (Metzker, 2010; Mardis, 2013) and the next next-generation sequencing technologies (next-NGS), DNA sequencing technologies are rapidly developing and moving towards to the direction with low-cost, high-speed and high-accuracy.

Traditional Approach

Sanger sequencing, the earliest or the first generation sequencing technology, was invented by Sanger, Nicklen, and Coulson (1977) (Sanger, Nicklen, & Coulson, 1977). The basic principle is: polyacrylamide gel electrophoresis can distinguish the single-stranded DNA molecules with only one base difference. Materials used in the first generation sequencing experiments are homogeneous single-stranded DNA molecules, called the template DNA. The

first step is to anneal the short oligonucleotide molecule in the same position on each template strand. The short oligonucleotide molecule is then treated as oligonucleotide primers to synthesize of new DNA stand, which is complementary to the template DNA. After sequencing, primers bind with single-stranded DNA template molecule, and DNA polymerase extends the primers with deoxynucleosidetriphosphates (dNTP). Extension reaction proceeds into four groups, and each group uses one of the four standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) to terminate the process. Then PAGE analysis is applied, and the desired sequence can be read from the resulting PAGE gel. Figure 1 shows the whole procedure of Sanger sequencing.

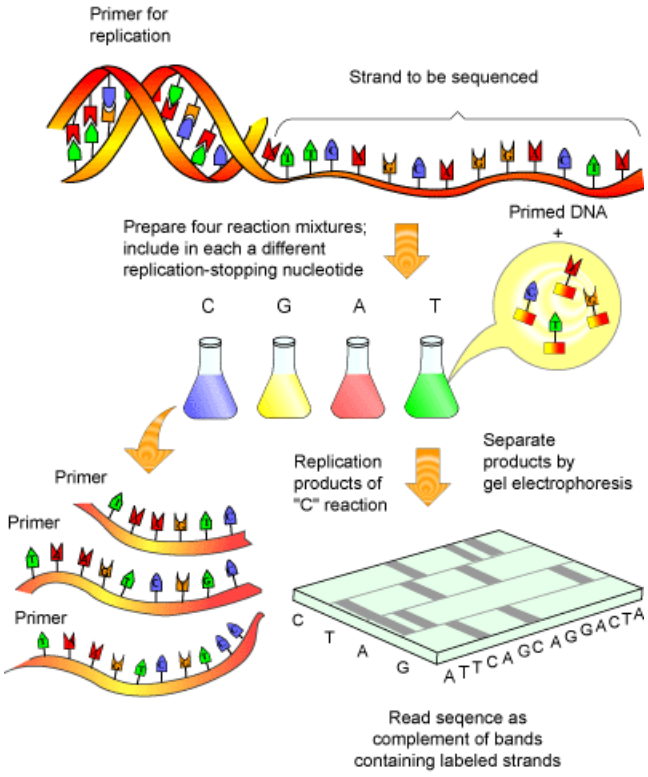


Figure 1. Procedure of Sanger Sequencing. http://www.eisenlab.org/FunFly/wp-content/uploads/2012/06/science-creative-quarterly-seq.gif

Next-generation Sequencing (NGS)

The Emergence of a massively parallel sequencing platform not only decreases the cost of DNA sequencing dramatically, but also allows many researchers able to sequence genomes, which was the privilege of the large DNA sequencing center before. Next-generation sequencing technology helps people with more comprehensive and in-depth analysis of genome, transcriptome, and protein interactions among various groups of data in relatively low cost. There are a number of next-generation sequencing products on the market, such as 454 (Margulies et al., 2005) genome sequencer produced by Roche Applied Science company, Illumine sequencing machine developed by illumine company in United States and Solexa technology company in United Kingdom, and SOLiD (htt) sequencing machine from Applied Biosystems company, etc. The basic principle for Illumine/Solexa Genome Analyzer sequencing is sequencing by synthesis. Based on Sanger sequencing technology, next-generation sequencing uses four different colors of fluorescent to label four types of dNTP, dATP, dCTP, dGTP, and dTDP. When the DNA polymerase synthesizes the complementary chain, the addition of different dNTPs will result in different fluorescence. The testing DNA sequence can be obtained by capturing the fluorescence signal through a specific software. Figure 2 shows the general flow for Illumina sequencing: (1) library preparation: DNA sequence is cut into fragments with several hundred nucleotides or less by ultrasonic wave or atomizing machine. DNA fragments are cut into blunt ends using polymerase and exonuclease, followed by the addition of a sticky

nucleotide to the end. Then, DNA fragments are ligated with adaptors. (2) cluster generation: Template molecules are put into chips for generating cloning clusters and sequencing of cluster cycle.

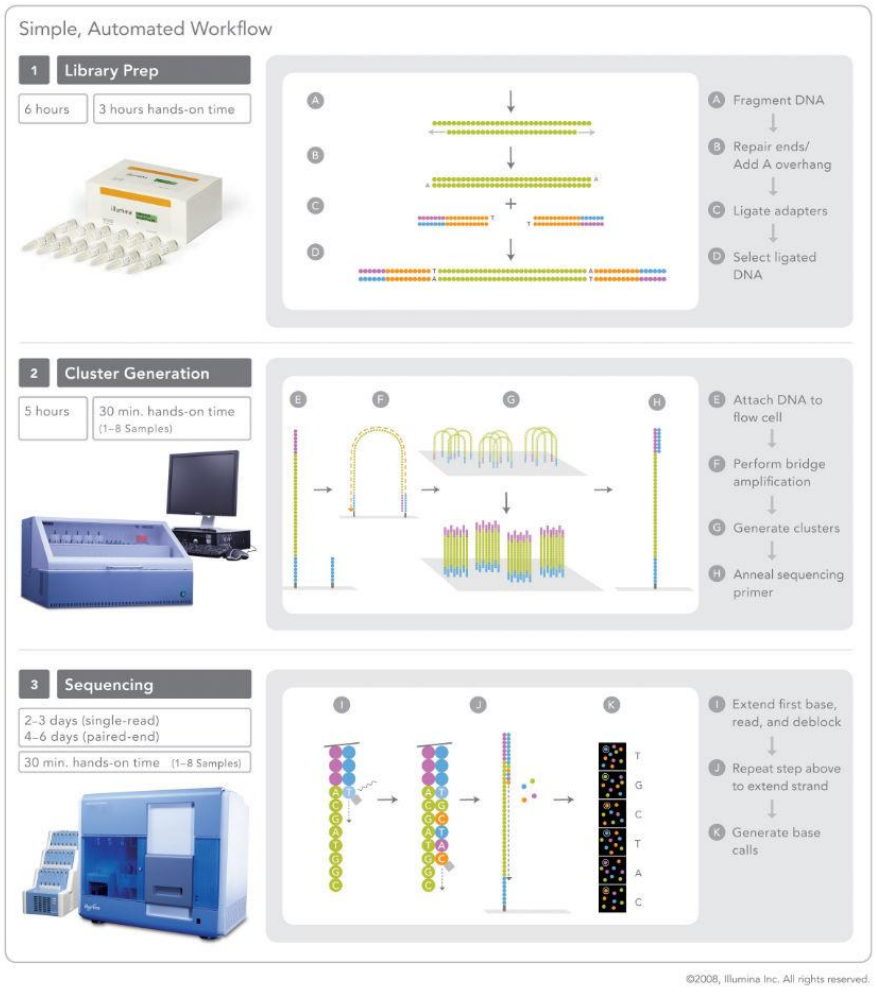


Figure 2. Illumina Genome Analyzer workflow. https://ccrod.cancer.gov/confluence/download/attachments/35947398/SimpleAutomatedWorkflow1.jpg?version=1&modificationDate=1239738295510&api=v2

Each chip has eight longitudinal silicon lanes. The inner surface of each lane has numeric fix single-strand adaptors. DNA fragments with adaptors denature into single-strand DNA fragment and then form bridge-like structures by connecting to the primers in the sequencing channels. A huge amount of DNA testing

fragments can be obtained by repeating the above procedures. (3) sequencing: There are three parts in this step, DNA polymerase combining with fluorescent terminator, fluorescent label cluster imaging and cutting the combined nucleotide and decomposition before next cycle begins.

Next-NGS Sequencing

Helicos single molecule sequencing, known as the next next-generation sequencing technologies, is SMRT technology (Osherovich, 2010) from PacificBioscience (Eid et al., 2009) and single-molecule nanopore sequencing technology from Oxford Nanopore Technologies Company. This sequencing technology is in the direction of high-throughput, low cost, and long read length. Unlike next-generation sequencing technology depending on the combination of solid surface and DNA template and sequencing by synthesis, the next next-generation sequencing technology is for single molecule DNA sequencing, and it does not require the PCR amplification process (Mayer, Farinelli, & Kawashima, 2013; Williams et al., 2006). The principles for different technologies are quite different. The workflow of Helico BioScience single molecule sequencing technology shows as Figure 3. It is based on the idea of sequencing by synthesis in the next-generation sequencing technology. First, the DNA sequence is randomly cut into small fragments with less than 1000nt, optimally between 100nt and 200nt. Each fragment is added by poly(A) tail at the 3’ by terminal transferase, and the poly(A) tail is labeled with fluorescence and resistance. The labeled fragments with poly(A) tail are hybridized with small fragments with poly(T) in the glass slide. The location for each hybridized template is obtained

by an imaging procedure. Polymerase and deoxynucleotides labeled with Cys fluorescence are added to synthesize DNA. Only one type of deoxynucleotides is added for each time. After removing non-synthesized dNTP and DNA polymerase, a template locus is observed to check whether there is fluoresced signal by imaging Cys. Then, add another type of deoxynucleotide and polymerase to build the next reaction. Through repeating the above steps, the DNA can be sequenced one base by one base.

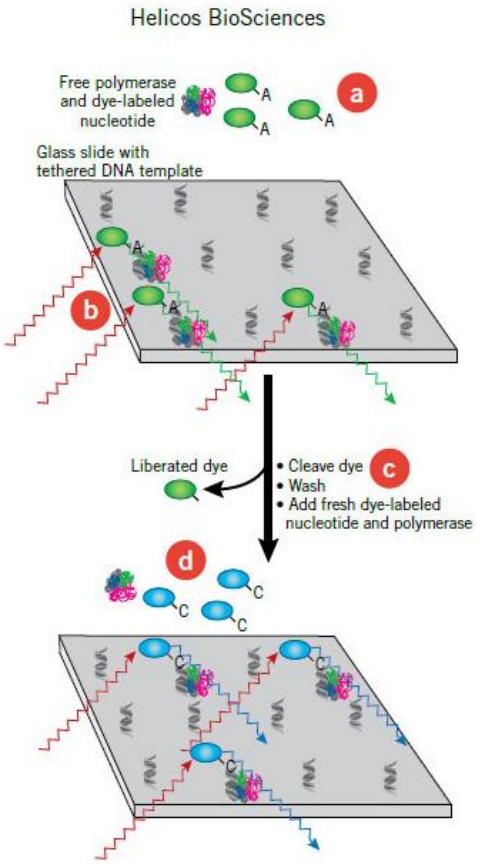


Figure 3. Helicos BioSciences workflow.
http://www.nature.com/scibx/journal/v3/n11/images/scibx.2010.331-F1.jpg

NGS Technologies Comparison

The first section of Chapter I introduces various technologies for sequencing DNA. Although the basic idea is similar for each technology, the methods are significantly different. Table 1 shows some statistics for three types of next-generation sequencing technology: Roche 454, Illumina GA, and AB SOLiD. In comparison with all three different methods, Roche 454 makes the longest read and also the fastest method with high accuracy; Illumina GA also has a wide range of read length from 50bp to 250bp and high throughput with median running speed, while AB SOLiD utilizes a shortest read.

Table 1

Comparison of Next-generation Sequencing Method (Liu et al., 2012)

Sequencing technology	454	Illumina	SOLiD
Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4
Read length	700bp	50-250bp	50+35 or 50+50bp
Accuracy	99.9%	98%	99.94%
Reads per run	1 million	up to 3 billion	1.2-1.4 billion
Time per run	24 hours	3~10 days	7 days for SE 14days for PE
Output data	0.7 Gb	600 Gb	120 Gb

length and lowest running time with highest accuracy. However, Roche 454 can only produce single-end read, while the other two, Illumina GA and AB SOLiD, can generate both single-end and paired-end read. In summary, Illumina GA is most widely applied due to its high throughput, low cost, and its capability of generating paired-end read with relatively high accuracy and speed.

NGS Sequence Analysis

NGS Sequence Read Type

In DNA sequencing technology, there are three types of reads: single-end reads, paired-end reads, and mate pair reads. Single-end reads are the result of sequencing one end of the fragments, while paired-end reads and mate pair reads obtain both ends of the DNA fragments while sequencing. The difference between paired-end and mate pair refers to how they make the sequencing library and how the DNA fragment is sequenced.

FASTA File Format

FASTA is a standard text-based format for sequencing. Each sequence contains two lines. The first line starts with a '>' character and is followed by the sequence

```
@sequence_id
GATTCCTGTAAGCTTAAAGCTCCATTGTACCCG
ATATACGCCTTT
```

identifier and/or description. The second line is the sequence containing A, C, G, T, or N (unknown base).

FASTQ File Format

Although the nucleotide is determined by collecting the fluorescence signal, the final sequence output is in another widely used file format called FASTQ (Cock et al., 2010). FASTQ format is a text-based file format. It contains all of the nucleotide sequences and its corresponding quality scores. Next follows an example of the FASTQ format.

```
@sequence_id
GATTCCTGTAAGCTTAAAGCTCCATTGTACCCG
ATATACGCCTTT
+
&??#55CCFF%%>>>>>6615%%+++***09@??=><
<=++@@AB
```

FASTQ format adopts four lines to represent a sequence. The first line starts with “@” character and is followed by the sequence identifier. The second line is nucleotide sequences letters. The third line begins with a “+” character and optional description. The fourth line is the quality score. Each score represents the quality of its corresponding base in the first line. Therefore, the number of qualities should be the same as the number of letters in the sequence. The quality score for the Illumina GA platform can be calculated by the following formula:

$$Q_{\text{solexa-prior to v.1.3}} = -10\log_{10}\frac{p}{1-p},$$

where p is the probability that the corresponding base is incorrect.

The quality score calculated by the above formula will be then encoded into a single ASCII character by some strategies: Phred+33 for Sanger (0, 40),

Solexa+64 for Solexa (-5, 40), Phred+64 for Illumina 1.3+ (0, 40), Phred+64 for Illumina 1.5+ (3, 40), and Phred+33 for Illumina 1.8+ (0, 41), etc.

NGS Sequence Assembly

Sequence assembly is to merge some short DNA sequence reads with certain overlapping bases into a longer DNA sequence in order to reconstruct the original structure of DNA. This process is vital because current sequencing technologies are unable to sequence the whole genome at one time. The whole genome needs to be cut into small fragments and then sequenced. There are two different types of assembly: de novo assembly and mapping assembly. De novo assembly is assembling short reads to create longer sequences, while mapping assembly is assembling reads to an existing backbone sequence template and then building a similar sequence as the backbone. A simple process of de novo assembly can be explained in Figure 4.

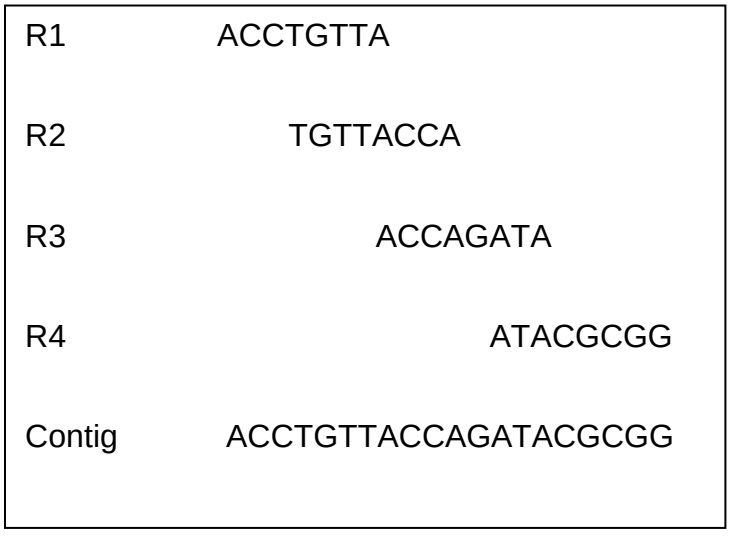


Figure 4. Example of de novo assembly.

R1, R2, R3, and R4 are four short sequence reads with overlapping. The longer sequence, namely “contig,” can be obtained by assembling these four reads.

The emergence of next-generation sequencing technology greatly promotes the development of sequence assembly technology (Miller, Koren, & Sutton, 2010; Li et al., 2012). There are a number of assembly tools that are free of charge: MIRA (Chevreus et al., 2004) is a general purpose assembler which can accept multiple platforms sequencing data and integrate them together. However, due to its speed limitation, MIRA is not suitable for assembling larger genomes. SOAPdenovo (Li et al., 2010) is an all-purpose genome assembler, which runs extremely fast using a medium amount of RAM and works well with short reads. Other free software include ABySS (Simpson et al., 2009), EULER (Chaisson, Brinza, & Pevzner, 2009), Ray (Boisvert, Laviolette, & Corbeil, 2010), and commercial software package, such as CLC and Newbler, etc.

NGS Sequence Alignment

Simply speaking, sequence alignment is to compare the similarity of two sequences. The theoretical basis of sequence alignment is Darwin's theory of evolution. If two sequences share high similarity, they are speculated to evolve from the same ancestor through the process of nucleotide replacement, sequence fragments, and missing and genetic variations. In sequence alignment, two or more sequences are put together in a way that the same nucleotide bases are aligned in the same column. Occasionally, gaps are inserted into the sequence in order to obtain the best alignment result.

There are a number of alignment tools, most of which utilize one of the alignment algorithms: Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) and Smith-Waterman algorithm (Smith & Waterman, 1981). These two

algorithms are both based on dynamic programming with the difference that Needleman-Wunsch algorithm is a global alignment technique, whereas Smith-Waterman algorithm is a general local alignment method.

Widely used alignment software includes BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990), BWA (Li & Durbin, 2009a), MOSAIK (htt1), BFAST (Homer et al., 2009), Bowtie (Langmead, Trapnell, Pop & Salzbeng, 2009), SOAP (Li, Li, Kristiansen, & Wang, 2008), and SSAHA (Ning, Cox, & Mullikin, 2001), etc.

NGS Limitation

Next-generation sequencing techniques provide higher throughput and a cheaper way of sequencing DNA than the traditional Sanger method. A high-throughput sequencing technique enables the genome to be sequenced in a day or less, or to sequence large genomes, such as the human genome. Another advantage is that RNA-seq is able to provide information about the entire transcriptome of a sample without knowing the genetic sequences of the organism in advance. However, NGS still has a lot of limitations:

- Accuracy: although the accuracy for NGS is relatively high, it is lower than traditional Sanger method due to its technique error and sequencing principle.
- Hard sequencing region: short sequencing length leads that some regions in genome are hard to be sequenced by next-generation sequencing.
- Storage: next-generation sequencing generates large amount of data, which gives rise to a big storage problem.

- Further analysis: data analysis can be time-consuming and may require special knowledge of bioinformatics to gain accurate information from sequence data.

Dissertation Organization

This dissertation is organized as follows: in Chapter I, the author introduces biological background, including different types of sequencing technologies, comparison for those technologies, basic analysis for NGS data, and NGS method limitations.

Chapter II introduces the SeqAssist, which consists of three parts: SA_RunStats, SA_Run2Run, and SA_Run2Ref. The SA_RunStats workflow generates basic statistics about an NGS dataset, including numbers of raw, cleaned, redundant and unique reads, redundancy rate, and a list of unique sequneces with length and read count. The SA_Run2Ref workflow estimates the breadth, depth, and evenness of genome-wide coverage of the NGS dataset at a nucleotide resolution. The SA_Run2Run workflow compares two NGS datasets to determine the redundancy (overlapping rate) between the two NGS runs.

Chapter III presents a novel and integrative SV discovery (SVDisc) pipeline that provides an all-in-one toolkit for investigators who are interested in identifying SVs in their studied species from genome re-sequencing data.

Chapter IV presents a new developed tool miRDisc, which is a new miRAN discovery algorithm to predict known and putative conserved/novel miRNAs from small RNA deep sequencing reads using assembled transcriptomes as the guidance for miRNA precursors.

Chapter V applies the developed tools to the experimental biological data and shows the results. And the last chapter, Chapter VI, is the conclusion and recommendations for future work.

CHAPTER II

SEQASSIST: A NOVEL TOOLKIT FOR PRELIMINARY ANALYSIS OF NEXT-GENERATION SEQUENCING DATA

Motivation

High throughput next-generation sequencing (NGS) technologies are capable of generating massive amounts of data in the form of paired-end or single-end reads with either fixed or variable lengths. The size of data files is often in the magnitude of mega- or giga-bytes (up to 1000 giga base pairs or Gb in a single sequencing run) and is likely to further increase in the coming years. While sequencing costs have dropped precipitously and sequencing speed and efficiency have raised exponentially, the development of computational tools for preliminary analysis of these gigantic datasets have lagged compared to the data generation. Hence, there is an increasing demand for efficient and user-friendly programs for preliminary sequencing data analysis.

At present, there are four commercially predominant NGS platforms, including Illumina/Solexa, Roche/454, ABI/SOLiD, and ABI/Ion Torrent (Mardis, 2013; Mardis, 2008). These massively parallel DNA sequencing technologies have been applied to transcriptome sequencing (RNA-Seq), *de novo* genome sequencing, and genome re-sequencing. RNA-Seq is a widely used approach to transcriptomic profiling (Martin & Wang, 2011; Wang, Gerstein, & Snyder, 2009b). Two representative efforts using *de novo* genome sequencing are the Genome 10K project to obtain the whole genome sequences for 10,000 vertebrate species (Bernardi et al., 2012; Scientists 10K Community of Scientists, 2009; Wong et al.,

2012), and the 5K Insect Genome Initiative (i5K) to sequence the genomes of 5,000 arthropod species (i5K Consortium, 2013; Levine, 2011). Genome re-sequencing is an experimental procedure that involves sequencing individual organisms whose genome is already known (Stratton, 2008). As a new genomics approach, genome re-sequencing has been applied to a wide range of fundamental and applied biological research including genetics, evolution, biomedicine, human diseases, and environmental health, etc., with good examples of the 1000 Genomes Project (Abecasis et al., 2012) and the Cancer Genomes project (Stephens et al., 2012).

Prior to the in-depth analysis of NGS deep sequencing data (differential gene expression and alternative splicing analysis for RNA-Seq studies, structural variants identification for genome re-sequencing studies, and genome assembly for *de nove* genome sequencing studies), investigators were often concerned about the following issues: (1) basic statistics of a sequencing run such as total numbers of raw, cleaned, and unique reads as well as the degree of reads redundancy; (2) sequencing library quality, i.e., whether the library truly represents the genome of the re-sequencing organism, and (3) the number of sequencing runs required, i.e., how many runs are necessary to attain a full representation of the sequencing library or to suffice a *de novo* genome assembly. To my best knowledge, there are currently no available tools that address these issues. Motivated by filling this gap and also driven by the demand for accelerating data-to-results turnaround, the author has developed a novel toolkit named SeqAssist (short for “Sequencing Assistant,” acronym: SA).

SeqAssist specifically addresses the aforementioned three issues and provides investigators who conduct RNA-Seq, *de novo* genome sequencing or genome re-sequencing experiments with a quick overview and preliminary analysis of their NGS data.

Current Method

There are a number of sequencers for next-generation sequencing technologies. The sequencers provide not only sequencing function, but also some basic data analysis tools, which mostly provide some statistics information for the generated sequence data. Take MiSeq as an example, it is developed by illumine company in 2011. It only needs 50ng DNA for the library preparation and takes several hours to finish sequencing and further analysis. Table 2 shows an output from the MiSeq analysis tool, which describes the depth of sequencing data. In this table, each column represents a chromosome or scaffold in the reference genome and each row stands for the depth. Therefore, each cell in the table means how many bases in a certain scaffold have the corresponding depth. Column two shows the total number with the depth of the whole chromosomes. For example, the second row shows the number of bases with depth 0. The total number shows in the second columns, and the number distributed into each scaffold shows from third column to the end. Such a summarized table is able to show the big picture as to how good the sequencing data is. However, sometimes researchers would like to know the exact depth for each specific position, average depth for each scaffold or a specific region, and the coverage breadth for each scaffold and the whole reference genome.

Table 2

Depth of Sequencing Data from MiSeq

Depth	Overall	scaffold_1	scaffold_2	scaffold_3	scaffold_4	scaffold_5
0	118075225	2048506	2010402	2125713	1984638	1388510
1	17849685	571008	470901	478648	299777	308718
2	11610660	404741	308756	299312	200604	201896
3	8387256	297577	233380	218970	144965	142161
4	6204350	228538	176372	163405	112646	103582
5	4618357	164393	132202	124325	90147	83751
6	3464223	124996	99472	93318	63783	65010
7	2577102	93685	74771	72695	47006	49029
8	1918093	68535	57973	52406	34634	38525
9	1438849	49300	42015	38406	26638	28452
10	1077096	34949	32037	27143	20317	22736
11	811467	25904	23336	20450	13631	17402
12	619723	19588	18206	16147	9778	12699
13	479845	15066	14494	11606	7877	10552
14	367114	10784	12057	9262	4942	7215
15	288928	7963	8781	6466	3254	5907
16	228195	5943	6244	4378	2608	5423
17	179629	4361	4831	2974	1861	3943
18	142854	3170	3412	2301	1041	3005
19	119515	3089	2393	1696	974	2293

Table 2 (continued).

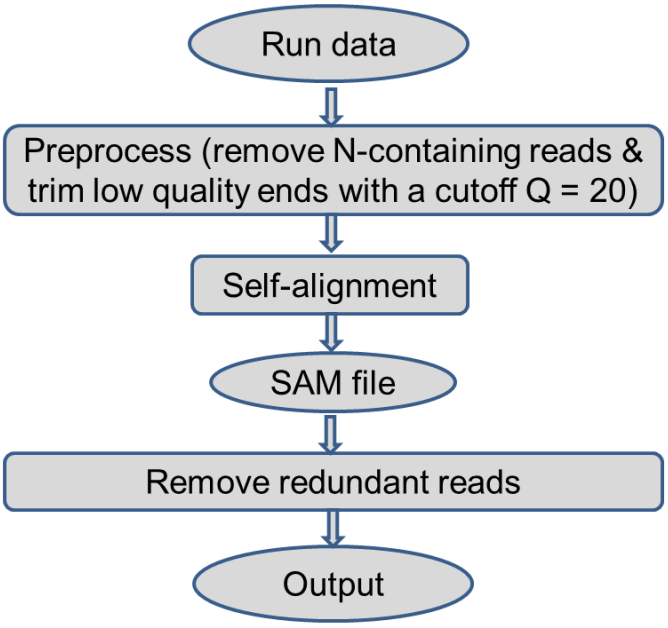
Depth	Overall	scaffold_1	scaffold_2	scaffold_3	scaffold_4	scaffold_5
20	96707	2375	2174	1259	891	1645
21	83427	1783	1690	1124	605	1515
22	70665	1296	1206	705	413	980
23	60673	935	828	475	286	849
24	51998	681	549	353	181	603
25	44900	440	651	245	156	348
26	39665	246	509	309	201	253
27	35062	267	269	308	189	150
28	29680	201	105	250	179	113

SeqAssist Toolkit

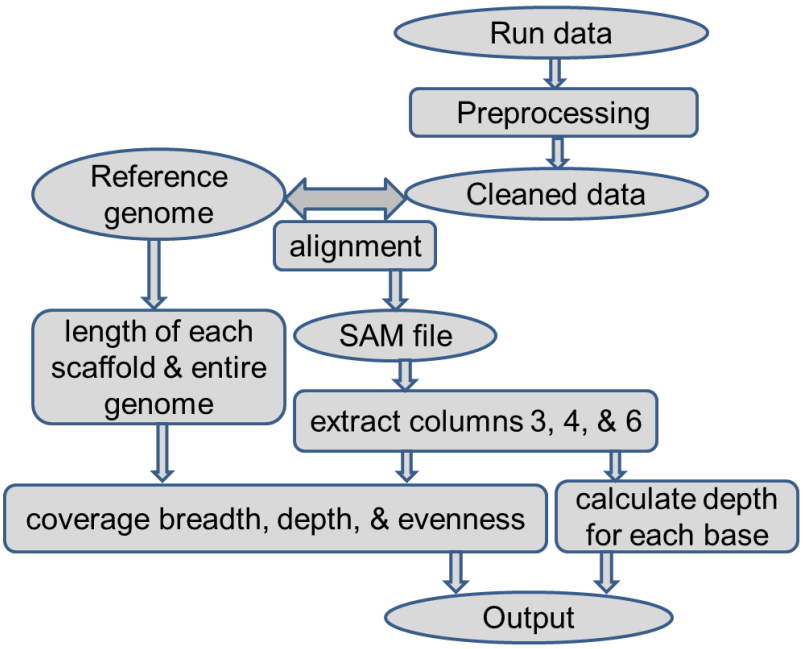
Overview of SeqAssist Pipeline

SeqAssist consists of three separate workflows: SA_RunStates, SA_Run2Ref, and SA_Run2Run. SA_RunStates generates the basic statistics such as the total number of raw and cleaned reads, length and copy number of unique sequences, and reads redundancy in a single sequencing run or a pooled dataset of several runs (see Figure 5a). SA_Run2Ref analyzes the breadth, depth, and evenness of genome-wide coverage of an individual or pooled sequencing dataset at a nucleotide resolution (see Figure 5b). Outputs from SA_Run2Ref can demonstrate what genomic loci are covered and how a genomic locus (gene), scaffold, or the entire genome is covered. SA_Run2Run

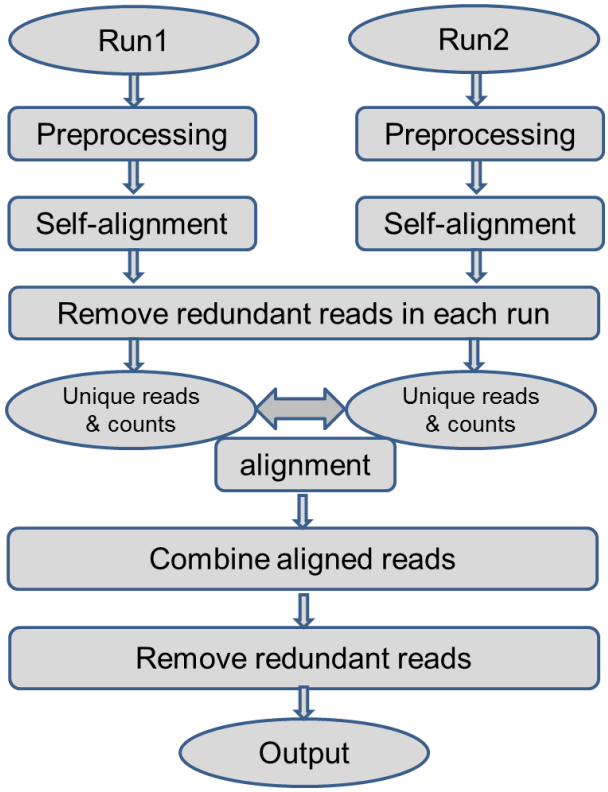
compares two separate sequencing datasets generated from the same DNA libraries, computes the basic statistics for each individual dataset, and estimates the redundancy rate between the two datasets (see Figure 5c). SA_Run2Run informs the user about the redundancy level both within each individual run and between two sequencing runs.



(a) SA_RunStats



(b) SA_Run2Ref



(c) SA_Run2Run

Figure 5. Workflow of SeqAssist pipeline: (a) SA_RunStats, (b) SA_Run2Ref, and (c) SA_Run2Run.

Dependency

BWA-MEM algorithm

BWA-MEM is alignment software which uses the maximal exact matches (MEM) as seed and extends the seed with gaps using Smith-Waterman algorithm. Smith-Waterman algorithm is a local sequence alignment algorithm, which compares all the possible common sequences and adopts the optimal solutions. An example of Smith-Waterman algorithm is shown in Figure 6.

sequence 1= ATCACA

sequence 2= ACACCA

match=+2, mismatch=-1

	-	A	T	C	A	C	A
-	0	0	0	0	0	0	0
A	0	2	1	0	2	1	3
C	0	1	1	3	2	4	3
A	0	3	2	2	5	4	6
C	0	2	2	4	4	7	6
C	0	1	1	6	5	9	8
A	0	3	2	5	8	8	11

sequence 1= ATCAC_A

sequence 2= A_CACCA

Figure 6. An example of Smith-Waterman algorithm.

SAM format

SAM (Sequence Alignment/Map format) (Li et al., 2009b) is a tab-delimited text format to store alignment or mapping results. There are two sections for SAM format: the header section (optional) and the alignment section. In the alignment section, each line represents one alignment result, which consists of eleven mandatory and some optional fields. Table 3 shows the eleven mandatory columns, such as query name, alignment flag, aligned reference name, alignment start position, mapping quality, CIGAR information, reference sequence name of the primary alignment of the NEXT read in template, position of primary alignment of NEXT read in the template, signed observed template length, sequence segment, and its associated quality score. CIGAR string explains how the sequence aligns to the reference genome.

Table 3

SAM Format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	Bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality

Table 3 (continued).

Col	Field	Type	Regexp/Range	Brief description
6	CIGAR	String	[*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	Observed Template LENgth
10	SEQ	String	[*|[A-Za-z=.]+	Segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

SA_RunStats Pipeline

SA_RunStats generates the basic statistics such as the total number of raw and cleaned reads, length and copy number of unique sequences, and reads redundancy in a single sequencing run or a pooled dataset of several runs. The input of this workflow is a FASTQ-formatted sequencing data file. The data file is preprocessed by first trimming off the adaptors and low quality read ends with a default cutoff of base-calling quality score (Q) of 20, followed by the removal of N-containing reads. Then, the cleaned reads are aligned with each other using BWA-MEM (acronym for Burrow-Wheeler Aligner-Maximal Exact Match) algorithm, one of the three Burrows-Wheeler Transform-based algorithms in the

BWA software package. Based on the alignment information in the BWA-MEM-generated SAM (acronym for Sequence Alignment/Map format) file (Li et al., 2009b), the number of unique reads is counted in which both identical and inclusive (i.e., redundant) reads are removed. Two reads are considered identical if they are a 100% match and are of equal length, while inclusive reads are defined as the sub-sequencing of a longer read and only the longest read is kept as the unique read. The redundancy rate is calculated as the percentage of redundant reads in the total number of unique cleaned reads (see Equation 2.1 for formula). The output of this workflow includes the total numbers of raw, cleaned, redundant and unique reads, and the redundancy rate. Also included in the output is a tab-delimited text file that lists all unique sequences along with their length and read count (copy number). This file can be used to further infer gene expression levels if the run data is produced for an RNA-Seq experiment.

$$\text{Redundancy rate}\ (\%) = \frac{\text{number of redundant reads}}{\text{total number of unique cleaned reads}} \times 100\% \qquad (2.1)$$

SA_Run2Ref Pipeline

SA_Run2Ref analyzes the breadth, depth, and evenness of genome-wide coverage of an individual or pooled sequencing dataset at a nucleotide resolution. Coverage breadth is defined as the percentage of a reference sequence (i.e., gene, scaffold/chromosome, or entire genome) that is covered by sequencing reads (see Equation 2.2 for formula); coverage depth is defined as the average times a reference sequence is covered (see Equation 2.3 for formula); and coverage evenness is defined as the coefficient of variance of scaffold coverage breadth (see Equation 2.4 for formula). Therefore, outputs from SA_Run2Ref can

inform what genomic loci are covered and how a genomic locus (gene), scaffold, or the entire genome is covered.

$$\text{Coverage breadth (\%)} = \frac{\text{number of reference bases mapped by sequencing reads}}{\text{length of the reference sequence in bases}} \times 100\% \quad (2.2)$$

$$\text{Coverage depth} = \frac{\text{total number of bases mapped to the reference}}{\text{length of the reference sequence in bases}} \quad (2.3)$$

$$\text{Coverage evenness} = \frac{\text{standard deviation of scaffold coverage breadth}}{\text{average scaffold coverage breadth}} \quad (2.4)$$

In the SA_Run2Ref workflow, cleaned reads are aligned against the reference genome sequence, generating an SAM file. Information stored in columns 3, 4, and 6 for each alignment in the SAM file represents mandatory fields RNAME (reference sequence name), POS (1-based leftmost mapping position), and CIGAR (CIGAR string), respectively (Li et al., 2009b). This information is extracted along with the length of each scaffold of the reference genome to compute scaffold coverage breadth and depth and genome coverage evenness. These statistics are provided in the output files, which also include a plain-text file that records the coverage depth of each individual base in the entire genome. This file can be used as an input for genome browser tools to visualize the coverage depth of any genomic regions. In the case that users conduct an RNA-Seq experiment and provide gene model sequences (instead of scaffold or chromosome sequences) as the input, the workflow will calculate the coverage breadth and depth for each gene model. This information can be readily transformed into a gene expression measurement.

SA_Run2Run Pipeline

SA_Run2Run compares two separate sequencing datasets generated for the same or different DNA libraries, computes the basic statistics for each individual dataset, and estimates the redundancy rate between the two datasets. SA_Run2Run informs the user about the redundancy level both within each individual run and between two sequencing runs.

Preprocessing

The inputs of SA_Run2Run are two experiment sequencing data called two runs that are both in fastq format. The following data pre-treatment steps as the preprocessing steps for the other two pipelines are applied prior to further analysis: (1) trim off adaptors; (2) remove low-quality bases from each end with the default base-calling quality score Q of 20; (3) trim off adaptors again in case the low-quality based cause mismatch with the adaptors; (4) remove N-containing reads. Besides the above four reads cleaning steps, the input files are converted from fastq format into fasta format (remove lines of additional information and sequence quality scores), which is the required format type for the use of BWA-MEM algorithm.

Repeat removal

For each dataset, the repeat reads are removed. The redundant reads have different definitions depending on whether the dataset shared the same length reads. For fixed length reads, the redundancy is considered as two or more reads sharing the same sequence. In terms of variable length reads, redundancy is the subsequence compared to their super sequences. Only the

super sequences (the longest sequence) are kept for further analysis. In this step, each run aligns itself when the identical and inclusive redundancy are removed using the alignment output SAM files.

Alignment

After cleaning the data and removing redundancy, the pipeline moves to the alignment phase. BWA-MEM is used in this step to align the two runs against each other. The standard output format for BWA-MEM is in the SAM format. Then, the 100% aligned reads are extracted from the BWA-MEM output by filtering out alignment with full-length match in CIGAR column and examining no-mismatch in the optional column. For the two runs, two alignment results are obtained for both directions: one is treated as reference, the other one is sequence, and vice-versa. Then, the alignment reads from the two alignment results are combined as the candidate of unique super sequences.

Repeat removal

All the alignment reads are extracted in the last step. However, there are redundant reads due to the same read aligned in multiple positions, the same sequences with different read names or super reads aligned in a different dataset. Hence in this step, self-alignment is executed by BWA-MEM again. During self-alignment, all the repeats are able to align with each other. Redundancy can be removed by the following strategy: databases are created to store all of the reference name so that the corresponding query name is already outputted but not in the database; if the query is neither in the database nor in the output, this alignment is outputed; if the query is not in the database but in the output, it is

pushed into database; if the query is in the database, the alignment is simple skipped. For example, suppose there are a1, a2, and a3 as three different reads with the same sequence. After aligning to each other, the result should be:

query	-> reference
a1	-> a1
a1	-> a2
a1	-> a3
a2	-> a1
a2	-> a2
a2	-> a3

Then the outputs are searched line by line. In the first line a1->a1, a1 is not in the database and thus it is outputted and is stored in the database. In the second line a1->a2, a2 will be stored in the database since a1 is already outputted. In the third line a1->a3, a3 will be also stored in the database since a1 is already in the output. From the fourth line to the end, a2 and a3 are both in the database and thus all these alignments would be skipped.

After all the steps, a tab-delimited text file is outputted without any redundancy. Each line is an alignment with a read aligned to itself. The number of lines is equal to the unique number of overlapping reads for the two input runs.

The output statistics from SA_Run2Run include the total numbers of raw reads, cleaned reads, and unique reads (after removing identical reads and inclusive reads), and numbers of total and unique overlapping reads. The redundancy rates within each dataset and between the two datasets can be

further derived from these statistics. Similar to the SA_RunStats output, a list of unique sequences along with their length and count number is provided for each run. However, different from the SA_RunStats output, the list generated by SA_Run2Run is broken into two files: one for overlapping reads and the other for non-overlapping reads. The SA_Run2Run workflow intends to guide the user in deciding whether to perform more runs on a sequencing library by looking at the percentage of reads in a new run covered by the reads in a previous run or the pooled reads of multiple previous runs.

Pipeline Testing

Testing Dataset

To test all SeqAssist workflows, a synthetic dataset was generated by the following steps:

- Clipping 10 distinct fragments with a length of 150 bp at different loci of the *Escherichia coli* str. K-12 substr. MG16551 genome (NCBI Reference Sequence Accession No. NC_000913.3, available at http://www.ncbi.nlm.nih.gov/nuccore/556503834?report=fasta) to construct 10 artificial chromosomes;
- Clipping 10 sequences of 75-100 bp in length from each artificial chromosome;
- Repeating each sequence 10 times. These steps result in a dataset of 1,000 reads and a reference genome consisting of 10 short artificial chromosomes, both of which are used to test the SA_RunStats and SA_Run2Ref workflows. The synthetic dataset is further split into two

halves to create Run1 and Run 2 that were used to test the SA_Run2Run workflow.

Testing Result

Table 4 illustrates the result for testing the dataset using SA_RunStats. Column C lists the expected result for the testing dataset, and column D shows the actual result obtained from the output files. From this table, it can be seen that the result from the workflow is exactly the same as the expected result.

Table 4

SA_RunStats Testing Result

	A Output statistics	B Synthetic data	C Expected result	D Result from workflow
1	Total number of reads	1000	1000	1000
2	Number of reads containing N	0		
3	Number of cleaned reads	B1-B2	1000	1000
4	Number of repeats/identical read (copy number)	10		
5	Number of inclusive reads (shorter reads that are part of longer ones)	40		
6	Unique number of reads (after removing identical repeats)	B1/B4	100	100
7	Unique number of reads (after removing identical & inclusive repeats)	B1/B4-B5	60	60

Table 5 and Table 6 list all the result outputs from SA_Run2Ref. Table 5 shows the length and coverage breadth for the 100 synthetic reads (10 chromosomes and 10 reads for each chromosome). According to the coverage breadth for each read in each chromosome, the coverage depth for each chromosome and for the whole dataset can be calculated. Finally, coverage depth for the 100 unique reads is about 58.54. Then based on coverage breadth and depth for each chromosome, the coverage breadth, coverage depth, and coverage evenness for the whole dataset can be obtained as shown in Table 6. Note that the expected result (column C) and the result from workflow (column D) are consistent.

Table 5

*Length and Coverage Breadth of 100 Synthetic Reads (10 chr *10 unique reads)*

	A	B	C	D	E	F	G	H	I	J	K	L
	Synthetic reads	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	All chr
8	Read1	76	99	88	95	83	91	87	92	75	91	877
9	Read2	80	77	91	100	86	100	98	76	98	98	904
10	Read3	82	77	91	100	81	94	100	86	100	85	896
11	Read4	96	77	98	98	94	84	86	80	89	99	901
12	Read5	86	96	100	93	97	81	78	85	81	100	897
13	Read6	100	100	100	87	93	79	80	99	91	95	924
14	Read7	75	79	76	79	100	88	87	80	93	83	840
15	Read8	77	95	76	88	78	92	79	81	76	81	823

Table 5 (continued).

	A	B	C	D	E	F	G	H	I	J	K	L
	Synthetic reads	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	All chr
16	Read9	86	96	79	88	90	87	77	84	78	81	846
17	Read10	100	88	90	75	98	83	96	80	83	80	873
18	Sum of 10 reads	858	884	889	903	900	879	868	843	864	893	8781
19	Chromosome coverage depth	57.20	58.93	59.27	60.20	60.00	58.60	57.87	56.20	57.60	59.53	58.54
20	Length of chromosome	150	150	150	150	150	150	150	150	150	150	1500
21	Length of chromosome covered by synthetic reads	150	150	150	150	150	150	150	149	148	148	1495
22	Breadth of chromosome coverage by synthetic reads	1	1	1	1	1	1	1	0.9933	0.987	0.9867	0.99666667

Table 6

SA_Run2Ref Testing Result

	A Ouput statistics	B Synthetic data	C Expected result	D Result from workflow
23	Number of synthetic chromosomes	10		
24	Length of each synthetic chromosome	150		
25	Sum of length of 100 synthetic reads (10 chr X 10 unique reads)	8781		
26	Genome coverage depth	(B25*B4)/(B23*B24)	58.54	58.54
27	Genome coverage breadth	L21/(B23*B24)	0.996666667	0.996666667
28	Genome coverage evenness	standard deviation (B19:K19)/L19	0.021038647	0.021038647

The results for SA_Run2Run are shown in Table 7. Since the sequencing dataset are evenly split into two smaller runs, the two smaller ones are completely identical. Then the result of basic statistics for each run should be half of the whole sequencing dataset, and when comparing the two runs, the overlapping part should be equal to each smaller run.

Table 7

SA_Run2Run Testing Result

	A Output statistics	B Synthetic data	C Expected result	D Result from workflow
29	Total number of reads in run1	500	500	500
30	Number of reads with N in run1	0		
31	Number of cleaned reads in run1	B29-B30	500	500
32	Identical repeats/chromosome in run1	5		
33	Inclusive repeats in run1	40		
34	Unique number of reads (after removing identical repeats) in run1	B29/B32	100	100
35	Unique number of reads (after removing inclusive repeats) in run1	B56/B32-B33	60	60
36	Total number of reads in run2	500	500	500
37	Number of reads with N in run2	0		
38	Number of cleaned reads in run2	B36-B37	500	500
39	Identical repeats/chromosome in run2	5		
40	Inclusive repeats in run2	40		
41	Unique number of reads (after removing identical repeats) in run2	B36/B39	100	100
42	Unique number of reads (after removing inclusive repeats) in run2	B36/B39-B40	60	60

Table 7 (continued).

	A Output statistics	B Synthetic data	C Expected result	D Result from workflow
43	Total overlapping reads in run1	run1=run2	500	500
44	Unique overlapping reads in run1	run1=run2	60	60
45	Total overlapping reads in run2	run1=run2	500	500
46	Unique overlapping reads in run2	run1=run2	60	60
47	Unique overlapping reads in two runs	run1=run2	60	60

All the results from three pipelines show the consistency of expected results from the pipeline and real data. Thus, the SeqAssist pipeline can be used to show the basic statistical information for the biological data based on different pipelines.

CHAPTER III

SVDISC: A NOVEL AND INTEGRATIVE PIPELINE FOR STRUCTURAL VARIANTS DISCOVERY USING GENOME RE-SEQUENCING DATA

Introduction

Genomic structural variation (SV) is the variation in DNA sequence structure within an organism's chromosome. SVs can be divided into two categories: (1) balanced rearrangements including inversions and translocations, and (2) unbalanced rearrangements or copy number variants (CNVs) including insertions, deletions, and duplications (Mills et al., 2011). Unlike point mutations, SVs vary widely from a few bp to as large as a few Mbp in size. Mounting evidence suggests that SVs are abundant in human genome and account for a much larger fraction of genetic variation than single nucleotide polymorphism (SNP), implying significant consequences of SVs on phenotypes (Abecasis et al., 2010; Abecasis et al., 2012; Feuk, Carson, & Scherer, 2006; Mills et al., 2011). For instance, recent studies have revealed the association of micro deletions with a number of genomic disorders such as learning disability (Shaw-Smith et al., 2006), Autism (Weiss et al., 2008), and mental retardation (Sharp et al., 2008). The two SV repository databases, Database of genomic structural variation (db Var) and Database of Genomic Variants archive (DGVa), have recorded over 7.7 million variant calls as of September 2012 (Lappalainen et al., 2013).

Researchers employing the sequencing approach make variant calls by either de novo assembling sequence reads ('AS') or aligning sequencing reads to a reference genome ("re-sequencing") (Mills et al., 2011). Due to the high depth

of genome coverage required by the AS strategy, the re-sequencing strategy has been more widely adopted, which consists of two main steps: (1) alignment of reads, and (2) prediction of SVs from alignment. Although the re-sequencing strategy is straightforward in principle, sensitive and specific SV deletion is actually difficult in practice (Alkan, Coe, & Eichler, 2011; Medvedev, Stanciu, & Brudno, 2009).

Algorithms that is used to predict a full spectrum of SV events from sequence alignment/mapping (SAM/BAM) (Li et al., 2009b) files have been fast growing. These algorithms can be generally classified into four categories (Mills et al., 2011; Suzuki, Yasuda, Shiraishi, Miyano, & Nagasaki, 2011): (1) discordant pair or read pair ("RP") analysis, (2) depth of coverage or read depth ("RD") analysis, (3) split read ("SR") analysis, and (4) integrated analysis such as DELLY (Rausch et al., 2012), a method integrating RP mapping with SR refinement, and Genome STRiP (Handsaker, Korn, Nemesh, & McCarroll, 2011) and GASVPro (Sindi, Onal, Peng, Wu, & Raphael, 2012), both combining information from RP and RD analyses (called "PD"). Briefly, the RP algorithms, e.g., VariationHunter (Hormozdiari et al., 2010), PEMer (Korbel et al., 2009), and BreakDancer (Chen et al., 2009), are based on analysis of abnormally mapping NGS read pairs; the RD algorithms, e.g., CNVnator (Abyzon, Urban, Snyder, & Gerstein, 2011a), SegSeq (Chiang et al., 2009), and Event-Wise Testing (Yoon, Xuan, Makaron, Ye, & Sebat, 2009), detect SV events by statistically analyzing the difference in the number of reads aligned to intervals of the reference genome; and the SR algorithms, e.g., Pindel (Ye, Schulz, Long, Apweiler, & Ning,

2009), SLOPE (Abel et al., 2010), and ClipCrop (Suzuki et al., 2011), identify SV breakpoints by anchoring the mapped read mate in the reference genome and split-aligning the prefix and suffix of the unmapped read mate independently to different locations.

In view of the current state of the art development of SV discovery tools, it has been realized that now is the time to integrate existing tools and develop a comprehensive pipeline that serves as a one-stop shop for SV identification. Meanwhile, there also exist increasing demands for such comprehensive tools from researchers who investigate SV contributions to phenotypic variations in a broad range of fields such as biomedicine, cancer genetics or genomics, toxicology, and ecology, as bioinformatics infrastructure often constitutes one of the biggest bottleneck factors, especially for research groups that a lack of bioinformatics support personnel.

Here, the author presents a novel and integrative SV discovery (SVDisc) pipeline that provides an all-in-one toolkit for investigators who are interested in identifying SVs in their studied species from genome re-sequencing data. The novelty of SVDisc lies in the fact that there is no similar pipeline or infrastructure available in the SV research community. It can detect all the common types of SVs with user-defined sizes (default size=50 bp), including insertions, deletions, duplications, inversions, intra-chromosomal, and inter-chromosomal translocations. Currently, SVDisc is a stand-alone downloadable package. The output includes: (1) a list of all identified SVs that are categorized into 6 different

SV types, (2) associated evidence of supporting sequences reads, and (3) functional annotation of identifies SVs.

Alignment Methods: BWA and MOSAIK

BWA Aligner

BWA (Burrows-Wheeler Alignment tool) is a new read alignment package based on a backward search with Burrows-Wheeler Transform (BWT). The Burrows-Wheeler Transform, invented by Michael Burrow and David Wheeler in 1994, is an algorithm used in data compression techniques, which permutes the order of the characters. Suppose there is a string $X=a_0a_1\ldots a_{n-1}$, it is always ended with symbol \$, and this symbol only appears at the end. Let $X[i]=a_i, i=0,1,\ldots,n-1$, be the i-th symbol of X, $X[i,j]=a_i\ldots a_j$ a substring and $X_i=X[i,n-1]$ a suffix of X. Suffix array S of X is a permutation of the integer $0\ldots n-1$ such that $S(i)$ is the start position of the i-th smallest suffix. The BWT of X is defined as $B[i]=\$$ when $S(i)=0$ and $B[i]=X[S(i)-1]$ otherwise. The length of string X is defined as Xv and therefore $|X|=|B|=n$. According to the principle of the Burrow-Wheeler Transform, the same substring or substring with the same prefix will be together since they will be sorted together. If string W is a substring of X, the following equations are defined:

$$\underline{R}(W)=\min\{k:W\text{ is the prefix of }X_{s(k)}\}$$

$$\overline{R}(W)=\max\{k:W\text{ is the prefix of }X_{s(k)}\}$$

Thus, the position region of all occurrences of W in X as prefix is $S(k):\underline{R}(W)\leq k\leq\overline{R}(W)$. There is a backward search followed by the Burrow-Wheeler Transform. Ferragina and Manzini proved that if W is a substring of X:

$$\underline{R}(aW)=C(a)+O(a,\underline{R}(W)-1)+1$$

$$\overline{R}(aW)=C(a)+O\left(a,\overline{R}(W)\right)$$

and that $\underline{R}(aW)\leq\overline{R}(aW)$ if and only if aW is a substring of X, where $C(a)$ is the number of symbols in $X[0,n-2]$ that are lexicographically smaller than the alphabet, and $O(a,i)$ is the number of occurrences of a in $B[0,1]$. This result makes it possible to test whether W is a substring of X.

MOSAIK Aligner

MOSAIK is a reference-guided assembler/aligner. It consists of four programs: MosaikBuild, MosaikAligner, MosaikSort, and MosaikAssembler. For our usage of alignment, only the first two programs are employed. MosaikBuild is able to convert multiple file formats, such as FASTA, FASTQ, Illumina Bustard, Illumina Gerald, and SRF files, into the compressed binary file formats. To speed up the whole process, Mosaik not only compresses the read files but also converts the reference sequences FASTA file to a binary format. After file transformation, MOSAIK perform alignment using MosaikAligner. Figure 7 shows the workflow of MosaikAligner.

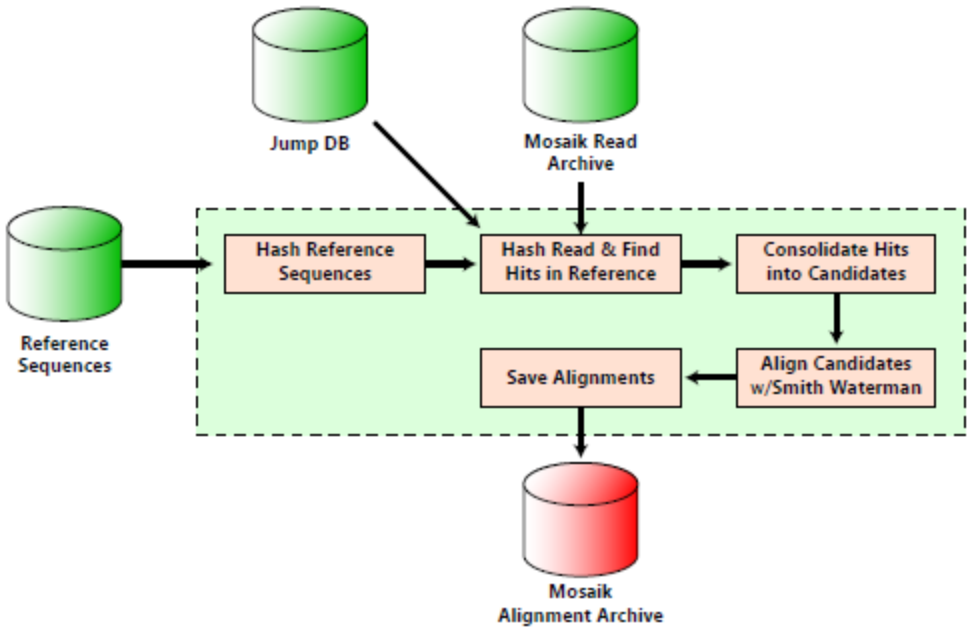


Figure 7. Workflow of MosaikAligner.

First of all, MosaikAligner hashes the reference genome. When searching a read, the read is hashed in similar jump databases. Then, MosaikAligner retrieves the reference position for each hash in the hash table. All of the hash positions are clustered together and evaluated with the Smith-Waterman algorithm.

Structural Variation Detection Methods: Pindel, BreakDancer, and CNVnator

Pindel

Pindel (Ye et al., 2009) is a split-read structural variation identification method for detecting large deletions and medium sized insertion by using a pattern growth approach. This method uses the one-end mapped paired-end reads. The mapped end is treated as anchor, and then the algorithm searches the minimum and maximum unique substrings of the unmapped end for both 5' and 3' in the setting region around the anchor position. The final break points are

identified after comparing all the unique substrings. The algorithm uses the pattern growth approach to detect the minimum and maximum unique substrings. This approach works as follows: Suppose there is a string P and a is a substring that starting from the leftmost of it and S_a is the projected database, which contains all locations of sequences that have the substring a. Then, another projected database $S_{a'}$ is calculated from S_a, which for each location of a, checks whether or not the base on its right-hand equals the newly appended character b. Any location without an appending item b is removed from the projected database. The algorithm checks one base by one until the unique minimum and maximum substrings are identified. Figure 8 shows an example of how the pattern growth approach works.

pattern:	TACGT
sequence:	TAGTTCATACGAATCT
pattern growth:	
T	TAGTTCATACGAATCT
TA	TAGTTCATACGAATCT
TAC	TAGTTCATACGAATCT
TACG	TAGTTCATACGAATCT

Figure 8. Pattern growth algorithms.

Suppose there is a short read “TACGT” and a sequence “TAGTTVATACGAATCT”. The purpose is to find out the unique minimum and maximum substring of the short read in the sequence. From the leftmost letter T, all of the positions of T in sequence are marked. Then, move to the next letter A and keep the positions where there is A following T. When moving to the third

letter C, only one position can be matched to TAC. TAC is called the unique minimum substring. Repeating the above process, TACG is matched and no longer substring can be matched. Then, TACG is the unique maximum substring.

BreakDancer

BreakDancer (BreakDancerMax) (Chen et al., 2009) is a typical read-pair method with a detection range from 100 basepair to 1 mega basepair. It provides five types of structural variations, including deletion, insertion, inversion, intra-chromosomal translocation, and inter-chromosome translocation.

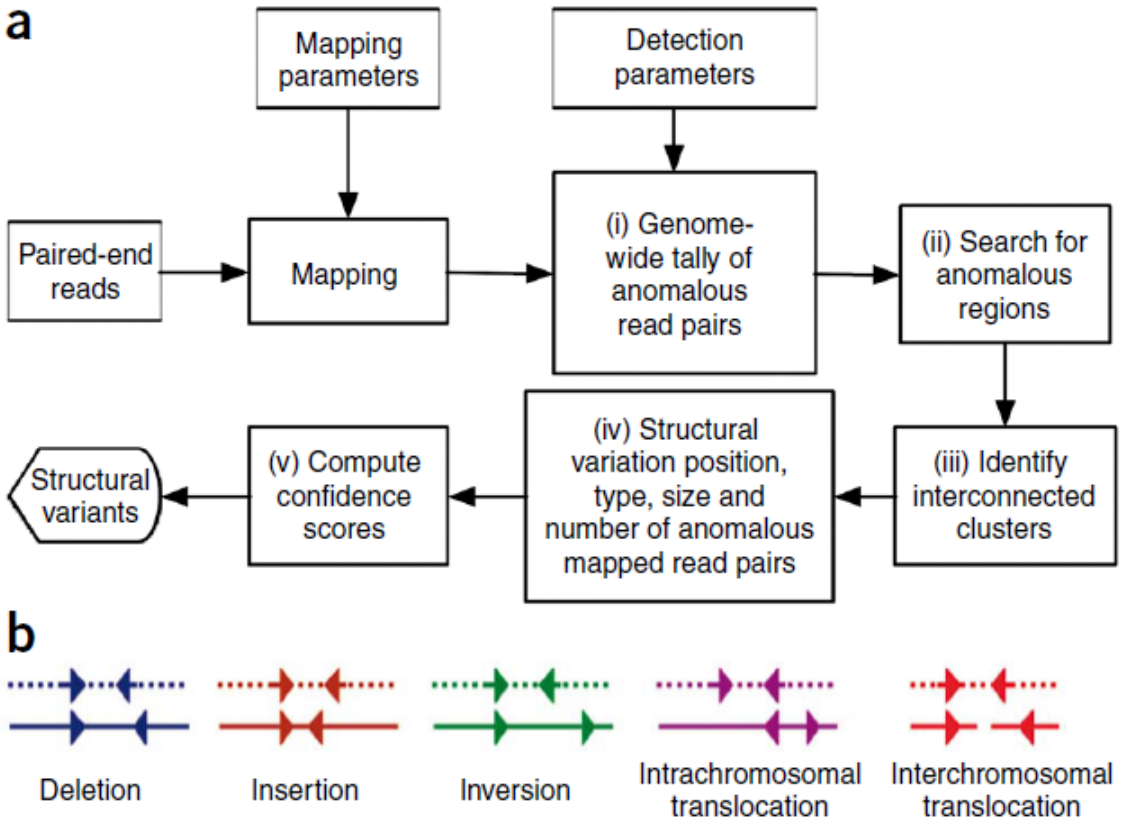


Figure 9. (a) workflow of BreakDancer and (b) anomalous read pair recognized by BreakDancer.

Figure 9a shows the workflow of BreakDancer. The algorithm first uses the mapped paired-end read to identify the anomalous read pairs according to

mapped distance and alignment orientation. All of the types of anomalous read pairs are shown in Figure 9b. Take deletion as an example; when the mapped paired-end is in the same orientation as the original sequence reads and the mapped distance is larger than insert size, a deletion is identified. Then, the algorithm searches for anomalous region and produces putative structural variant by combining two or more interconnected anomalous read pairs. Finally, a confidence score is estimated for each variant based on a Poisson model that takes into consideration the number of supporting anomalous read pairs, the size of the anchoring regions, and the coverage of the genome.

CNVnator

CNVnator (Abyzon et al., 2011a) is a read depth approach. This method first divides the entire reference genome into consecutive nonoverlapping bins of equal size. For each bin, the read depth (RD) signal is calculated as a number of placed reads with centers in bin boundaries and is corrected by the following formula to remove bias,

$$RD^{i}_{corrected} = \frac{\overline{RD}_{global}}{\overline{RD}_{gc}} RD^{i}_{raw}$$

where i is bin index, RD^{i}_{raw} is raw RD signal for a bin, $RD^{i}_{corrected}$ is corrected RD signal for the bin, $\overline{RD}_{global}$ is average RD signal over all bins, and $\overline{RD}_{gc}$ is the average RD signal over all bins with the same GC content as in the bin. Then, the algorithm uses PDF (probability density function) to calculate mean-shift vector, which is used to determine the directions of the RD signal for each bin.

Finally, the break points are determined where two neighboring vectors have opposite directions but do not point to each other.

SVDisc Pipeline

Overview

Figure 10 shows the flow chart of SVDisc. There are four main components in this pipeline: (1) preprocessing, including four steps: remove N-containing reads, trim adaptors, remove low quality bases and trim adaptors again, (2) structural variation detection, including sequence alignment and SV identification, (3) breakpoints revision, precise determine the locus of breakpoints, and (4) SV integration: obtain consensus SVs from outputs of multiple SV detection tools.

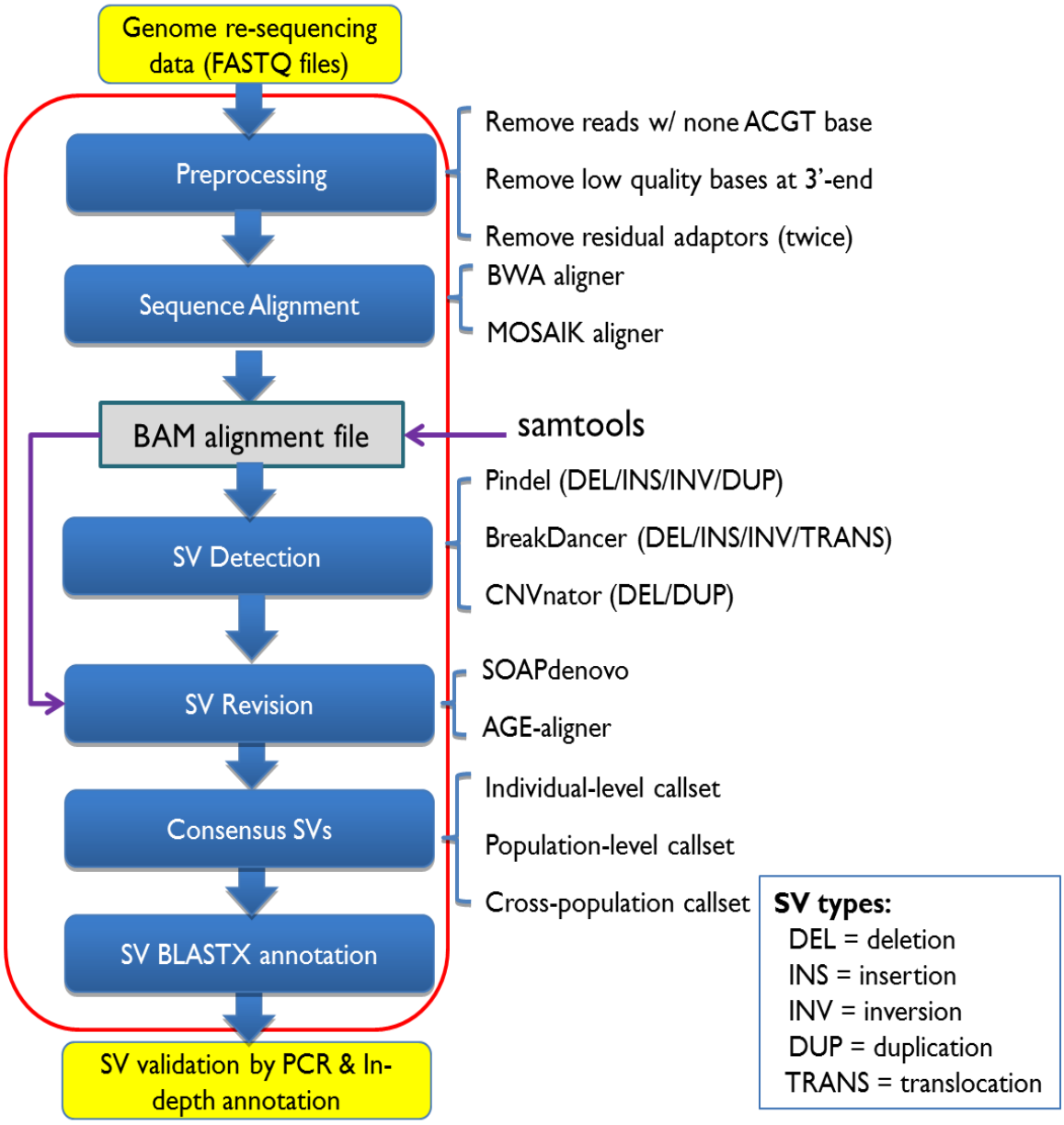


Figure 10. SVDisc workflow.

Preprocessing

The input of the pipeline is genome re-sequencing data. As discussed in Chapter I, during sequencing process, the fragments of sequence are appended with adaptors at both ends in order to fix them to the sequencer channel surface. Besides, the sequence reads contain ambiguous base calling (i.e., N or non-

A/C/T/G bases), which is hard to confirm the exact type of nucleotide during the sequencing procedure, and some low quality bases, which have low probability that the base captures the correct type of nucleotide. All of these will affect further analysis. So a four-step preprocessing procedure is designed to remove these bias: (1) remove N-containing reads in pairs; (2) trim adaptors, cutadapt is used to remove full or partial adaptors from both 5' and 3'; (3) remove low quality bases, contiguous low quality bases are remove from both ends; and (4) trim adaptors again, low quality may lead to the mismatch of adaptors to the reads, so adaptors are trimmed again after removing low quality bases.

Alignment

In this step, the cleaned reads are aligned to a reference genome. Two aligners, the Burrows-Wheeler aligner's Smith-Waterman Alignment (BWA-SW) and MOSAIK (see more detailed at https://code.google.com/p/mosaik-aligner/), are chosen regarding their compatibility to NGS read formats, alignment speed, memory footprint, accuracy, and output SAM/BAM format that are acceptable by subsequent SV discovery programs.

Structural Variation Detection

In this step, the complementary algorithms for SV detection are implemented, including Pindel (an SR method), BreakDancer (a RP method), and CNVnator (a RP method). These three methods were selected because of their relatively higher maturity, availability, and feasibility to be implemented, in comparison with other algorithms. As a prototype pipeline, the SVDisc does not include integrative algorithms because it was the author's belief that the

combination of selected algorithms representing different categories mentioned above would outperform any single integrative algorithms. For future upgrades, more methods will be added and will further look into the integrative algorithms such as Genome STRiP.

Breakpoints Revision

SOAPdenovo

SOAPdenovo (short for Oligounucleotide Analysis Package de novo assembly) (Li et al., 2009a) is a short-read assembly method, which adopts De Bruijn Graph (Li et al., 2012) to assemble short reads. An n-dimensional De Bruijn graph of m symbols is a directed graph, which represents overlaps between sequences of symbols. Considering all possible combinations of length n sequences, the graph totally has m^n vertices. If there is a set of m symbols $S:=\{s_1,\ldots,s_m\}$ then the set of vertices is:

$V=S^n=\{(s_1,\ldots,s_1,s_1),(s_1,\ldots,s_1,s_2),\ldots,(s_1,\ldots,s_1,s_m),(s_1,\ldots,s_2,s_1),\ldots,(s_m,\ldots,s_m,s_m)\}$.

If one of the vertices can be expressed as another vertex by shifting all of its symbols by one place to the left and adding a new symbol at the end of this vertex, then the latter has a directed edge to the former vertex. Thus, the set of arcs is:

$E=\{\big((v_1,v_2,\ldots,v_n),(v_2,\ldots,v_n,s_i)\big):i=1,\ldots,m\}$.

Figure 11 shows the basic principle of the De Bruijn Graph.

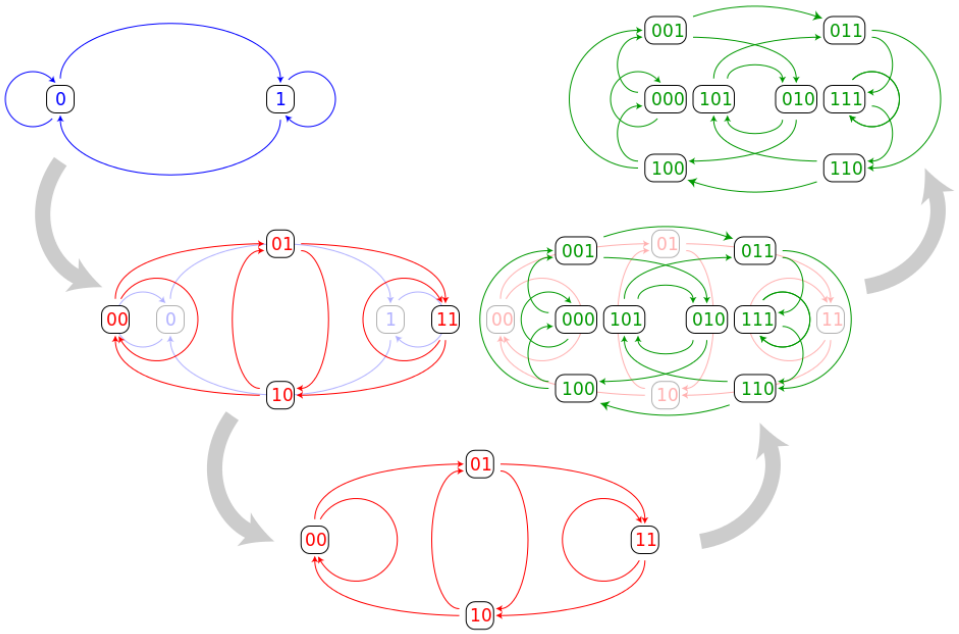


Figure 11. De Bruijn graph. http://upload.wikimedia.org/wikipedia/commons/thumb/9/9d/DeBruijn-as-line-digraph.svg/954px-DeBruijn-as-line-digraph.svg.png

SOAPdenovo firstly cuts all the reads into the length of K-mer. Then, it generates the De Bruijn graph based on the cut reads and finds out the longest pathway, which shifts one base between two nodes. The contig is obtained by routing the pathway. Then, the raw contigs from graph perform four steps: remove tips, solve the tiny repeats, merge bubbles, and finally link to generate the scaffolds.

AGE

AGE (Abyzov & Gerstein, 2011b) is a dynamic-programming algorithm for defining the precise location of structural variations. It finds the optimal solution by aligning the $5^{'}$ and $3^{'}$ ends of two given sequencing at the same time and introducing a “large-gap jump” between the local end alignments to maximize the total alignment score. Suppose there are two given sequences: N and M. The

maximum $M^L(n,m)$ in the leading submatrix $[0,n]x[0,m]$ of S^L, where $n \leq N$ and $m \leq M$, anchors the best local alignment for n and m nucleotides at the 5' ends. Similarly, the maximum $M^R(n+1,m+1)$ in the trailing submatrix $[n+1,N+1]x[m+1,M+1]$ of S^R, anchors the best local alignment for N-n and M-m nucleotides at the 3' ends:

$$M^L(n,m) = \max\left(S^L(n',m')\right),\ n' \leq n, m' \leq m$$

$$M^R(n,m) = \max\left(S^R(n',m')\right),\ n' \geq n, m' \geq m.$$

The total score of aligning n and m nucleotides at the 5' ends and N-n and M-m nucleotides at the 3' ends is $M^L(n,m)+M^R(n+1,m+1)$. The optimal alignment has the highest score, and thus it maximizes the sum:

$$BS = \max(M^L(n,m)+M^R(n+1,m+1)).$$

Breakpoint revision

After preprocessing, alignment, and structural variation detection, the raw structural variations are obtained. However, due to the limitation of algorithm or the allowed error of the method strategy, the breakpoints obtained may not be the extract location of SV, but the location near the actual breakpoints. Thus, another step is added to precisely confirm the locus of breakpoints.

First, the two breakpoints of candidate SV are extended with upstream and downstream of average insert size. Then, all of the supporting reads that fall into the extended region are extracted. All of the supporting reads are assembled into contig using SOAPdenovo. At last, AGE is used to explore SV in the contig compared to the same region of reference genome. For the AGE output, different

strategies are set for different types of SV due to their characters that filter out the optimal breakpoint.

For deletion structural variation, the following conditions are used to filter out results:

- contig length is greater than or equal to single reads length;
- mismatch ratio is less than 10%

$$\text{mismatch ratio} = \frac{\text{identical aligned bases}}{\text{contig length-unaligned bases}};$$

- if there are still more than one results, keep the one with highest age score.

For insertion structural variation, the following conditions are used to filter out the result:

- contig length is greater than or equal to single reads length;
- aligned bases for the upstream and downstream of the SV are greater than or equal to 5 bases;
- length changed ratio is less than or equal to 50%

$$\text{length changed ratio} = \frac{\text{length changed in reference genome}}{\text{length changed in contig}};$$

- if there are still more than one result, keep the one with highest age score.

For inversion and duplication, the following strategies are used:

- contig length is greater than or equal to single reads length;
- aligned bases for the upstream and downstream of the SV are greater than or equal to 5 bases;
- if there are still more than one results, keep the one with highest age score.

SV Integration

In SV detection phase, two alignment tools and three SV identification tools are used. So for a single sample sequence input, six different SV outputs are obtained. Each detection method outputs multiple types of SV results. So this step is to integrate all the output to get a consensus SV with strong evidences.

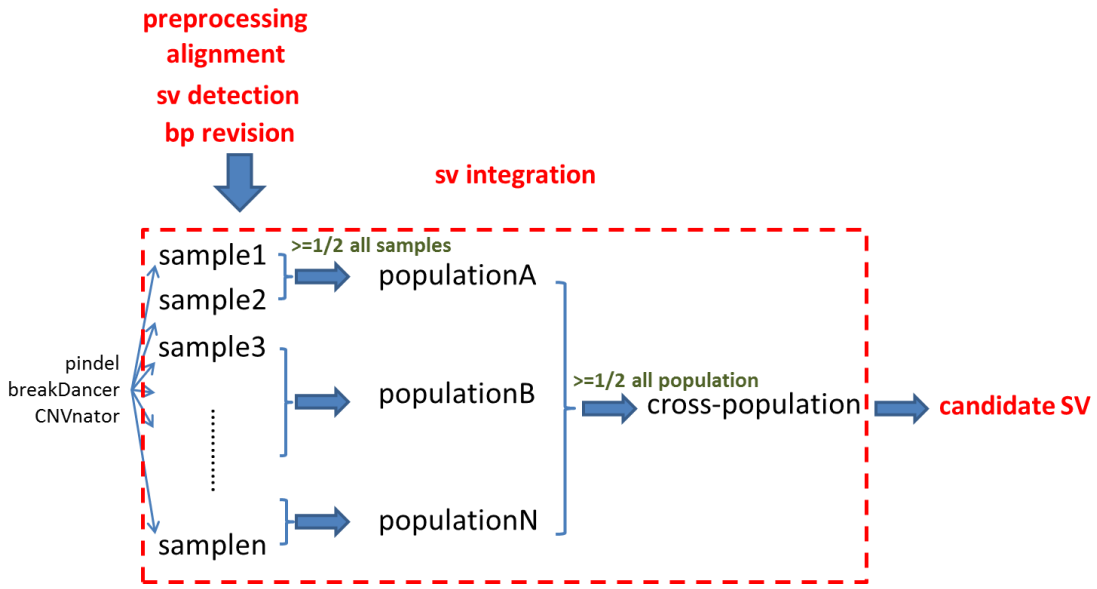


Figure 12. Workflow for consensus SV of deletion.

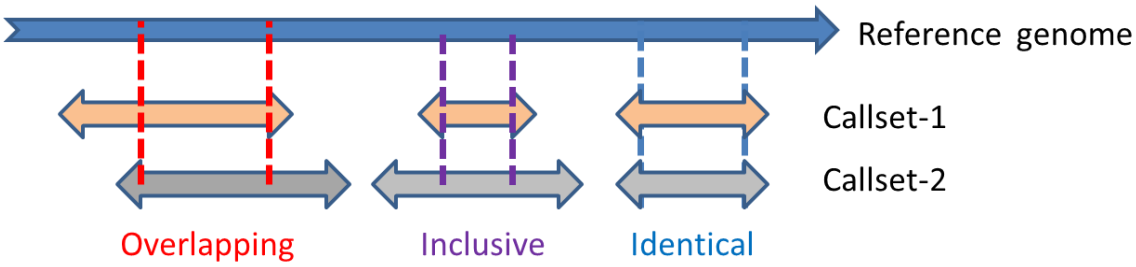


Figure 13. Callset integration.

The workflow of deletion consensus SV is shown in Figure 12. There are three levels in the integration phase: sample level, population level, and cross-population level. First, for the six revised results from three different detection

methods of each sample, identical, inclusive and overlapping SVs are explored. Figure 13 shows the strategy of callset integration. Identical structural variation is two or more SVs that share the same region. For those identical SVs, only one is kept. Inclusive structural variation is one or more small regions that is fully covered by a larger region. For those inclusive SVs, the small region is kept. Overlapping structural variation is when two or more regions have certain length of overlap, but each of them still has its own region. For those overlapping SVs, the overlapped region is kept. Then, the consensus SVs for each sample are obtained. In the population level and cross-population level, the same strategy as the sample level is used. The consensuses SVs from sample level are the inputs of population level integration, while the consensus SVs from population level are the inputs of cross-population level integration. The threshold for the population level and cross-population level are set as half of the input callsets. The population level consensus SVs are from at least half of the number of total samples, and the cross-population level consensus SVs are from at least half of the number of total populations. Finally, consensus SVs from different detection methods, different samples, and different populations are obtained.

For the other three types of SVs, insertion, inversion, and duplication, the consensus SVs are those that shared the same break points in each level based on their characters of structural variation. Then, the SVs with the same regions are combined, and only the SVs from at least half of their resource are kept. From the population level, the consensus SVs, which arefrom at least half of the

total number of samples, are kept. While those from at least half of total number of populations are kept for cross-populaiton level.

Annotation

BLAST(short for Basic Local Alignment Search Tool) is an algorithm to compare two sequences. Due to the type of the compared sequence, there are different types of BLAST.

Table 8

Different BLAST Programs

BLAST Program	Description
Nucleotide blast	Search a nucleotide database using a nucleotide query
Protein blast	Search protein database using a protein query
Blastx	Search protein database using a translated nucleotide query
Tblastn	Search translated nucleotide database using a protein query
Tblastx	Search translated nucleotide database using a translated nucleotide query

In the pipeline, the annotation phase uses BLASTX to execute function annotation, which searches the protein database and detects functions using a translated nucleotide query. Each candidate SV is annotated using its function from the BLASTX search.

Experimental Validation

For all candidate SVs, a biological experiment is designed to validate whether the SV is a true SV. Researchers are interested in those which have a function annotation, since such SV has a high probability that will affect the phenotype of an organism.

After all of the steps, structural variation candidates with function annotation are identified from the input experimental biological dataset. Through the experimental validation, true structural variations are determined.

CHAPTER IV

MIRDISC: A NOVEL COMPUTATIONAL PROGRAM FOR MICRORNA DISCOVERY FROM SHORT DEEP SEQUENCING READS

Introduction

MicroRNAs (miRNA) are a large family of small, non-coding RNAs with an average length of 22 nucleotides that regulate gene expression through near-perfect Watson-Crick pairing to the 3'-untranslated or coding regions (plants only) of target mRNAs (Ambros, 2004; Bartel, 2004; He & Hannon, 2004).

Figure 14 depicts the procedure of miRNA formation. miRNA is not directly transcribed from DNA, but encoded by DNA in nuclear first transcribed under the action of RNA polymerase II. The polymerase often binds a promoter found near the DNA sequence encoding that will become the hairpin loop of the pre-miRNA. The transcript is capped with a specially modified nucleotide at the 5' end and polyadenylated with a pol(A) tail (multiple adenosines). Then, double-stranded pri-miRNA is cut into 70 nt stem-loop intermediate with phosphate group at 5' end and two-nucleotide overhang at the end of 3' through RNA polymerase III Drosha-DGCR complex. Such a resulting sequence is called precursor miRNA (pre-miRNA). Then, the pre-miRNA combines with the transporter protein Exportin-5 and is exported to the cytoplasm by Ran-GTP. Last, polymerase Dicer recognition 5' of terminal phosphate and 3' of overhang from pre-miRNA and cut the double-helix strands at two nucleotides away from the stem loop, resulting in a dimer miRNA: miRNA, whose structure is similar to a dimer siRNA.

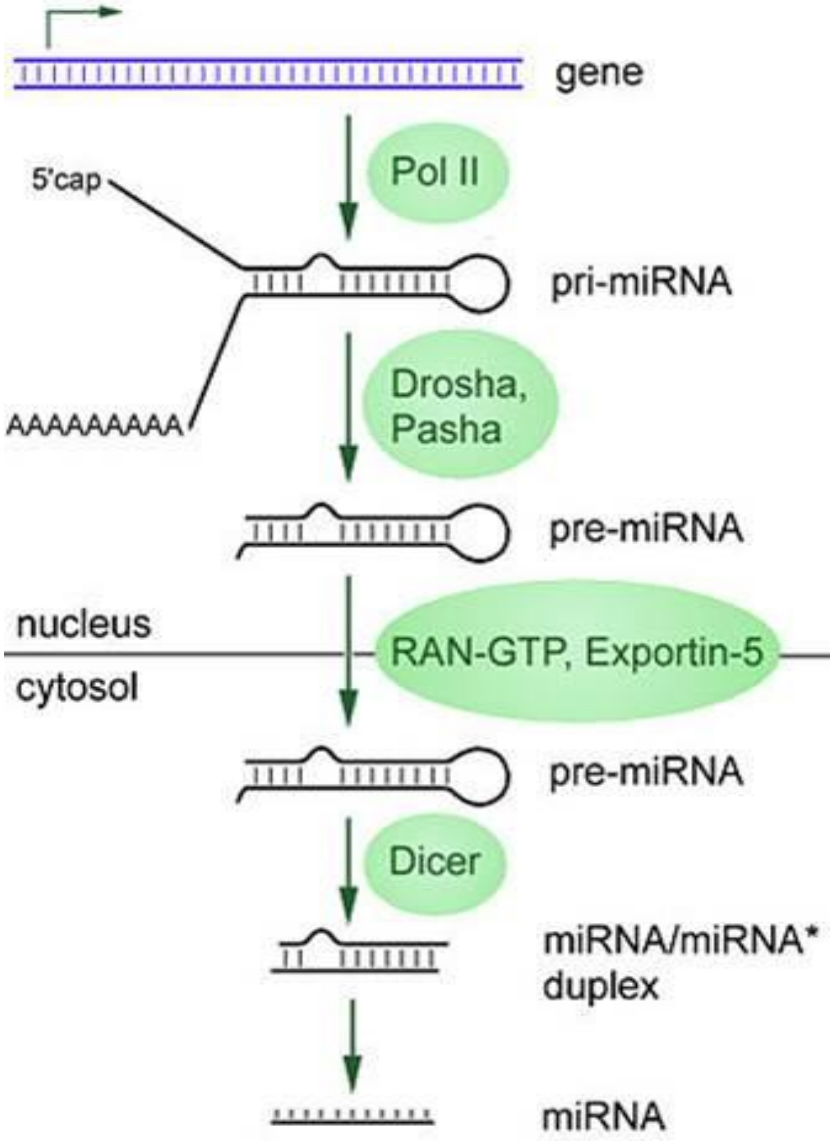


Figure 14. miRNA generation.
http://upload.wikimedia.org/wikipedia/commons/9/95/MiRNA-biogenesis.jpg

In association with miRNA protein effector components, they mediate sequence-specific posttranscriptional and transcriptional gene regulation, and hence control mRNA translation, stability, and localization and feed into a process that controls transposons and heterochromatin structures (Bartel, 2004; He & Hannon, 2004). The discovery and characterization of miRNAs have led to

a rapid expansion of research directed at elucidating their expression patterns and regulatory functions (Maroney, Chamnongpol. Souret, & Nilsen, 2007). It is now clear that miRNAs play important roles in almost all biological processes in eukaryotic organisms, including normal development, cellular response to toxicants and human diseases such as cancer, heart disease, and neurodegenerative disorders (Jiang et al., 2009; Taylor & Gant, 2008; Weinberg & Wood, 2009).

Current Methods

A key part of research involving miRNAs is to identify novel or unknown miRNA in the organism of interest. Since the discovery of the two founding miRNAs lie-4 (Lee, Feinbaum, & Ambros, 1993) and let-7 (Pasquinelli et al., 2000) in *Caenorhabditis elegans* in the 1990s, 21264 hairpin miRNA precursors expressing 25141 mature miRNA products in 193 species have been registered as of August 2012 in Release 19 of miRBase (http://www.mirbase.org/), an online repository of miRNA nomenclature, sequence data, annotation, and target prediction (Griffiths-Jones, 2004; Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006; Griffiths-Jones, Saini, van Dongen, & Enright, 2008). The exponential growth of miRBase entries in the past decade (starting in 2002 with Release 1.0 hosting 218 miRNA precursors in 5 species) has been, to a large degree, attributed to the computational identification of conserved and novel miRNAs. In general, these *in silico* miRNA discovery methods can be divided into two categories.

The first category includes tools that predict mature miRNAs and/or miRNA precursors (pre-miRNA) from genome sequences or cloned sequences of model organisms based on evolutionary sequence conservation and machine learning algorithms (Wu, Wei, Liu, Li, & Rayner, 2011; Yousef, Showe, & Showe, 2009). Examples are phylogenetic shadowing (Berezikov et al., 2005), MiRscan (Lim et al., 2003b; Lim, Glasner, Yekta, Burge, & Bartel, 2003a), MiRseeker (Lai, Tomancak, Williams, & Rubin, 2003), miRAlign (Wang et al., 2005), MirEval (Ritchie, Theodule, & Gautheret, 2008), miRPara (Wu et al., 2011), miRank (Xue et al., 2005), miPred (Jiang et al., 2007) and proMiR II (Nam, Kim, Kim, & Zhang, 2006). The main drawbacks of these tools are either that they are limited to conserved miRNAs and organisms with completed genome sequences, or they tend to have a high rate of false positive and false negative predictions (Hackenberg, Stum, Langenberger, Falcon-Perez, & Aransay, 2009; Hendrix, Levine, & Shi, 2010).

The second category includes programs for miRNA prediction from massive amounts of small RNA reads generated by next-generation deep sequencing technologies such as Illumina/Solexa, 454, SOLiD, and Ion Torrent. Unlike the conventional time-consuming approach of cloning and Sanger sequencing (Bentwich et al., 2005), high throughput sequencing data allows for the detection of more lowly abundant miRNAs with unprecedented sensitivity (Friedlander, Mackowiak, Li, Chen, & Rajewsky, 2012). Methods in this category take miRNA biogenesis into consideration, including miRanalyzer (Hackenberg et al., 2009), miRTRAP (Hendrix et al., 2010), MIReNA (Mathelier & Carbone,

2010), miREAP (Zhai et al., 2011), mirTool (Zhu et al., 2010), miRDeep (Friedlander et al., 2008), and its variants such as miRDeep2 (Friedlaender, Mackowiak, Li, Chen, & Rajewsky, 2012), miRDeep* (An, Lai, Lehman, & Nelson, 2013), miRDeep-P (Yang & Li, 2011), and miRDeepFinder (Xie, Xiao, Chen, Xu, & Zhang, 2012). The core algorithm developed in miRDeep was based on a probabilistic model of miRNA biogenesis to score the compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor (Friedlander et al., 2008). This algorithm has been not only inherited by all miRDeep variants but applied to other comprehensive tools such as mirTool (Zhu et al., 2010), MIReNA (Mathelier & Carbone, 2010), and deepBase (Yang, Shao, Zhou, Chen, & Qu, 2010) with or without modifications.

Despite the existence of such a large variety of computational programs, accurately identifying miRNAs from deep sequenced RNAs remains challenging. The existing algorithms for identification or prediction of miRNAs all rely on the availability of a reference genome, which severely limits their applicability. Given the facts that miRNAs have been reported in fewer than 200 organisms and even fewer animals and plants have had their genomes fully sequenced, there exists a gap in computational tools that detect miRNAs in eukaryotic organisms that only have transcriptomic and small RNA data generated from NGS. In this study, a new tool, called miRDisc (microRNA Discovery) is developed to fill this gap. Based on the miRNA biogenesis principle, a transcriptome provides a better guidance than genome for miRNA discovery.

miRDeep2 is a completely overhauled tool which discovers miRNA genes by analyzing sequenced RNAs and especially identifies both novel and conserved miRNAs with high accuracy in seven specials (Griffiths-Jones, 2004). MIREAP is a tool which can be used to identify both known and novel microRNAs form small RNA libraries deeply sequenced by Solexa/454/Solid technology. The MIREAP algorithm is employed to obtain all candidate precursors with hairpin-like structures that were perfectly mapped by sequencing tags (Berezikov et al., 2005; Yousef et al., 2009). miRanalyzer is a tool for detecting known and predicting novel miRNAs in high-throughput sequencing experiments. The miRanalyzer, including both the web-based interface and the stand alone package works for detecting known miRNAs from miRBase and predicting new miRNAs, especially in 31 species. Although widely used, all these methods have crucial shortcomings. The reference files for the query sequences mapping against to extract the precursors are genomes, which are composed of not only exons but also introns. Since introns are removed by RNA splicing while the final RNA sequences and part of the query sequences may be mapped to the intros sections in the genomes, the precursors extracted from the genomes are not exactly the correct ones, which leads to a low level of performance for identifying miRNAs. Additionally, miRDeep2 and miRanalyzer are explicitly designed for certain species only, without good performance for other species. For all these aforementioned reasons, new software which is capable of overcoming these shortages needs to be developed.

MiRDisc Package

miRDisc is a new method developed to identify miRNAs, especially to identify the miRNAs in transcriptome enriching species.

The workflow of miRDisc which is presented in Figure 15 includes two pipelines for identifying both novel and conserved miRNAs. The steps in the green boxes specify the future experiment work and are excluded from the miRDisc flowchart. The left pipeline without color marked is developed to discover novel and conserved miRNA, respectively, whereas the gray-marked right pipeline disclosures only conserved miRNAs. The results in both pipelines are merged for the experimental validation.

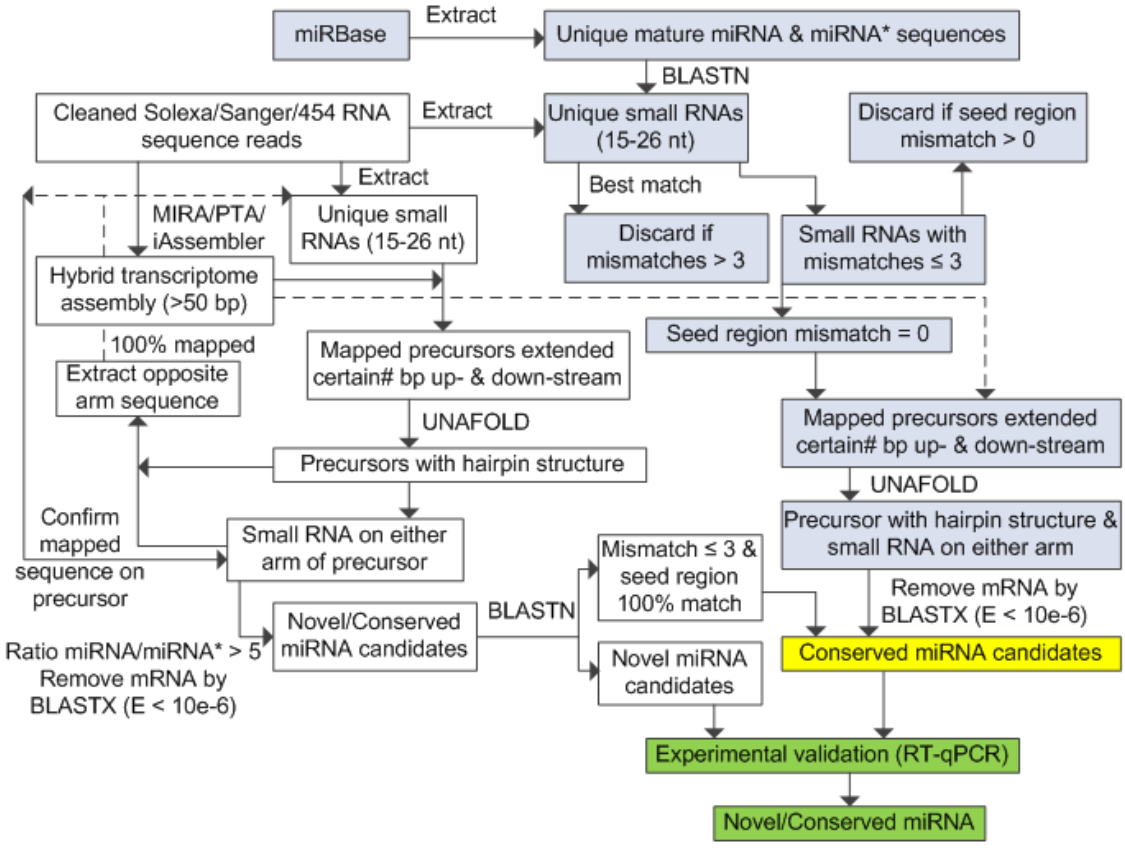


Figure 15. Workflow of miRDisc: the pipeline for discovering both novel and conserved miRNAs.

Generating Transcriptome and Short Sequences

The high throughput sequences are detected by Solexa, 454, and Sanger technologies, and pre-cleaned off the introns. The sequences are then assembled to form transcriptome with widely used assembly tools, MIRA (Lim et al., 2003b), PTA, or iAssembler (Lim et al., 2003a). Since a miRNA molecule has 22 nucleotides in average (Ambros, 2004; He & Hannon, 2004), sequences with length in range of 15~26 are extracted out as candidates for identifying the miRNAs.

Mapping to Extract Precursors and Folding for the Hairpin Structures

Based on the formation process of the miRNA, the short sequences are mapped to the trancriptome and extended up-stream and down-stream for a certain number of nucleotides, respectively, or extended to the end of the transcriptome if not having enough number of nucleotides in transcriptome for extending, and therefore, to extract the precursors (Lai et al., 2003; Wang et al., 2005). Without introns in the transcriptome, the correctness of the precursors extracted from it is guaranteed. Of course, genomes can be also taken as the reference file for query sequences mapping against, but the miRDisc package is emphatically designed to identify the miRNA candidates especially in trancriptome enriching species. According to the precursor data in miRBase, all species can be distributed to five classifications; Metazoa (Ritchie et al., 2008), Chromalveolata, Mycetozoa, Viridiplantae, and Virus, in which the length range varies. According to the statistical result, for instance, the precursors after

extended 60 nucleotide up-stream and down-streams will be in the range which covers 99% of precursors of Metazoa.

Unifies Nucleic Acid Folding (abbreviated as UNAFold) software package integrates a number of programs to simulate the folding process for one or two single-stranded nucleic acid sequences and form a hairpin structures which is composed of two arms and a steam-loop between both arms (Xue et al., 2005). One precursor may be folded in multiple ways and, thus, generates more than one hairpin structures.

Extracting Mature and Star Sequences

After the precursors are folded into the hairpin structures, a complementary sequence of the original query sequence can be found in the hairpin structure. Since mature sequence and the star sequence exist in pair and are located on two arms of the hairpin structure, respectively (Jiang et al., 2007), the complementary sequence located on the steam-loop or broken at some place is out of consideration. In order to avoid finding the complementary for some sequences which actually does not exist by chance, mismatches are restricted. 12 accumulation mismatches and 6 continuous mismatches are preferred as the upper bounds for mismatches between the query sequence and its complementary sequence. Furthermore, for both ends of the query sequences, at least one of the two nucleotides at each end must be complemented.

Not only located on the opposite arms with the query, the complementary should also exist in the original sequence file such that the complementary sequences can be considered as existing in pair with the query sequences. In

order to verify the existence, complementary sequences are mapped back to the query sequence file with no mismatch allowed. Once existence is verified in the original sequence data, the sequences which include or are equal to the complementary sequences are extracted out from the original sequence data. And then, the extracted sequences are mapped back to the precursors to double check whether these extracted sequences are still complementary to the query sequences or not.

Based on the verification results above, the mature sequences and the star sequences have been found out (Jiang et al., 2007; Nam et al., 2006), but the star sequences which have lower level stability than those of the mature sequences are preferentially degraded. Therefore, the mature sequences in the original sequences data have a greater copy number than that of the star sequences. Again, in order to avoid mistaking the sequences existing in pair by chance in the original sequences data as the mature and star sequences, the ratio between the paired sequences with copy number are restricted in a certain range, such as less than or equal to 1/6 or greater than or equal to 6 are discarded. The sequences with greater copy number in the remaining paired sequences are the mature sequences.

Eliminating Coding RNAs and Distinguishing the Novel and Conserved miRNAs

BLASTX (Hackenberg et al., 2009; Hendrix et al., 2010) is introduced for aligning the mature sequences to the NCBI database. The sequences with high possibilities translated to protein are eliminated, and the E-value here are restricted as 10e-6.

After applying the BLASTX, the remaining sequences are miRNA candidates, including both the novel and the conserved miRNA candidates. The miRBase (Bentwich et al., 2005) database provides the latest released miRNAs in various species. To separate the novel and conserved miRNA candidates, the non-coding sequences after the BLASTX process are mapped to the published miRNA in the miRBase database. In order to increase the accuracy of novel and conserved miRNA identification, the perfect match in seed region of the remaining sequences and the published miRNA sequences are required (Friedlander et al., 2012; Mathelier & Carbone, 2010; Zhai et al., 2011). For animal species, the seed region encompasses the $5'$ bases 2-7 of the miRNA, including 7 nucleotides. Besides, the number of total mismatches between the query sequence and known miRNAs are restricted to a certain range, say 3 for instance. The sequences with seed region perfectly matched with the published miRNAs and with E-value less than or equal to 10 are considered as the conserved miRNA candidates, whereas other sequences are considered as the novel miRNA candidates.

The Pipeline for Discovering Only Conserved miRNAs

This pipeline is shown as the right part of the flow chart in Figure 15 and some procedures are similar as the left pipeline.

Align Sequences to the miRBase Database

The input query sequences are selected from the Solex, 454, and Sanger detected sequences, with length range of 15~26. The miRBase database is introduced again here to map the query sequences to the published miRNAs,

and the seed region and total mismatch are restricted to ensure the accuracy of mapping. The sequences with mismatches exceeding the restriction and with imperfect match are explicitly not taken as the conserved sequences and therefore discarded.

Mapping to Extract Precursors and Folder for the Hairpin Structure

As specified in the left pipeline, the short sequence are mapped to the transcriptome and extended up-stream and down-stream to extract the precursors, and the extending length is also taken into consideration inevitably here. After folded to the hairpin structure with UNAFold software package, the location of the sequences in the hairpin structure should be checked to ensure that they are located on either arm but not the stem-loop. The sequences locating on the stem-loop or complementing with themselves are excluded.

Rejecting the Coding Sequences

The BLASTX is also applied in the right pipeline to reject the coding sequences with E-value of 10e-6. The remaining sequences after BLASTX process are considered as the conserved miRNA candidates.

The results from both pipelines are then merged together to two files: one for novel miRNA candidates and the other for conserved miRNA candidates. Both the novel and conserved miRNA candidates identified by miRDisc need to be further validated by the experiments, for example, RT_aPCR (Friedlander et al., 2008; Zhu et al., 2010), as green-marked in the work flow in Figure 15.

CHAPTER V

NGS DATA ANALYSIS: CASE STUDY

Data Analysis Using SeqAssist

The crustacean genus *Daphnia*, a sentinel sensitive to many toxicants, is used for monitoring and assessing the ecological impact and for establishing regulatory criteria by government agencies, such as the U.S. Environmental Protection Agency, that regulates Army sites. Numerous studies have demonstrated significant variation in chemical sensitivity among natural and laboratory-raised populations. Moreover, various researchers have observed temporal sensitivity drifting to heavy metals in laboratory strains where genetic impoverishment is caused by isolation, inbreeding, and artificial selection. These two types of variations cause a problematic interpretation of the chemical effect levels measured in inter- or intra- laboratory comparisons and require the introduction of uncertainty factors in evaluation of Army sites. The researchers are interested in identifying and discovering how the genotype affects the phenotype for the *Daphnia* species. Understanding this fact is helpful to clarify differences in chemical sensitivity between populations of a single model species to explain variability in toxicological experiments and reduce uncertainty in evaluation of contaminated Army sites.

The genus *Daphnia* is ideal for use as a biological model to study phenotypic plasticity as it is able to adapt physiologically to wide ranges of pH, toxins, oxygen concentrations, food, and temperature. It has a short life cycle (30~100 days), short generation time (>= 6 days), and a unique reproductive

strategy which add to the strengths of this model to investigate phenotypic plasticity across many generations.

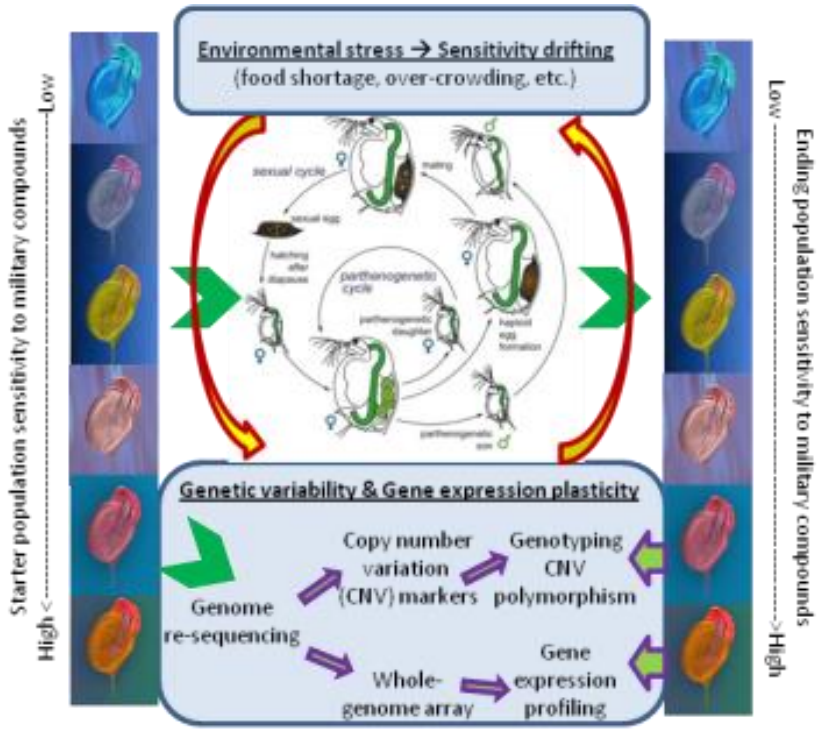


Figure 16. Changes in genetic variation in the phenotype using 8 different populations of varying chemical sensitivity.

In order to investigate the relationship between genotype and phenotype, the author choose several *Daphnia* populations from different locations and several individuals for each population to analyze the structural variations (shows in Figure 16). First, the dataset set of sequences is analyzed by SeqAssist to discover the basic statistical information, which can be used to measure the quality of the dataset. Then, the SVDisc pipeline is applied to the dataset to discover structural variations for each individual, each population, and different populations. Last, the genetic information, structural variations, phenotypic information, and the chemical sensitivity are integrated together to determine the relationship.

Experimental Design and Dataset Generation

The following experiment is designed by Dr. Ping Gong at the Environmental Laboratory of U.S. Army Engineer Research and Development Center. *Daphnia pulex* is obtained from multiple sources. Table 9 shows basic information for all the experimental populations, including their experimental population code, type (lab or field), source, acquisition time, and whether it is a selected population. Sources include sustained laboratory cultures and recently collected natural populations to acquire a diversity of clones. Gravid females will be cultured at ERDC, and neonates obtained from each culture (but not each cultures themselves) are subjected to acute and chronic chemical exposures to determine their relative sensitivity, by an initial assessment of intra-treatment variability using multiple clones within each population. Then 8 cultures with a gradient of sensitivities measured by endpoints for genome re-sequencing are selected. A clone from each of the 8 populations is used for further analysis.

Table 9

Basic Information for All Testing Population

#	Population Code	Type	Source	Acquisition	Selected?
1	ECT	Lab	EPA-ORD	Mar-12	Y
2	TCO	Lab	Canada	Aug-12	Y
3	HSL	Field	Missouri	Sep-12	Y
4	STL	Field	St. Louis, MO	Mar-12	Y
5	CA2	Field	UK	Dec-12	Y
6	W3.6A	Field	UK	Dec-12	Y

Table 9 (continued).

#	Population Code	Type	Source	Acquisition	Selected?
7	ABS	Lab	Colorado	Dec-11	Y
8	BEL	Field	Belgium	Dec-11	Y
9	SRL	Field	Canada	Aug-12	N
10	BEL	Field	Belgium	Oct-11	N
11	IL	Field	Champaign, IL	Jun-12	N
12	ABS	Lab	EPA-ORD	Dec-11	N
13	MI	Field	Houghton, MI	Nov-11	N
14	NY	Field	New York	Oct-11	N
15	LD3.24	Field	UK	Dec-12	N
16	LD3.2	Field	UK	Dec-12	N
17	D8.7A	Field	UK	Dec-12	N
18	D8.4A	Field	UK	Dec-12	N
19	W1.7A	Field	UK	Dec-12	N

Daphnia Culturing and Sensitivity Screening

Daphnia pulex is cultured under ideal (Culture A) and stressful (Culture B) conditions as summarized in Figure 17. Universal conditions include the use of reconstituted hard water as the culture medium, a 16-h light: 8-h dark photoperiod at 23±1℃ and a feeding ration of 1:1 green algae (Selenastrum capricornutum), and yeast-cereal leaves-trout chow (YCT). In the “ideal” laboratory culture condition, females from each population will be maintained

under optimized conditions to sustain parthenogenesis, or asexually reproducing female clones (sexual reproduction will not be allowed). Such conditions include maintaining algae at 2.3×10^5 cells/ml and a density of adult female clones of less than 15 individuals per liter culture medium. In the “stressful” condition, females originating from the same population will be subjected to overcrowding stress (e.g., >30 adult females per liter of culture medium) which will induce brooding of males, sexual reproduction, and ephippium.

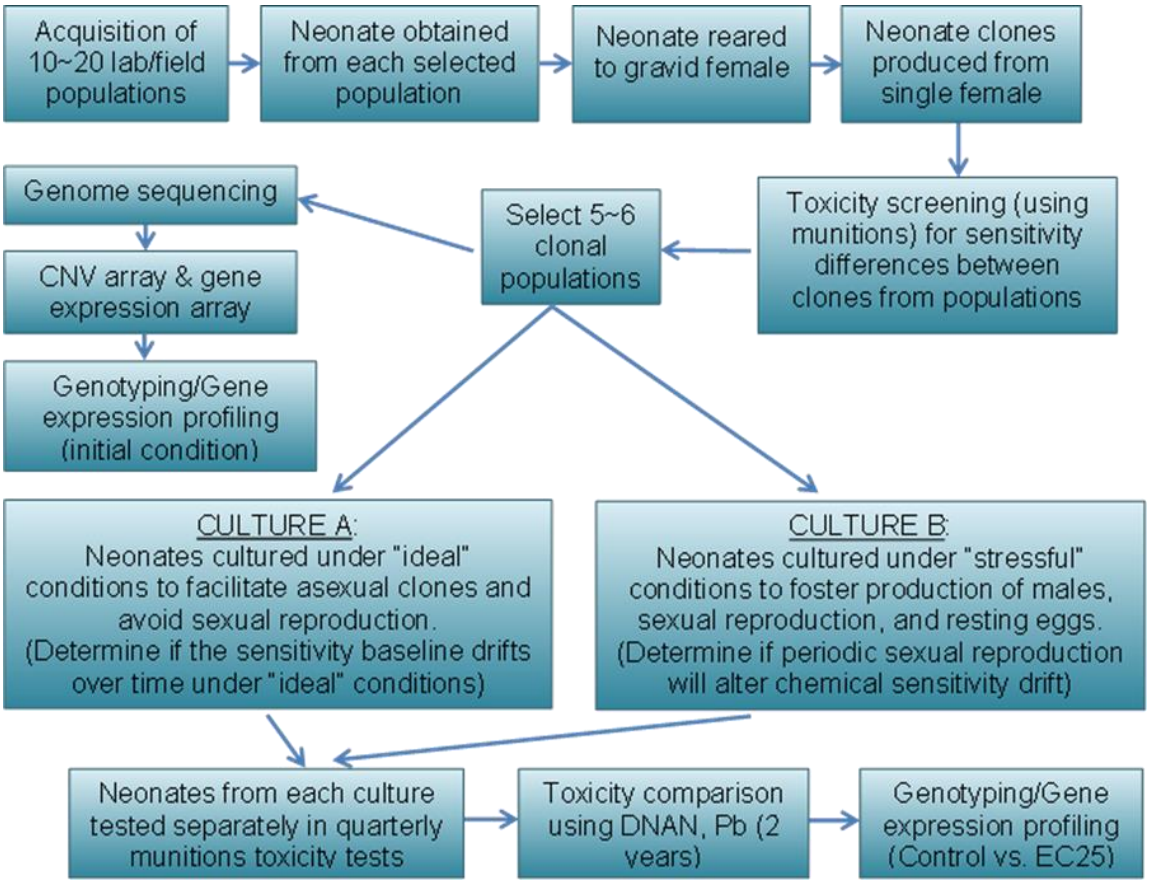
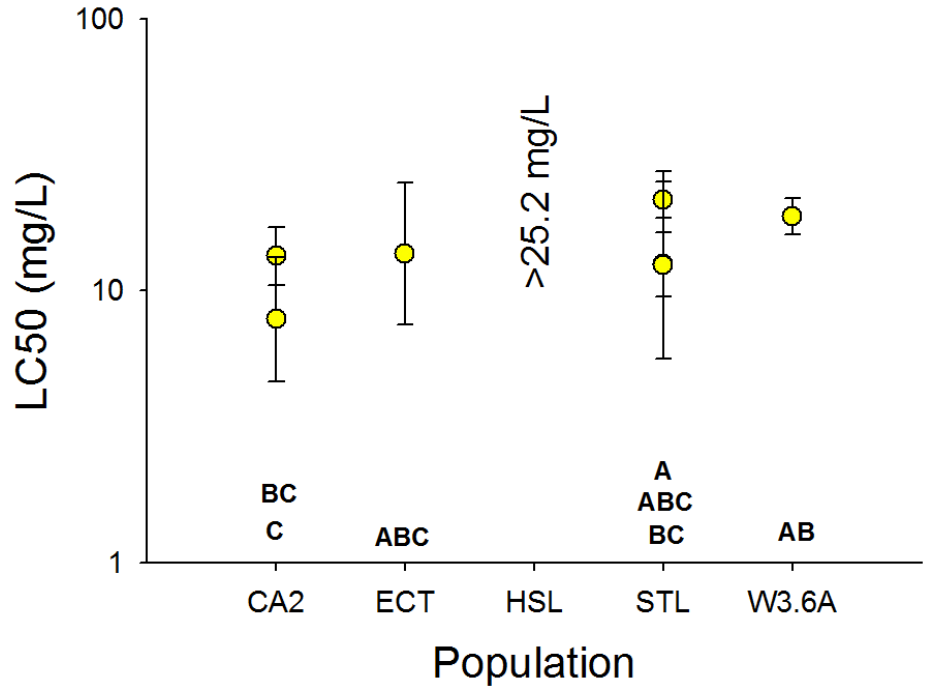


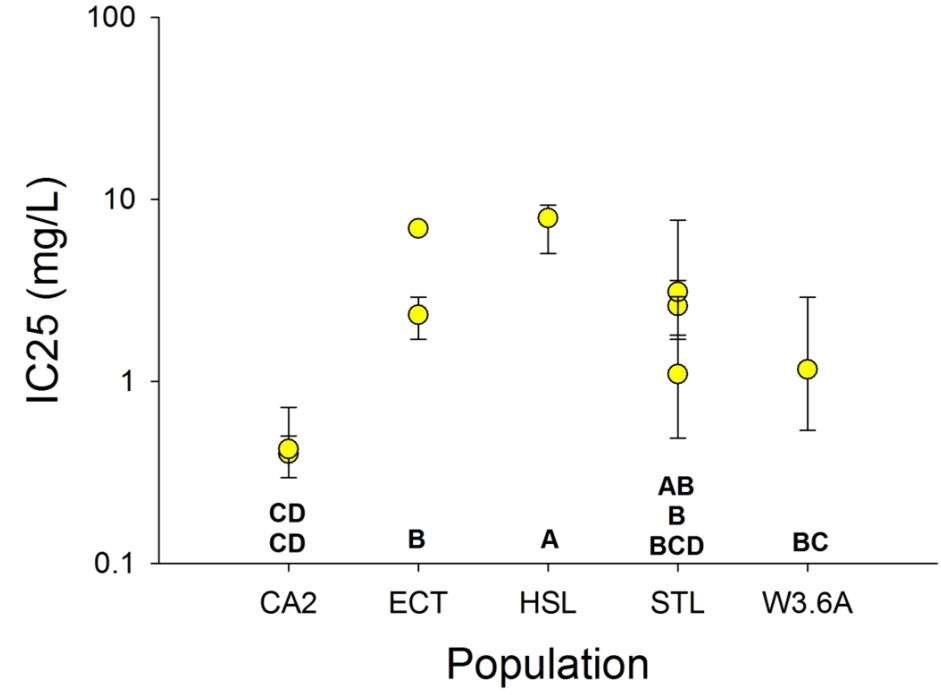
Figure 17. Overview of the procedure to initiate cultures from different *Daphnia pulex* populations and the conduct of toxicity screening for determining differences in chemical sensitivity to support the proposed objectives.

Ephippia are allowed to settle to the bottom of the culture vessel and hatch sexually reproduced *Daphnia pulex* to contribute to the further generations

of the population. Other stresses such as lower food rations and temperature alterations are not employed as they are known by this research group to reduce reproductive output and change sensitivity to chemicals, respectively. For each *D. pulex* culture, quarterly acute lethality (48-hour) and chronic reproduction and growth (21-days) experiments will be conducted by exposing the neonates obtained from cultures to five concentrations of munitions compound (DANA and Pb) in accordance with nationally recommended guidance, with consideration to munitions handling. Figure 18 shows the effect of DANA to the reproduction of several populations. From those two figures, it can be seen that different populations show different sensitivities to the DANA. Some populations have a weak tolerance to the chemical, and half of individuals are killed at very low concentration (CA2 and W3.6A). While other populations show good tolerance to the DNAN, e.g., for HSL, all the individuals are alive when the concentration is as high as 25.2mg/l in the LC50 test.



(a)



(b)

Figure 18. Chemical sensitivity test.

Genome Re-sequencing

The isoclinal animals from each of the 8 starter strains are reared to large numbers in filtered cultures medium, and then treated with 500mg/L of Tetracycline to reduce bacterial contamination and with 4.5 micron copolymer microsphere beads (Duke Scientific cat# 7505A) to clear the gut. High molecular weight DNA is isolated by Genomic-tips using the manufacturer's protocol for animal tissues (Qiagen). The genomic DNA is further sheared using the TruSeq DNA Sample Prep Kits (Illumina) to prepare DNA libraries with insert sizes from 300-500 bp for paired-end sequencing on the Illumina/Solexa MiSeq system.

Dataset Generation

In the genome re-sequencing process, some runs just cover one population, while some runs contain several populations. Furthermore, in order to obtain higher coverage depth and coverage breadth, the same library of one population are sequenced for multiple times in different runs. Table 10 lists all the sequenced populations and all samples for each population. Table 11 describes all the sequencing runs with samples in each run and the average insert size for each run.

Table 10

Daphnia pulex Populations and Samples

Population	Samples
ABS=A	A3, A7, A13(non-pooled), A14, A28(non-pooled)
BEL=B	B1, B2(non-pooled), B9, B13, B16(non-pooled)

Table 10 (continued).

Population	Samples
ECT=E	ECT1, ECT2, ECT3, ECT4, ECT5, E7(non-pooled), E7_rerun(non-pooled), E12(non-pooled)
STL=SL	SL1, SL2, SL3, SL4, SL5
TCO	TCO1, TCO2, TCO3, TCO4, TCO5
HSL	HSL1, HSL2, HSL3, HSL4, HSL5
CA2	CA2_1, CA2_2, CA2_3, CA2_4, CA2_5
W3.6A=W	W1, W2, W3, W4, W5

Table 11

Daphnia pulex Sequence Runs

MiSeq Run	Type	Library/Sample	Insert Size
1	Single	A28	2×148
2	Single	A13	2×151
3	Single	E12	2×151
4	Single	B16	2×151
5	Single	B2	2×151
6	Single	E7	2×151
7	Single	E7(rerun)	2×151
8	Pooled(36)	all pooled samples	2×151

Table 11 (continued).

MiSeq Run	Type	Library/Sample	Insert Size
9	Pooled(36)	all pooled samples	2×151
10	Pooled(36)	all pooled samples	2×151
11	Pooled(36)	all pooled samples	2×251
12	Pooled(36)	all pooled samples	2×251
13	Pooled(36)	all pooled samples	2×251
14	Pooled(6)	SL1,ECT2,TCO2,TCO5,W3,CA2_5	2×251
15	Pooled(6)	SL2,ECT3,TCO3,TCO4,HSL1,A14	2×251
16	Pooled(6)	SL3,ECT4,W4,W5,HSL2,A7	2×251
17	Pooled(6)	SL4,ECT5,W2,CA2_1,B1,B9	2×251
18	Pooled(6)	SL5,HSL3,HSL5,CA2_4,B13,A3	2×251
19	Pooled(6)	ECT1,TCO1,HSL4,W1,CA2_2,CA2_3	2×251
20	Pooled(36)	all pooled samples	2×151
21	Pooled(36)	all pooled samples	2×151
22	Pooled(36)	all pooled samples	2×151

As shown in Table 11, there are total 22 runs, including 8 populations and 42 individual samples. Among these runs, 7 runs are sequenced only one individual sample, 6 runs are sequenced pooled sample, which contain 6 individual samples, and 9 runs are sequenced pooled sample, which contain 36

individual samples. Here, E7 is sequenced twice (E7 and E7_rerun) using the same library in order to validate the quality of the experimental design.

Statistical Analysis

Preprocessing

Prior to the data analysis, first N-containing reads are removed in pairs. In other words, the paired-end reads are discarded as long as there is N in any end. Table 12 shows the number of reads before and after cleaning N-containing reads, and the cleaned percentage based on different populations. From the table, it can be seen that the N-containing reads occupy a small portion of the dataset. All of the cleaned percentages are in the range between 1% and 2%.

Table 12

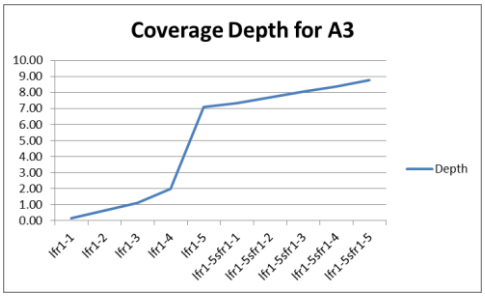
Summary of Preprocessing Result for Different Population

Population	Raw Reads	Cleaned Reads	Cleaned Percentage
ABE	32,560,089	31,991,979	1.745%
BEL	31,598,909	31,147,005	1.430%
CA2	31,302,336	30,852,836	1.436%
ECT	49,194,373	48,566,971	1.275%
HSL	32,222,297	31,798,828	1.314%
SL	27,829,938	27,462,408	1.321%
TCO	36,354,251	35,821,735	1.465%
W3.6A	31,833,885	31,402,360	1.356%

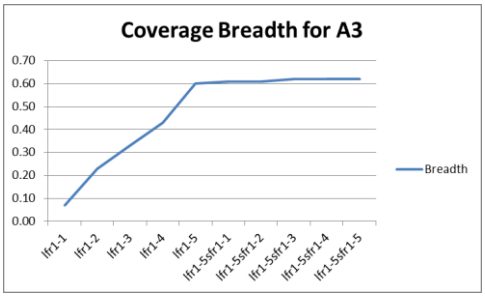
Analysis of coverage depth, coverage breadth, and evenness

After preprocessing, SeqAssist Run2Ref is applied to all the dataset, including 8 populations and multiple individuals for each population. The SeqAssist Run2Ref first aligns the reads to reference genome, and then calculates the coverage depth, coverage breadth, and evenness based on the alignment result. Since there are multiple runs for each individual, reseachers would like to see how the additional run affects the statistical result for the individual. Then, the following strategy is set to each individual: first time, the Run2Ref is applied to the first run, then the second run is added to the dataset and Run2Ref is performed to the combined dataset of the first run and second run. The following process follows the same strategy. Each time, one additional run is added to the dataset to test the statistical status for the combined dataset. All of the results are collected together and for each individual; a plot is drawn to depict the changing trend of coverage depth, coverage breadth, and evenness as different runs, separately. All of the statistical results and plots are shown in Figure 19 - 26.

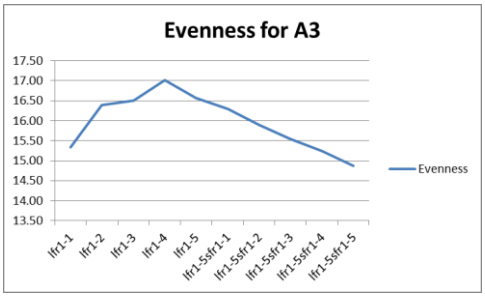
Run	Depth
lfr1-1	0.15
lfr1-2	0.65
lfr1-3	1.12
lfr1-4	1.98
lfr1-5	7.11
lfr1-5sfr1-1	7.33
lfr1-5sfr1-2	7.72
lfr1-5sfr1-3	8.08
lfr1-5sfr1-4	8.38
lfr1-5sfr1-5	8.80



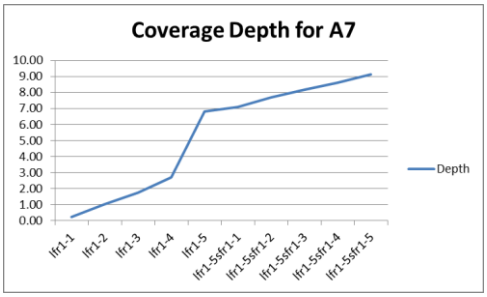
Run	Breadth
lfr1-1	0.07
lfr1-2	0.23
lfr1-3	0.33
lfr1-4	0.43
lfr1-5	0.60
lfr1-5sfr1-1	0.61
lfr1-5sfr1-2	0.61
lfr1-5sfr1-3	0.62
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.62



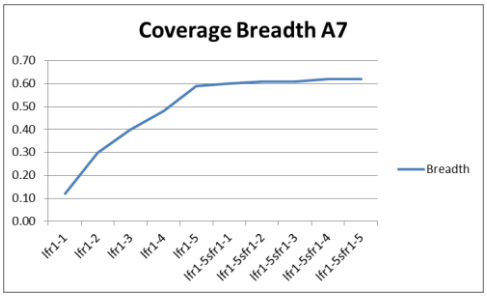
Run	Evenness
lfr1-1	15.34
lfr1-2	16.39
lfr1-3	16.51
lfr1-4	17.01
lfr1-5	16.57
lfr1-5sfr1-1	16.30
lfr1-5sfr1-2	15.89
lfr1-5sfr1-3	15.54
lfr1-5sfr1-4	15.24
lfr1-5sfr1-5	14.88



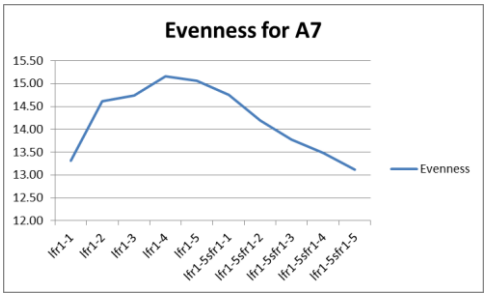
Run	Depth
lfr1-1	0.25
lfr1-2	1.03
lfr1-3	1.76
lfr1-4	2.72
lfr1-5	6.85
lfr1-5sfr1-1	7.12
lfr1-5sfr1-2	7.72
lfr1-5sfr1-3	8.20
lfr1-5sfr1-4	8.61
lfr1-5sfr1-5	9.15



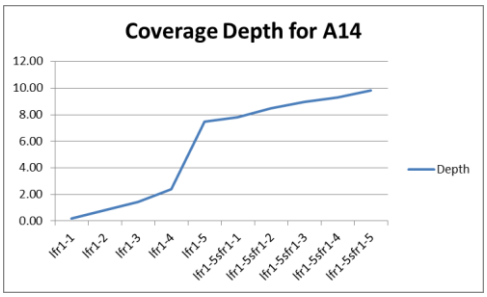
Run	Breadth
lfr1-1	0.12
lfr1-2	0.30
lfr1-3	0.40
lfr1-4	0.48
lfr1-5	0.59
lfr1-5sfr1-1	0.60
lfr1-5sfr1-2	0.61
lfr1-5sfr1-3	0.61
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.62



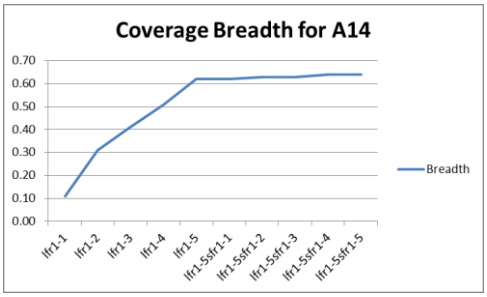
Run	Evenness
lfr1-1	13.32
lfr1-2	14.61
lfr1-3	14.74
lfr1-4	15.16
lfr1-5	15.06
lfr1-5sfr1-1	14.76
lfr1-5sfr1-2	14.19
lfr1-5sfr1-3	13.78
lfr1-5sfr1-4	13.48
lfr1-5sfr1-5	13.12



Run	Depth
lfr1-1	0.19
lfr1-2	0.83
lfr1-3	1.42
lfr1-4	2.39
lfr1-5	7.47
lfr1-5sfr1-1	7.81
lfr1-5sfr1-2	8.49
lfr1-5sfr1-3	8.94
lfr1-5sfr1-4	9.32
lfr1-5sfr1-5	9.85



Run	Breadth
lfr1-1	0.11
lfr1-2	0.31
lfr1-3	0.41
lfr1-4	0.51
lfr1-5	0.62
lfr1-5sfr1-1	0.62
lfr1-5sfr1-2	0.63
lfr1-5sfr1-3	0.63
lfr1-5sfr1-4	0.64
lfr1-5sfr1-5	0.64



Run	Evenness
lfr1-1	12.24
lfr1-2	12.78
lfr1-3	12.91
lfr1-4	13.29
lfr1-5	13.29
lfr1-5sfr1-1	13.00
lfr1-5sfr1-2	12.51
lfr1-5sfr1-3	12.24
lfr1-5sfr1-4	12.03
lfr1-5sfr1-5	11.75

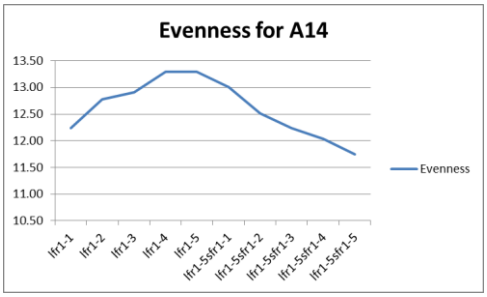
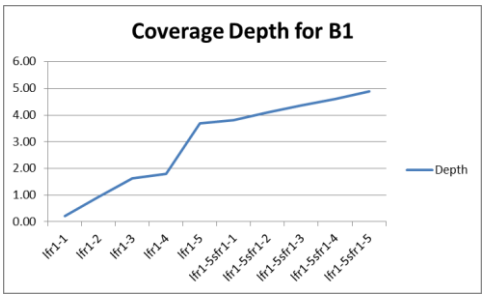
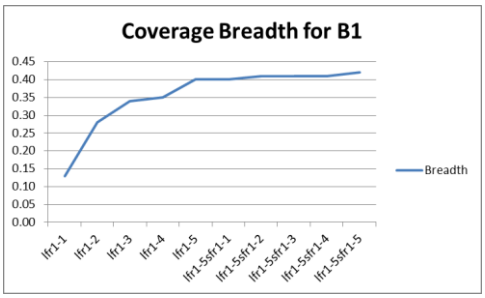


Figure 19. Distribution of coverage depth, coverage breadth, and evenness for population ABE.

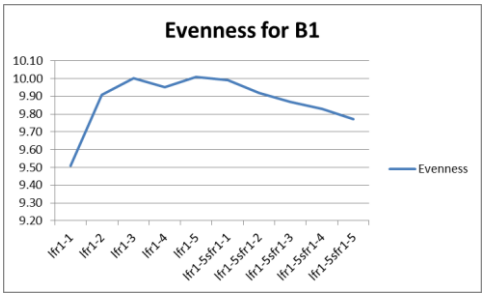
Run	Depth
lfr1-1	0.22
lfr1-2	0.94
lfr1-3	1.62
lfr1-4	1.80
lfr1-5	3.70
lfr1-5sfr1-1	3.82
lfr1-5sfr1-2	4.09
lfr1-5sfr1-3	4.35
lfr1-5sfr1-4	4.59
lfr1-5sfr1-5	4.88



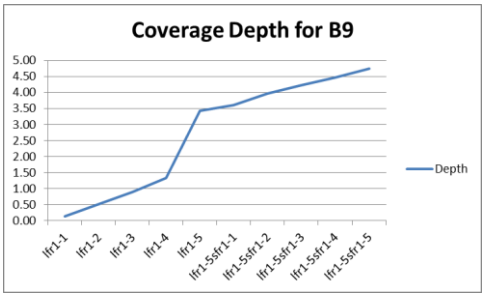
Run	Breadth
lfr1-1	0.13
lfr1-2	0.28
lfr1-3	0.34
lfr1-4	0.35
lfr1-5	0.40
lfr1-5sfr1-1	0.40
lfr1-5sfr1-2	0.41
lfr1-5sfr1-3	0.41
lfr1-5sfr1-4	0.41
lfr1-5sfr1-5	0.42



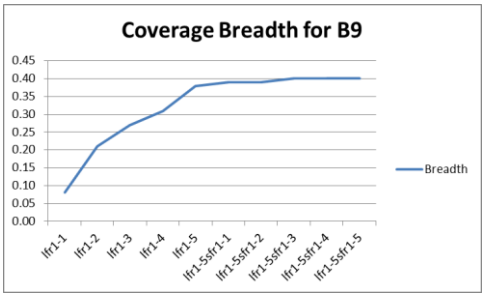
Run	Evenness
lfr1-1	9.51
lfr1-2	9.91
lfr1-3	10.00
lfr1-4	9.95
lfr1-5	10.01
lfr1-5sfr1-1	9.99
lfr1-5sfr1-2	9.92
lfr1-5sfr1-3	9.87
lfr1-5sfr1-4	9.83
lfr1-5sfr1-5	9.77



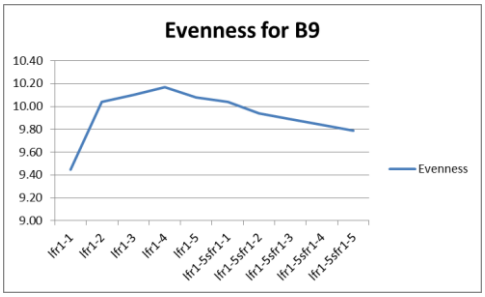
Run	Depth
lfr1-1	0.13
lfr1-2	0.52
lfr1-3	0.89
lfr1-4	1.33
lfr1-5	3.44
lfr1-5sfr1-1	3.62
lfr1-5sfr1-2	3.97
lfr1-5sfr1-3	4.23
lfr1-5sfr1-4	4.47
lfr1-5sfr1-5	4.76



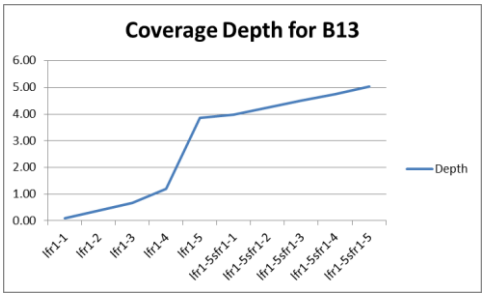
Run	Breadth
lfr1-1	0.08
lfr1-2	0.21
lfr1-3	0.27
lfr1-4	0.31
lfr1-5	0.38
lfr1-5sfr1-1	0.39
lfr1-5sfr1-2	0.39
lfr1-5sfr1-3	0.40
lfr1-5sfr1-4	0.40
lfr1-5sfr1-5	0.40



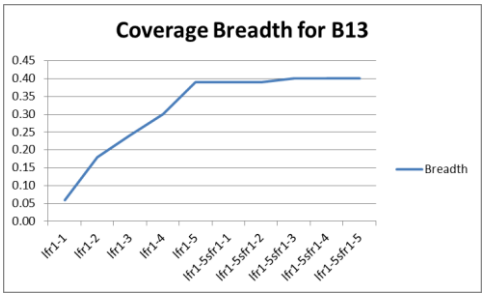
Run	Evenness
lfr1-1	9.45
lfr1-2	10.04
lfr1-3	10.10
lfr1-4	10.17
lfr1-5	10.08
lfr1-5sfr1-1	10.04
lfr1-5sfr1-2	9.94
lfr1-5sfr1-3	9.89
lfr1-5sfr1-4	9.84
lfr1-5sfr1-5	9.79



Run	Depth
lfr1-1	0.10
lfr1-2	0.39
lfr1-3	0.67
lfr1-4	1.20
lfr1-5	3.85
lfr1-5sfr1-1	3.99
lfr1-5sfr1-2	4.25
lfr1-5sfr1-3	4.50
lfr1-5sfr1-4	4.74
lfr1-5sfr1-5	5.03



Run	Breadth
lfr1-1	0.06
lfr1-2	0.18
lfr1-3	0.24
lfr1-4	0.30
lfr1-5	0.39
lfr1-5sfr1-1	0.39
lfr1-5sfr1-2	0.39
lfr1-5sfr1-3	0.40
lfr1-5sfr1-4	0.40
lfr1-5sfr1-5	0.40



Run	Evenness
lfr1-1	9.31
lfr1-2	9.65
lfr1-3	9.80
lfr1-4	9.85
lfr1-5	9.95
lfr1-5sfr1-1	9.90
lfr1-5sfr1-2	9.80
lfr1-5sfr1-3	9.76
lfr1-5sfr1-4	9.70
lfr1-5sfr1-5	9.65

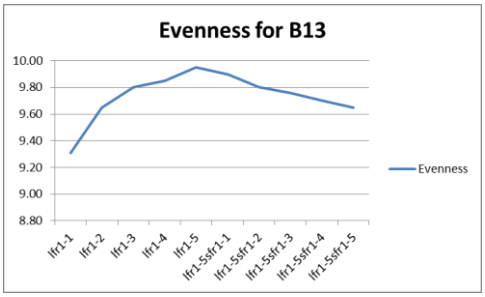
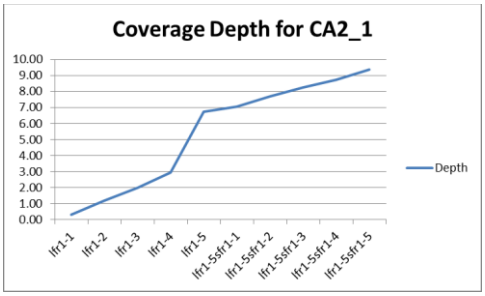
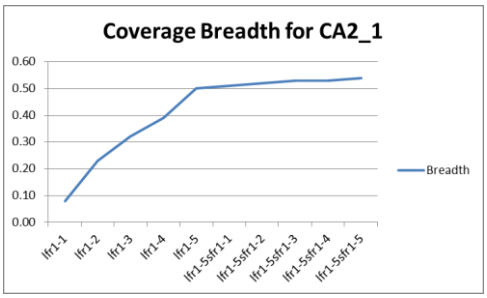


Figure20. Distribution of coverage depth, coverage breadth, and evenness for population BEL.

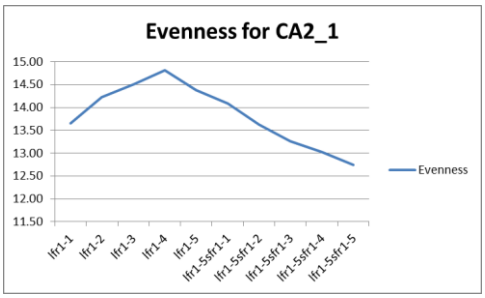
Run	Depth
lfr1-1	0.30
lfr1-2	1.18
lfr1-3	2.00
lfr1-4	2.97
lfr1-5	6.76
lfr1-5sfr1-1	7.08
lfr1-5sfr1-2	7.71
lfr1-5sfr1-3	8.27
lfr1-5sfr1-4	8.75
lfr1-5sfr1-5	9.37



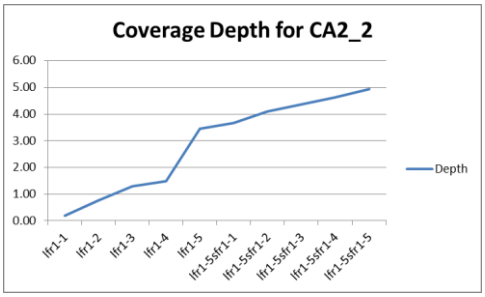
Run	Breadth
lfr1-1	0.08
lfr1-2	0.23
lfr1-3	0.32
lfr1-4	0.39
lfr1-5	0.50
lfr1-5sfr1-1	0.51
lfr1-5sfr1-2	0.52
lfr1-5sfr1-3	0.53
lfr1-5sfr1-4	0.53
lfr1-5sfr1-5	0.54



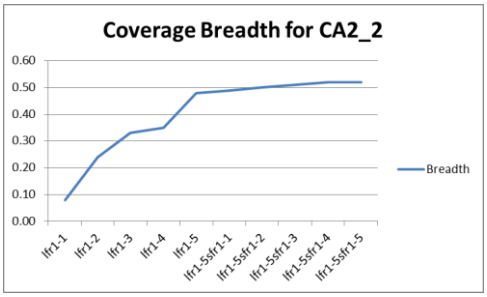
Run	Evenness
lfr1-1	13.65
lfr1-2	14.23
lfr1-3	14.51
lfr1-4	14.81
lfr1-5	14.38
lfr1-5sfr1-1	14.08
lfr1-5sfr1-2	13.63
lfr1-5sfr1-3	13.26
lfr1-5sfr1-4	13.02
lfr1-5sfr1-5	12.75



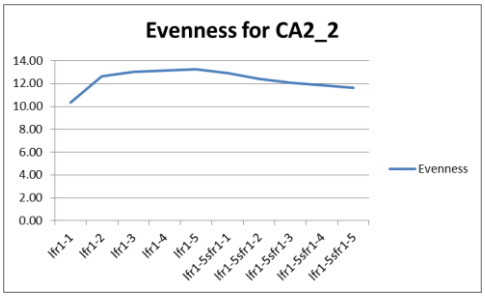
Run	Depth
lfr1-1	0.19
lfr1-2	0.76
lfr1-3	1.29
lfr1-4	1.49
lfr1-5	3.46
lfr1-5sfr1-1	3.66
lfr1-5sfr1-2	4.09
lfr1-5sfr1-3	4.37
lfr1-5sfr1-4	4.62
lfr1-5sfr1-5	4.93



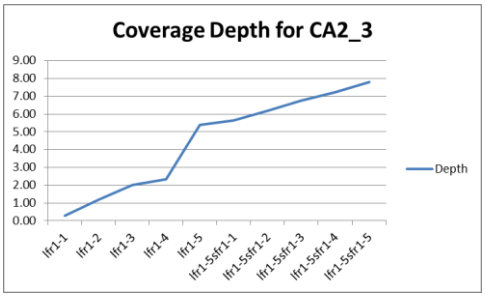
Run	Breadth
lfr1-1	0.08
lfr1-2	0.24
lfr1-3	0.33
lfr1-4	0.35
lfr1-5	0.48
lfr1-5sfr1-1	0.49
lfr1-5sfr1-2	0.50
lfr1-5sfr1-3	0.51
lfr1-5sfr1-4	0.52
lfr1-5sfr1-5	0.52



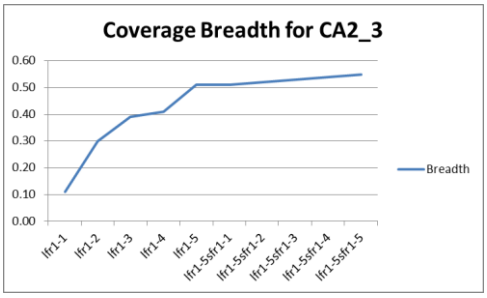
Run	Evenness
lfr1-1	10.37
lfr1-2	12.63
lfr1-3	13.04
lfr1-4	13.13
lfr1-5	13.23
lfr1-5sfr1-1	12.91
lfr1-5sfr1-2	12.40
lfr1-5sfr1-3	12.07
lfr1-5sfr1-4	11.88
lfr1-5sfr1-5	11.64



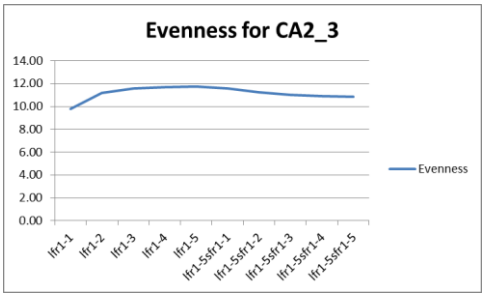
Run	Depth
lfr1-1	0.29
lfr1-2	1.19
lfr1-3	2.02
lfr1-4	2.33
lfr1-5	5.38
lfr1-5sfr1-1	5.65
lfr1-5sfr1-2	6.20
lfr1-5sfr1-3	6.75
lfr1-5sfr1-4	7.21
lfr1-5sfr1-5	7.81



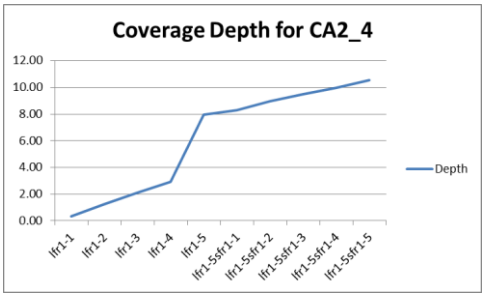
Run	Breadth
lfr1-1	0.11
lfr1-2	0.30
lfr1-3	0.39
lfr1-4	0.41
lfr1-5	0.51
lfr1-5sfr1-1	0.51
lfr1-5sfr1-2	0.52
lfr1-5sfr1-3	0.53
lfr1-5sfr1-4	0.54
lfr1-5sfr1-5	0.55



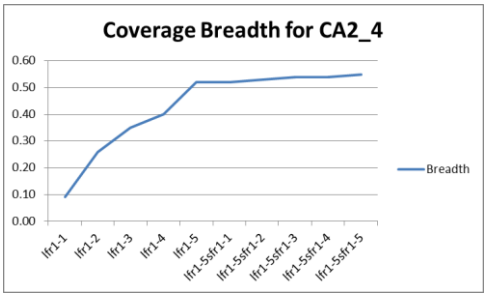
Run	Evenness
lfr1-1	9.79
lfr1-2	11.16
lfr1-3	11.55
lfr1-4	11.66
lfr1-5	11.75
lfr1-5sfr1-1	11.55
lfr1-5sfr1-2	11.22
lfr1-5sfr1-3	11.00
lfr1-5sfr1-4	10.90
lfr1-5sfr1-5	10.84



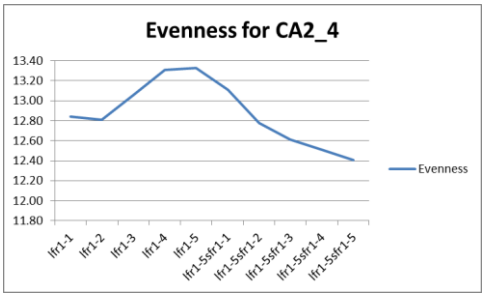
Run	Depth
lfr1-1	0.32
lfr1-2	1.24
lfr1-3	2.10
lfr1-4	2.92
lfr1-5	7.94
lfr1-5sfr1-1	8.27
lfr1-5sfr1-2	8.98
lfr1-5sfr1-3	9.51
lfr1-5sfr1-4	9.96
lfr1-5sfr1-5	10.55



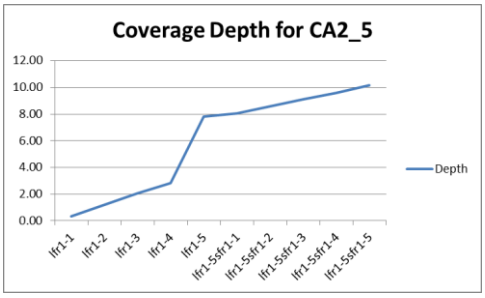
Run	Breadth
lfr1-1	0.09
lfr1-2	0.26
lfr1-3	0.35
lfr1-4	0.40
lfr1-5	0.52
lfr1-5sfr1-1	0.52
lfr1-5sfr1-2	0.53
lfr1-5sfr1-3	0.54
lfr1-5sfr1-4	0.54
lfr1-5sfr1-5	0.55



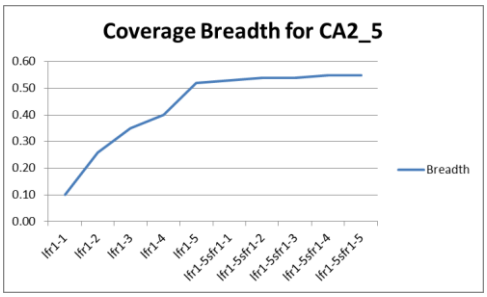
Run	Evenness
lfr1-1	12.84
lfr1-2	12.81
lfr1-3	13.06
lfr1-4	13.31
lfr1-5	13.33
lfr1-5sfr1-1	13.11
lfr1-5sfr1-2	12.78
lfr1-5sfr1-3	12.61
lfr1-5sfr1-4	12.51
lfr1-5sfr1-5	12.41



Run	Depth
lfr1-1	0.32
lfr1-2	1.22
lfr1-3	2.06
lfr1-4	2.81
lfr1-5	7.82
lfr1-5sfr1-1	8.07
lfr1-5sfr1-2	8.56
lfr1-5sfr1-3	9.11
lfr1-5sfr1-4	9.58
lfr1-5sfr1-5	10.18



Run	Breadth
lfr1-1	0.10
lfr1-2	0.26
lfr1-3	0.35
lfr1-4	0.40
lfr1-5	0.52
lfr1-5sfr1-1	0.53
lfr1-5sfr1-2	0.54
lfr1-5sfr1-3	0.54
lfr1-5sfr1-4	0.55
lfr1-5sfr1-5	0.55



Run	Evenness
lfr1-1	13.67
lfr1-2	12.53
lfr1-3	12.74
lfr1-4	12.97
lfr1-5	12.93
lfr1-5sfr1-1	12.75
lfr1-5sfr1-2	12.40
lfr1-5sfr1-3	12.11
lfr1-5sfr1-4	11.89
lfr1-5sfr1-5	11.70

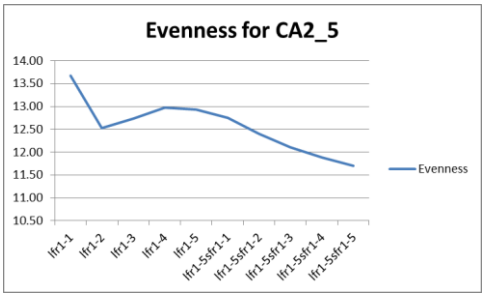
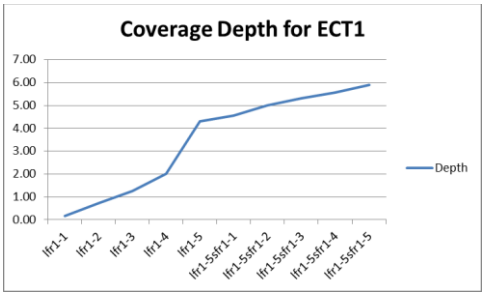
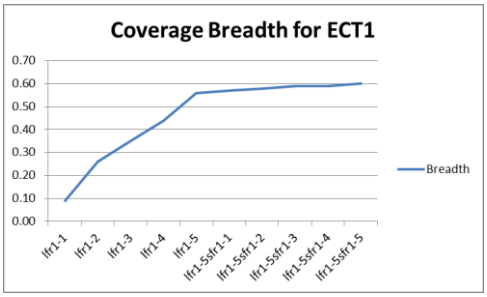


Figure 21. Distribution of coverage depth, coverage breadth, and evenness for population CA2.

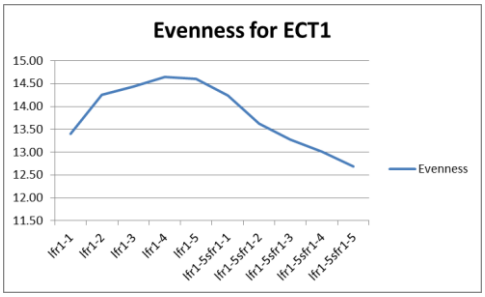
Run	Depth
lfr1-1	0.18
lfr1-2	0.73
lfr1-3	1.25
lfr1-4	2.01
lfr1-5	4.31
lfr1-5sfr1-1	4.55
lfr1-5sfr1-2	5.01
lfr1-5sfr1-3	5.30
lfr1-5sfr1-4	5.56
lfr1-5sfr1-5	5.89



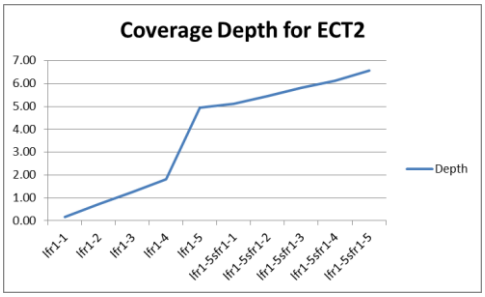
Run	Breadth
lfr1-1	0.09
lfr1-2	0.26
lfr1-3	0.35
lfr1-4	0.44
lfr1-5	0.56
lfr1-5sfr1-1	0.57
lfr1-5sfr1-2	0.58
lfr1-5sfr1-3	0.59
lfr1-5sfr1-4	0.59
lfr1-5sfr1-5	0.60



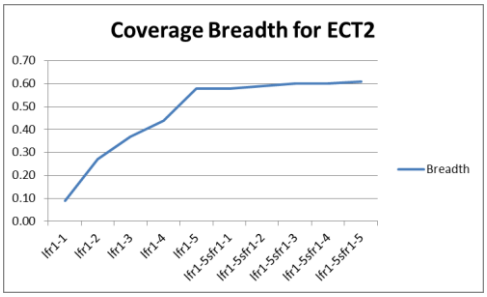
Run	Evenness
lfr1-1	13.40
lfr1-2	14.26
lfr1-3	14.43
lfr1-4	14.65
lfr1-5	14.61
lfr1-5sfr1-1	14.24
lfr1-5sfr1-2	13.63
lfr1-5sfr1-3	13.27
lfr1-5sfr1-4	13.01
lfr1-5sfr1-5	12.69



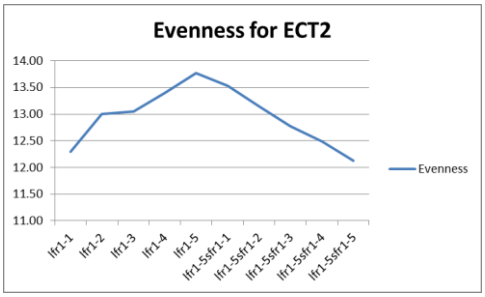
Run	Depth
lfr1-1	0.17
lfr1-2	0.74
lfr1-3	1.26
lfr1-4	1.82
lfr1-5	4.94
lfr1-5sfr1-1	5.13
lfr1-5sfr1-2	5.44
lfr1-5sfr1-3	5.81
lfr1-5sfr1-4	6.12
lfr1-5sfr1-5	6.56



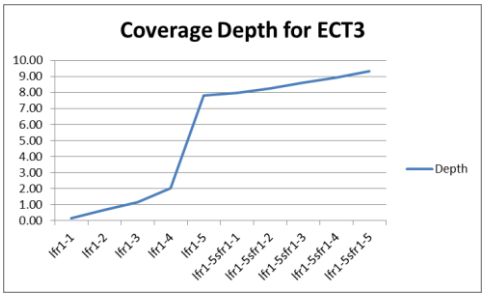
Run	Breadth
lfr1-1	0.09
lfr1-2	0.27
lfr1-3	0.37
lfr1-4	0.44
lfr1-5	0.58
lfr1-5sfr1-1	0.58
lfr1-5sfr1-2	0.59
lfr1-5sfr1-3	0.60
lfr1-5sfr1-4	0.60
lfr1-5sfr1-5	0.61



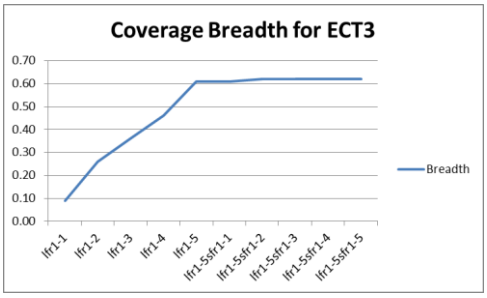
Run	Evenness
lfr1-1	12.29
lfr1-2	13.00
lfr1-3	13.05
lfr1-4	13.40
lfr1-5	13.77
lfr1-5sfr1-1	13.53
lfr1-5sfr1-2	13.15
lfr1-5sfr1-3	12.77
lfr1-5sfr1-4	12.48
lfr1-5sfr1-5	12.13



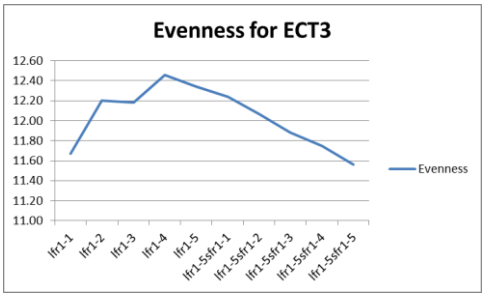
Run	Depth
lfr1-1	0.16
lfr1-2	0.68
lfr1-3	1.15
lfr1-4	2.02
lfr1-5	7.83
lfr1-5sfr1-1	8.00
lfr1-5sfr1-2	8.28
lfr1-5sfr1-3	8.63
lfr1-5sfr1-4	8.93
lfr1-5sfr1-5	9.34



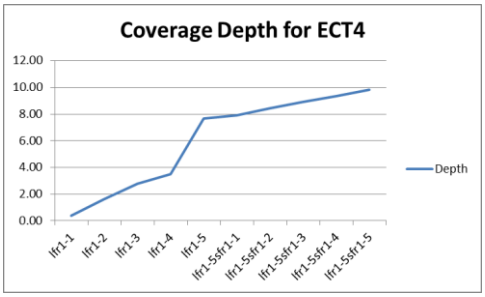
Run	Breadth
lfr1-1	0.09
lfr1-2	0.26
lfr1-3	0.36
lfr1-4	0.46
lfr1-5	0.61
lfr1-5sfr1-1	0.61
lfr1-5sfr1-2	0.62
lfr1-5sfr1-3	0.62
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.62



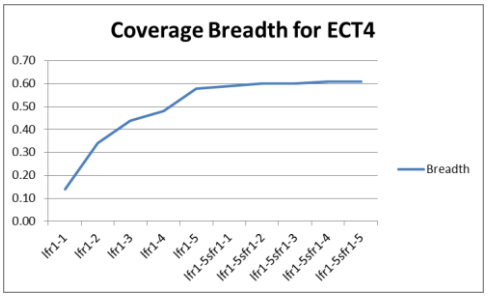
Run	Evenness
lfr1-1	11.67
lfr1-2	12.20
lfr1-3	12.18
lfr1-4	12.46
lfr1-5	12.34
lfr1-5sfr1-1	12.24
lfr1-5sfr1-2	12.07
lfr1-5sfr1-3	11.88
lfr1-5sfr1-4	11.75
lfr1-5sfr1-5	11.56



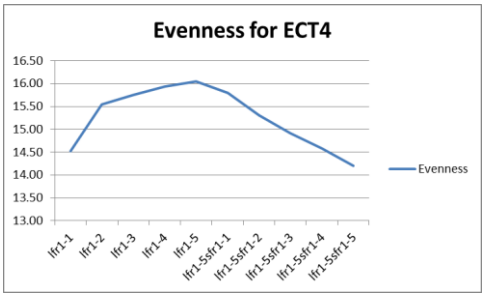
Run	Depth
lfr1-1	0.39
lfr1-2	1.64
lfr1-3	2.80
lfr1-4	3.48
lfr1-5	7.67
lfr1-5sfr1-1	7.93
lfr1-5sfr1-2	8.45
lfr1-5sfr1-3	8.92
lfr1-5sfr1-4	9.33
lfr1-5sfr1-5	9.85



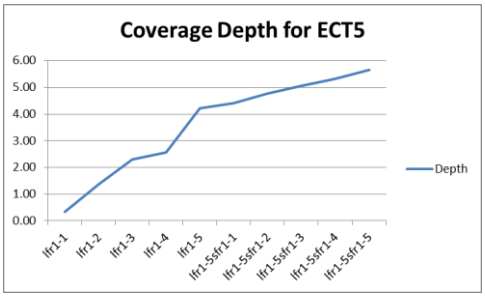
Run	Breadth
lfr1-1	0.14
lfr1-2	0.34
lfr1-3	0.44
lfr1-4	0.48
lfr1-5	0.58
lfr1-5sfr1-1	0.59
lfr1-5sfr1-2	0.60
lfr1-5sfr1-3	0.60
lfr1-5sfr1-4	0.61
lfr1-5sfr1-5	0.61



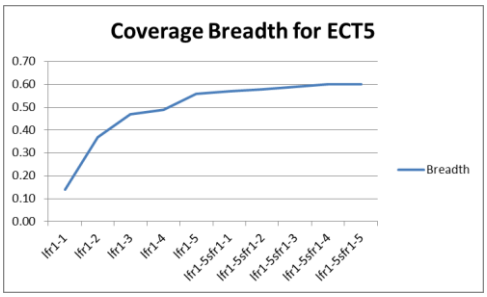
Run	Evenness
lfr1-1	14.53
lfr1-2	15.54
lfr1-3	15.75
lfr1-4	15.93
lfr1-5	16.05
lfr1-5sfr1-1	15.79
lfr1-5sfr1-2	15.30
lfr1-5sfr1-3	14.91
lfr1-5sfr1-4	14.58
lfr1-5sfr1-5	14.20



Run	Depth
lfr1-1	0.33
lfr1-2	1.36
lfr1-3	2.31
lfr1-4	2.57
lfr1-5	4.22
lfr1-5sfr1-1	4.41
lfr1-5sfr1-2	4.77
lfr1-5sfr1-3	5.06
lfr1-5sfr1-4	5.31
lfr1-5sfr1-5	5.65



Run	Breadth
lfr1-1	0.14
lfr1-2	0.37
lfr1-3	0.47
lfr1-4	0.49
lfr1-5	0.56
lfr1-5sfr1-1	0.57
lfr1-5sfr1-2	0.58
lfr1-5sfr1-3	0.59
lfr1-5sfr1-4	0.60
lfr1-5sfr1-5	0.60



Run	Evenness
lfr1-1	14.61
lfr1-2	15.04
lfr1-3	15.00
lfr1-4	15.13
lfr1-5	14.82
lfr1-5sfr1-1	14.49
lfr1-5sfr1-2	13.92
lfr1-5sfr1-3	13.50
lfr1-5sfr1-4	13.18
lfr1-5sfr1-5	12.80

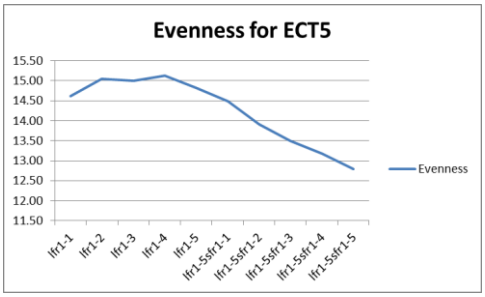
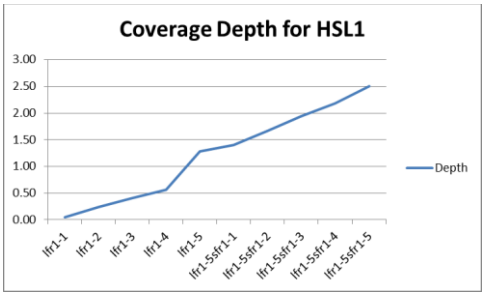
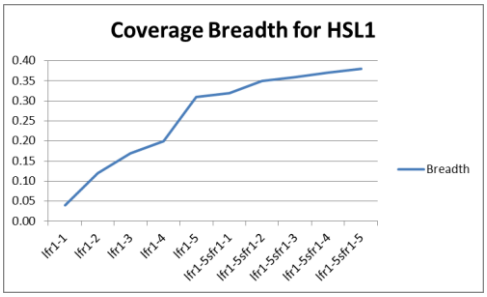


Figure 22. Distribution of coverage depth, coverage breadth, and evenness for population ECT.

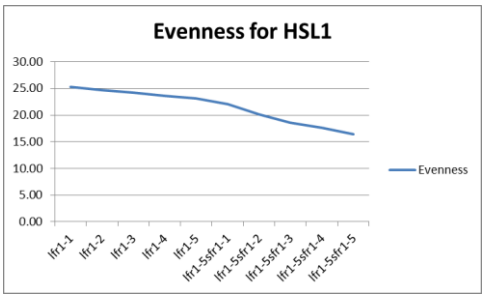
Run	Depth
lfr1-1	0.05
lfr1-2	0.24
lfr1-3	0.41
lfr1-4	0.56
lfr1-5	1.28
lfr1-5sfr1-1	1.40
lfr1-5sfr1-2	1.67
lfr1-5sfr1-3	1.94
lfr1-5sfr1-4	2.18
lfr1-5sfr1-5	2.51



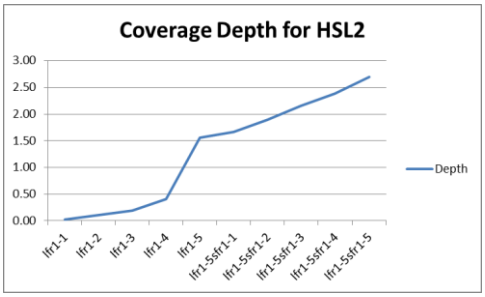
Run	Breadth
lfr1-1	0.04
lfr1-2	0.12
lfr1-3	0.17
lfr1-4	0.20
lfr1-5	0.31
lfr1-5sfr1-1	0.32
lfr1-5sfr1-2	0.35
lfr1-5sfr1-3	0.36
lfr1-5sfr1-4	0.37
lfr1-5sfr1-5	0.38



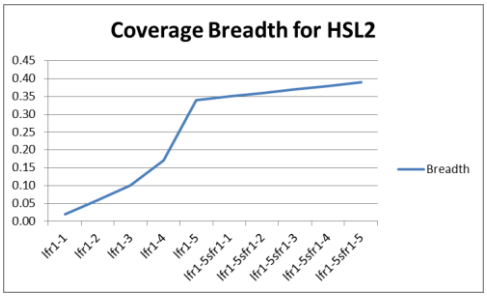
Run	Evenness
lfr1-1	25.29
lfr1-2	24.63
lfr1-3	24.15
lfr1-4	23.55
lfr1-5	23.09
lfr1-5sfr1-1	22.08
lfr1-5sfr1-2	20.18
lfr1-5sfr1-3	18.63
lfr1-5sfr1-4	17.58
lfr1-5sfr1-5	16.46



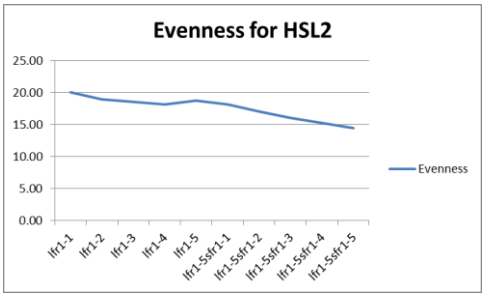
Run	Depth
lfr1-1	0.02
lfr1-2	0.11
lfr1-3	0.19
lfr1-4	0.41
lfr1-5	1.56
lfr1-5sfr1-1	1.67
lfr1-5sfr1-2	1.89
lfr1-5sfr1-3	2.16
lfr1-5sfr1-4	2.38
lfr1-5sfr1-5	2.70



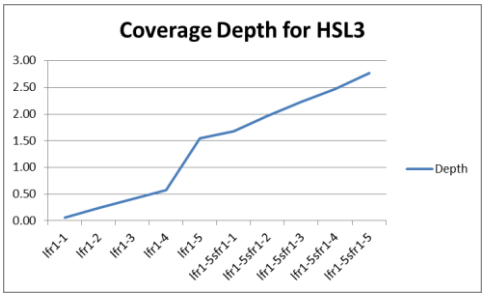
Run	Breadth
lfr1-1	0.02
lfr1-2	0.06
lfr1-3	0.10
lfr1-4	0.17
lfr1-5	0.34
lfr1-5sfr1-1	0.35
lfr1-5sfr1-2	0.36
lfr1-5sfr1-3	0.37
lfr1-5sfr1-4	0.38
lfr1-5sfr1-5	0.39



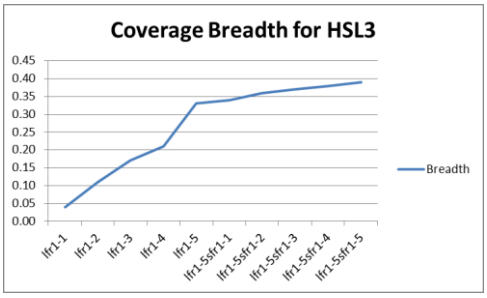
Run	Evenness
lfr1-1	20.06
lfr1-2	18.96
lfr1-3	18.62
lfr1-4	18.20
lfr1-5	18.81
lfr1-5sfr1-1	18.21
lfr1-5sfr1-2	17.10
lfr1-5sfr1-3	16.06
lfr1-5sfr1-4	15.31
lfr1-5sfr1-5	14.49



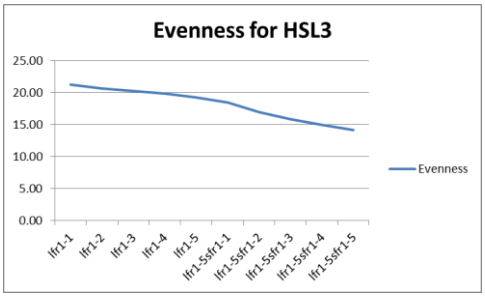
Run	Depth
lfr1-1	0.06
lfr1-2	0.24
lfr1-3	0.41
lfr1-4	0.57
lfr1-5	1.55
lfr1-5sfr1-1	1.68
lfr1-5sfr1-2	1.97
lfr1-5sfr1-3	2.23
lfr1-5sfr1-4	2.47
lfr1-5sfr1-5	2.77



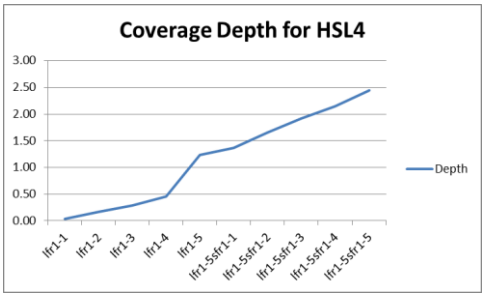
Run	Breadth
lfr1-1	0.04
lfr1-2	0.11
lfr1-3	0.17
lfr1-4	0.21
lfr1-5	0.33
lfr1-5sfr1-1	0.34
lfr1-5sfr1-2	0.36
lfr1-5sfr1-3	0.37
lfr1-5sfr1-4	0.38
lfr1-5sfr1-5	0.39



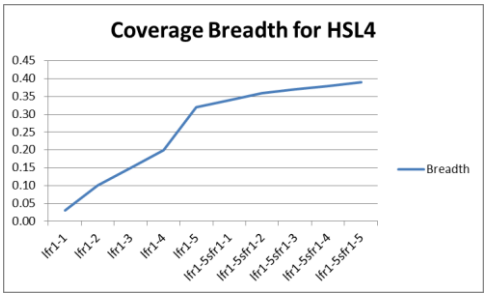
Run	Evenness
lfr1-1	21.31
lfr1-2	20.63
lfr1-3	20.32
lfr1-4	19.86
lfr1-5	19.24
lfr1-5sfr1-1	18.46
lfr1-5sfr1-2	16.95
lfr1-5sfr1-3	15.83
lfr1-5sfr1-4	14.99
lfr1-5sfr1-5	14.20



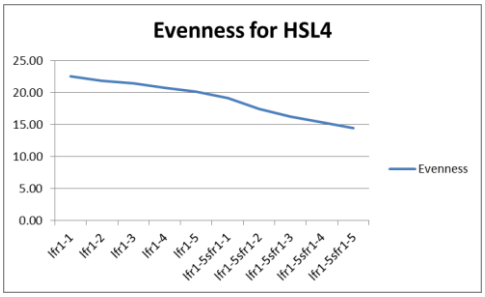
Run	Depth
lfr1-1	0.04
lfr1-2	0.17
lfr1-3	0.29
lfr1-4	0.46
lfr1-5	1.24
lfr1-5sfr1-1	1.37
lfr1-5sfr1-2	1.65
lfr1-5sfr1-3	1.92
lfr1-5sfr1-4	2.14
lfr1-5sfr1-5	2.45



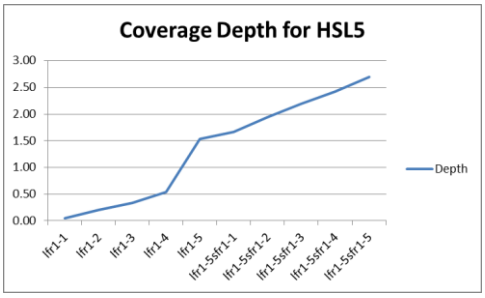
Run	Breadth
lfr1-1	0.03
lfr1-2	0.10
lfr1-3	0.15
lfr1-4	0.20
lfr1-5	0.32
lfr1-5sfr1-1	0.34
lfr1-5sfr1-2	0.36
lfr1-5sfr1-3	0.37
lfr1-5sfr1-4	0.38
lfr1-5sfr1-5	0.39



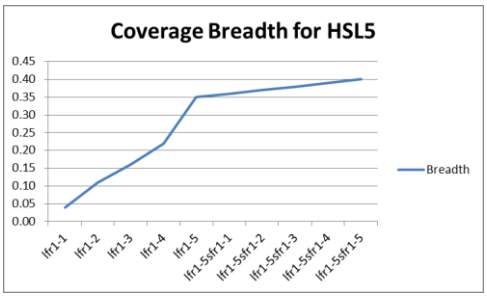
Run	Evenness
lfr1-1	22.56
lfr1-2	21.91
lfr1-3	21.52
lfr1-4	20.80
lfr1-5	20.19
lfr1-5sfr1-1	19.20
lfr1-5sfr1-2	17.50
lfr1-5sfr1-3	16.24
lfr1-5sfr1-4	15.36
lfr1-5sfr1-5	14.50



Run	Depth
lfr1-1	0.05
lfr1-2	0.20
lfr1-3	0.33
lfr1-4	0.54
lfr1-5	1.53
lfr1-5sfr1-1	1.66
lfr1-5sfr1-2	1.94
lfr1-5sfr1-3	2.19
lfr1-5sfr1-4	2.42
lfr1-5sfr1-5	2.70



Run	Breadth
lfr1-1	0.04
lfr1-2	0.11
lfr1-3	0.16
lfr1-4	0.22
lfr1-5	0.35
lfr1-5sfr1-1	0.36
lfr1-5sfr1-2	0.37
lfr1-5sfr1-3	0.38
lfr1-5sfr1-4	0.39
lfr1-5sfr1-5	0.40



Run	Evenness
lfr1-1	20.03
lfr1-2	19.04
lfr1-3	18.85
lfr1-4	18.42
lfr1-5	17.72
lfr1-5sfr1-1	17.09
lfr1-5sfr1-2	15.89
lfr1-5sfr1-3	15.09
lfr1-5sfr1-4	14.45
lfr1-5sfr1-5	13.81

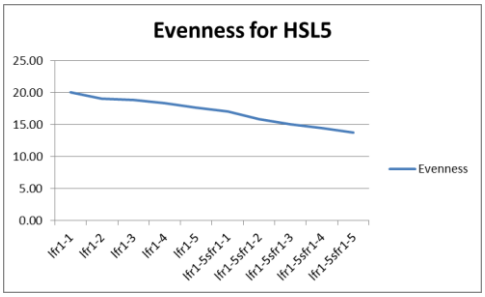
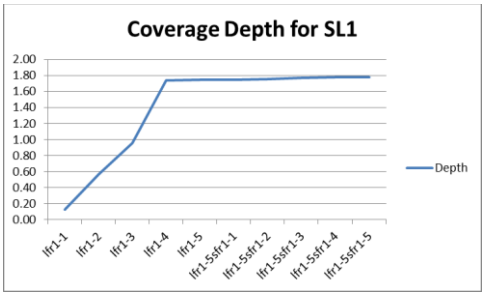
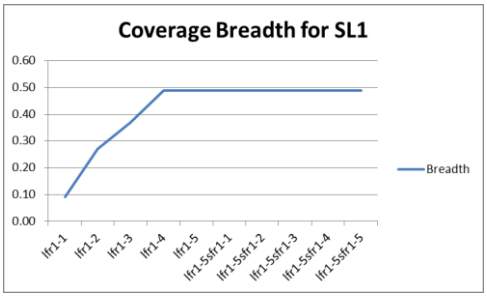


Figure 23. Distribution of coverage depth, coverage breadth, and evenness for population HSL.

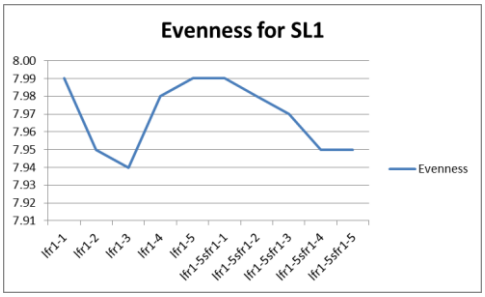
Run	Depth
lfr1-1	0.13
lfr1-2	0.57
lfr1-3	0.96
lfr1-4	1.74
lfr1-5	1.75
lfr1-5sfr1-1	1.75
lfr1-5sfr1-2	1.76
lfr1-5sfr1-3	1.77
lfr1-5sfr1-4	1.78
lfr1-5sfr1-5	1.78



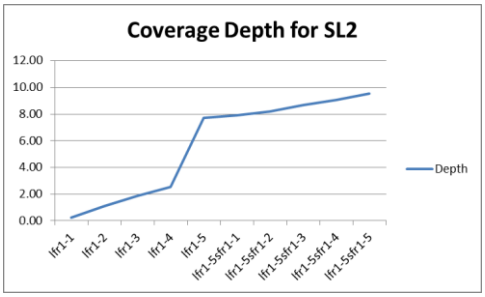
Run	Breadth
lfr1-1	0.09
lfr1-2	0.27
lfr1-3	0.37
lfr1-4	0.49
lfr1-5	0.49
lfr1-5sfr1-1	0.49
lfr1-5sfr1-2	0.49
lfr1-5sfr1-3	0.49
lfr1-5sfr1-4	0.49
lfr1-5sfr1-5	0.49



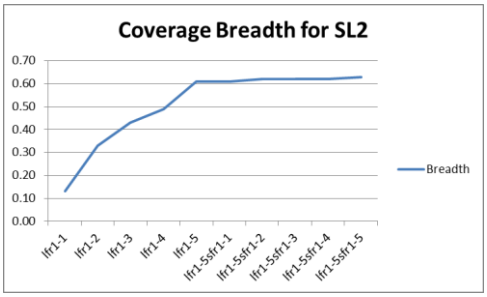
Run	Evenness
lfr1-1	7.99
lfr1-2	7.95
lfr1-3	7.94
lfr1-4	7.98
lfr1-5	7.99
lfr1-5sfr1-1	7.99
lfr1-5sfr1-2	7.98
lfr1-5sfr1-3	7.97
lfr1-5sfr1-4	7.95
lfr1-5sfr1-5	7.95



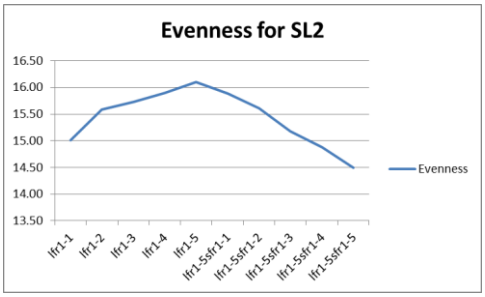
Run	Depth
lfr1-1	0.26
lfr1-2	1.08
lfr1-3	1.85
lfr1-4	2.53
lfr1-5	7.72
lfr1-5sfr1-1	7.89
lfr1-5sfr1-2	8.21
lfr1-5sfr1-3	8.66
lfr1-5sfr1-4	9.06
lfr1-5sfr1-5	9.56



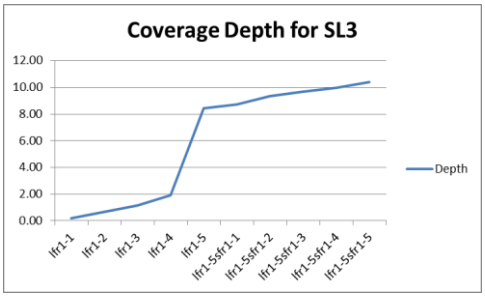
Run	Breadth
lfr1-1	0.13
lfr1-2	0.33
lfr1-3	0.43
lfr1-4	0.49
lfr1-5	0.61
lfr1-5sfr1-1	0.61
lfr1-5sfr1-2	0.62
lfr1-5sfr1-3	0.62
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.63



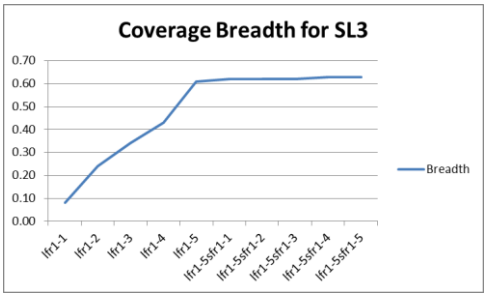
Run	Evenness
lfr1-1	15.01
lfr1-2	15.59
lfr1-3	15.73
lfr1-4	15.90
lfr1-5	16.10
lfr1-5sfr1-1	15.89
lfr1-5sfr1-2	15.61
lfr1-5sfr1-3	15.18
lfr1-5sfr1-4	14.88
lfr1-5sfr1-5	14.49



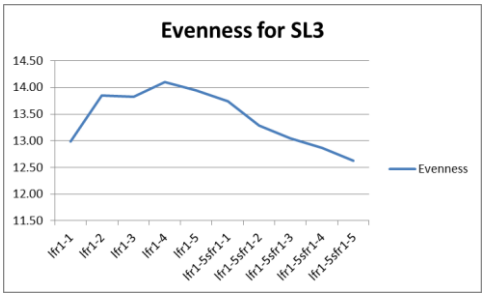
Run	Depth
lfr1-1	0.17
lfr1-2	0.68
lfr1-3	1.16
lfr1-4	1.93
lfr1-5	8.45
lfr1-5sfr1-1	8.70
lfr1-5sfr1-2	9.33
lfr1-5sfr1-3	9.68
lfr1-5sfr1-4	9.99
lfr1-5sfr1-5	10.39



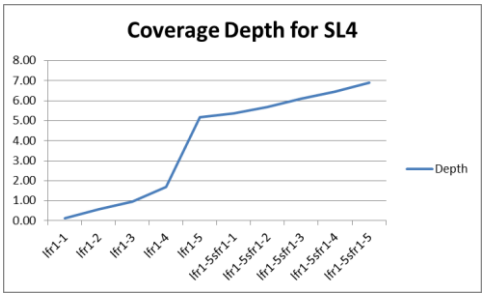
Run	Breadth
lfr1-1	0.08
lfr1-2	0.24
lfr1-3	0.34
lfr1-4	0.43
lfr1-5	0.61
lfr1-5sfr1-1	0.62
lfr1-5sfr1-2	0.62
lfr1-5sfr1-3	0.62
lfr1-5sfr1-4	0.63
lfr1-5sfr1-5	0.63



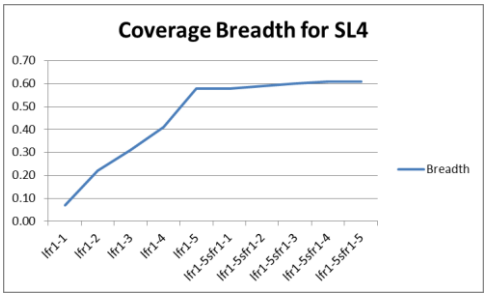
Run	Evenness
lfr1-1	12.98
lfr1-2	13.85
lfr1-3	13.83
lfr1-4	14.10
lfr1-5	13.95
lfr1-5sfr1-1	13.74
lfr1-5sfr1-2	13.29
lfr1-5sfr1-3	13.05
lfr1-5sfr1-4	12.86
lfr1-5sfr1-5	12.63



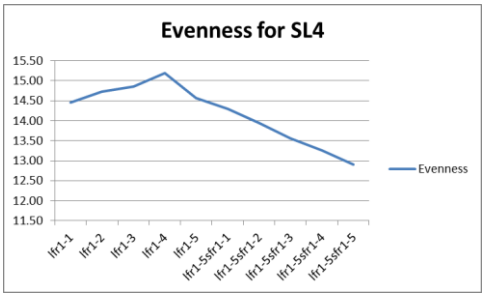
Run	Depth
lfr1-1	0.14
lfr1-2	0.56
lfr1-3	0.95
lfr1-4	1.69
lfr1-5	5.18
lfr1-5sfr1-1	5.37
lfr1-5sfr1-2	5.70
lfr1-5sfr1-3	6.10
lfr1-5sfr1-4	6.44
lfr1-5sfr1-5	6.89



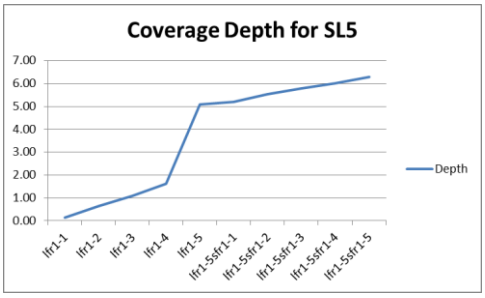
Run	Breadth
lfr1-1	0.07
lfr1-2	0.22
lfr1-3	0.31
lfr1-4	0.41
lfr1-5	0.58
lfr1-5sfr1-1	0.58
lfr1-5sfr1-2	0.59
lfr1-5sfr1-3	0.60
lfr1-5sfr1-4	0.61
lfr1-5sfr1-5	0.61



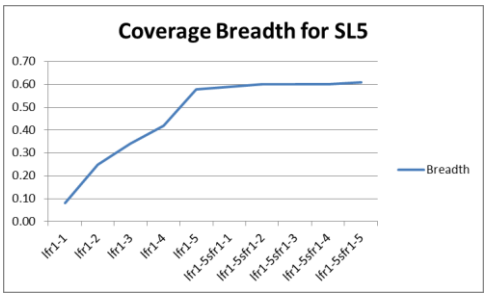
Run	Evenness
lfr1-1	14.45
lfr1-2	14.73
lfr1-3	14.85
lfr1-4	15.19
lfr1-5	14.56
lfr1-5sfr1-1	14.30
lfr1-5sfr1-2	13.94
lfr1-5sfr1-3	13.56
lfr1-5sfr1-4	13.25
lfr1-5sfr1-5	12.90



Run	Depth
lfr1-1	0.15
lfr1-2	0.64
lfr1-3	1.08
lfr1-4	1.62
lfr1-5	5.09
lfr1-5sfr1-1	5.19
lfr1-5sfr1-2	5.54
lfr1-5sfr1-3	5.79
lfr1-5sfr1-4	6.00
lfr1-5sfr1-5	6.28



Run	Breadth
lfr1-1	0.08
lfr1-2	0.25
lfr1-3	0.34
lfr1-4	0.42
lfr1-5	0.58
lfr1-5sfr1-1	0.59
lfr1-5sfr1-2	0.60
lfr1-5sfr1-3	0.60
lfr1-5sfr1-4	0.60
lfr1-5sfr1-5	0.61



Run	Evenness
lfr1-1	13.25
lfr1-2	13.48
lfr1-3	13.62
lfr1-4	13.88
lfr1-5	13.65
lfr1-5sfr1-1	13.49
lfr1-5sfr1-2	13.11
lfr1-5sfr1-3	12.87
lfr1-5sfr1-4	12.67
lfr1-5sfr1-5	12.44

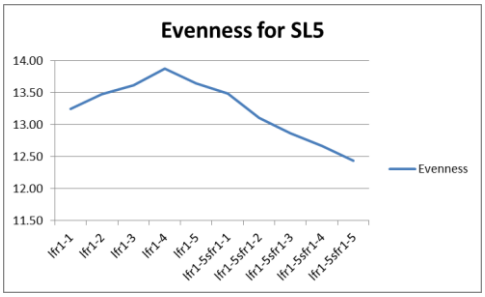
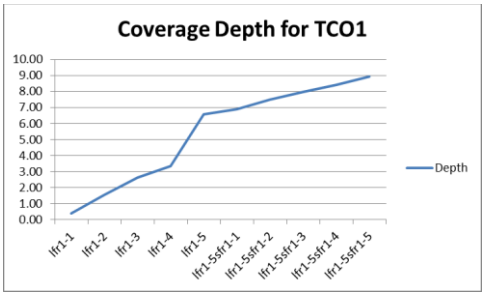
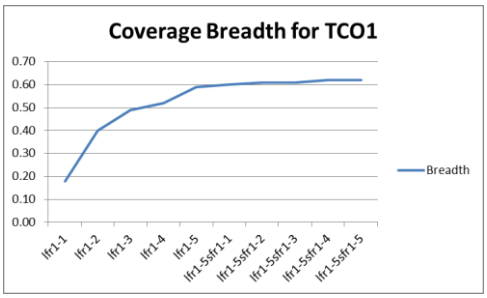


Figure 24. Distribution of coverage depth, coverage breadth, and evenness for population SL.

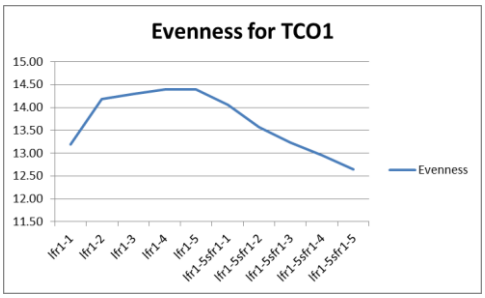
Run	Depth
lfr1-1	0.39
lfr1-2	1.55
lfr1-3	2.64
lfr1-4	3.37
lfr1-5	6.60
lfr1-5sfr1-1	6.91
lfr1-5sfr1-2	7.51
lfr1-5sfr1-3	7.99
lfr1-5sfr1-4	8.42
lfr1-5sfr1-5	8.96



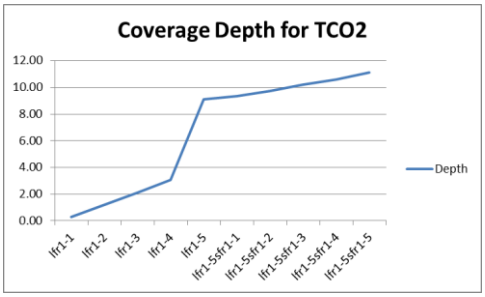
Run	Breadth
lfr1-1	0.18
lfr1-2	0.40
lfr1-3	0.49
lfr1-4	0.52
lfr1-5	0.59
lfr1-5sfr1-1	0.60
lfr1-5sfr1-2	0.61
lfr1-5sfr1-3	0.61
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.62



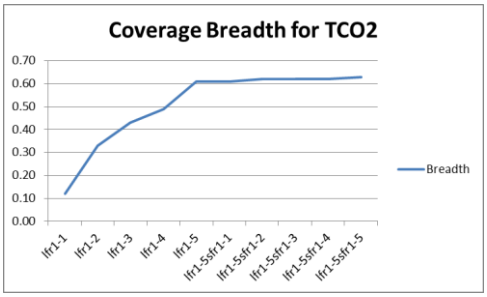
Run	Evenness
lfr1-1	13.19
lfr1-2	14.19
lfr1-3	14.30
lfr1-4	14.39
lfr1-5	14.39
lfr1-5sfr1-1	14.06
lfr1-5sfr1-2	13.57
lfr1-5sfr1-3	13.24
lfr1-5sfr1-4	12.96
lfr1-5sfr1-5	12.64



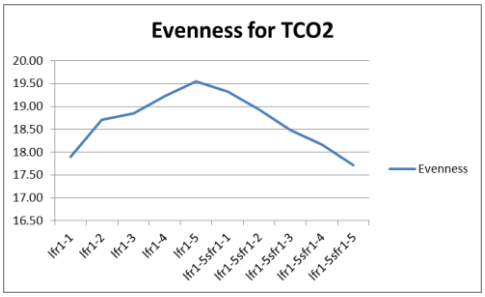
Run	Depth
lfr1-1	0.28
lfr1-2	1.22
lfr1-3	2.10
lfr1-4	3.07
lfr1-5	9.12
lfr1-5sfr1-1	9.34
lfr1-5sfr1-2	9.74
lfr1-5sfr1-3	10.19
lfr1-5sfr1-4	10.57
lfr1-5sfr1-5	11.10



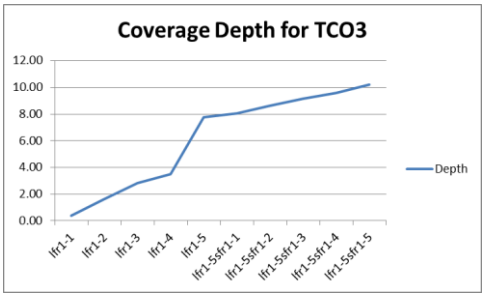
Run	Breadth
lfr1-1	0.12
lfr1-2	0.33
lfr1-3	0.43
lfr1-4	0.49
lfr1-5	0.61
lfr1-5sfr1-1	0.61
lfr1-5sfr1-2	0.62
lfr1-5sfr1-3	0.62
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.63



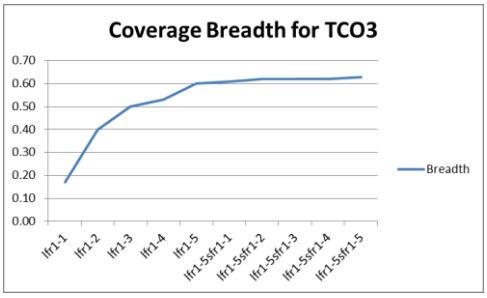
Run	Evenness
lfr1-1	17.90
lfr1-2	18.71
lfr1-3	18.85
lfr1-4	19.22
lfr1-5	19.55
lfr1-5sfr1-1	19.32
lfr1-5sfr1-2	18.93
lfr1-5sfr1-3	18.48
lfr1-5sfr1-4	18.16
lfr1-5sfr1-5	17.72



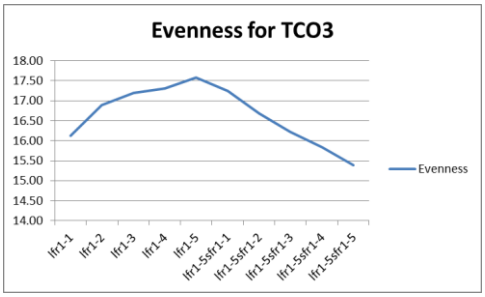
Run	Depth
lfr1-1	0.40
lfr1-2	1.64
lfr1-3	2.81
lfr1-4	3.51
lfr1-5	7.77
lfr1-5sfr1-1	8.05
lfr1-5sfr1-2	8.62
lfr1-5sfr1-3	9.14
lfr1-5sfr1-4	9.60
lfr1-5sfr1-5	10.20



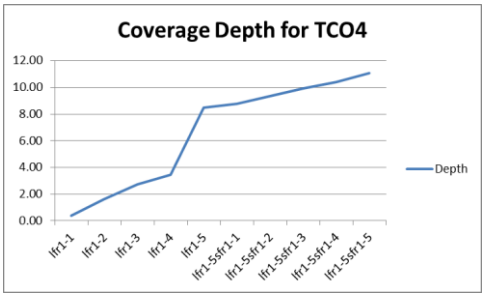
Run	Breadth
lfr1-1	0.17
lfr1-2	0.40
lfr1-3	0.50
lfr1-4	0.53
lfr1-5	0.60
lfr1-5sfr1-1	0.61
lfr1-5sfr1-2	0.62
lfr1-5sfr1-3	0.62
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.63



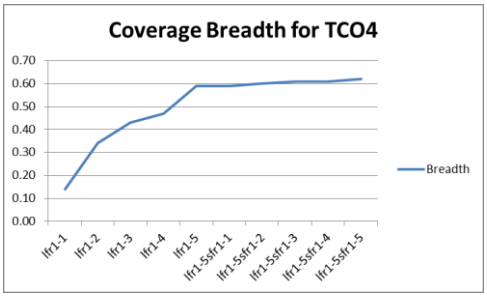
Run	Evenness
lfr1-1	16.12
lfr1-2	16.89
lfr1-3	17.19
lfr1-4	17.30
lfr1-5	17.58
lfr1-5sfr1-1	17.24
lfr1-5sfr1-2	16.68
lfr1-5sfr1-3	16.22
lfr1-5sfr1-4	15.84
lfr1-5sfr1-5	15.39



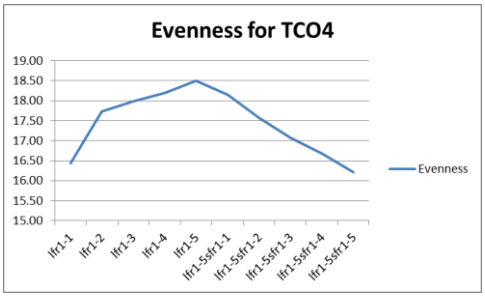
Run	Depth
lfr1-1	0.39
lfr1-2	1.61
lfr1-3	2.75
lfr1-4	3.45
lfr1-5	8.49
lfr1-5sfr1-1	8.77
lfr1-5sfr1-2	9.36
lfr1-5sfr1-3	9.93
lfr1-5sfr1-4	10.42
lfr1-5sfr1-5	11.05



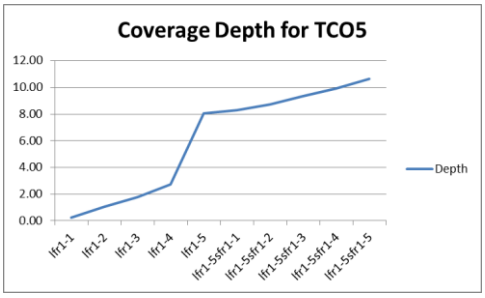
Run	Breadth
lfr1-1	0.14
lfr1-2	0.34
lfr1-3	0.43
lfr1-4	0.47
lfr1-5	0.59
lfr1-5sfr1-1	0.59
lfr1-5sfr1-2	0.60
lfr1-5sfr1-3	0.61
lfr1-5sfr1-4	0.61
lfr1-5sfr1-5	0.62



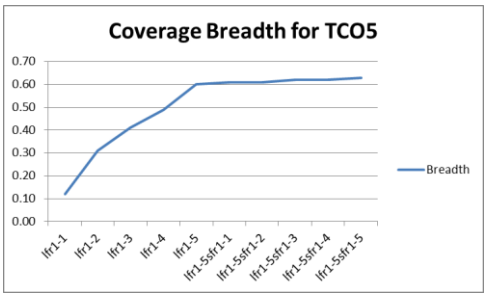
Run	Evenness
lfr1-1	16.44
lfr1-2	17.74
lfr1-3	17.98
lfr1-4	18.19
lfr1-5	18.50
lfr1-5sfr1-1	18.15
lfr1-5sfr1-2	17.58
lfr1-5sfr1-3	17.08
lfr1-5sfr1-4	16.68
lfr1-5sfr1-5	16.21



Run	Depth
lfr1-1	0.25
lfr1-2	1.05
lfr1-3	1.79
lfr1-4	2.75
lfr1-5	8.05
lfr1-5sfr1-1	8.28
lfr1-5sfr1-2	8.74
lfr1-5sfr1-3	9.37
lfr1-5sfr1-4	9.91
lfr1-5sfr1-5	10.62



Run	Breadth
lfr1-1	0.12
lfr1-2	0.31
lfr1-3	0.41
lfr1-4	0.49
lfr1-5	0.60
lfr1-5sfr1-1	0.61
lfr1-5sfr1-2	0.61
lfr1-5sfr1-3	0.62
lfr1-5sfr1-4	0.62
lfr1-5sfr1-5	0.63



Run	Evenness
lfr1-1	15.12
lfr1-2	16.27
lfr1-3	16.33
lfr1-4	16.66
lfr1-5	17.06
lfr1-5sfr1-1	16.82
lfr1-5sfr1-2	16.38
lfr1-5sfr1-3	15.83
lfr1-5sfr1-4	15.41
lfr1-5sfr1-5	14.93

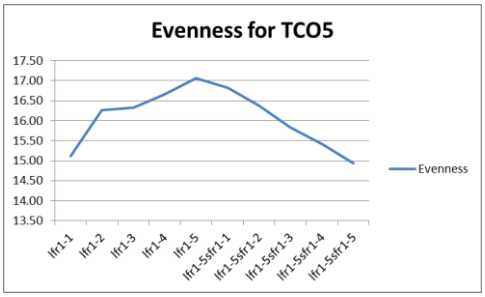
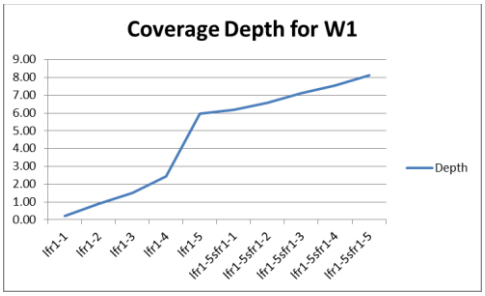
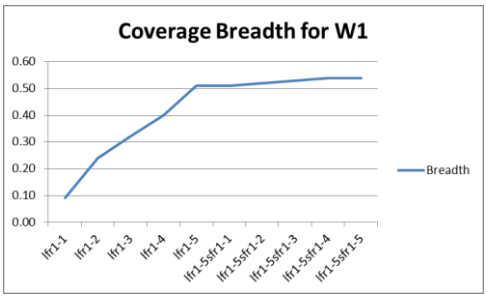


Figure 25. Distribution of coverage depth, coverage breadth, and evenness for population TCO.

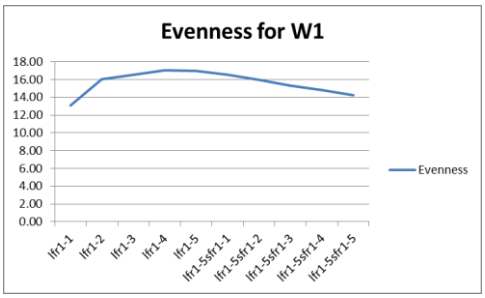
Run	Depth
lfr1-1	0.21
lfr1-2	0.89
lfr1-3	1.51
lfr1-4	2.43
lfr1-5	5.98
lfr1-5sfr1-1	6.20
lfr1-5sfr1-2	6.59
lfr1-5sfr1-3	7.11
lfr1-5sfr1-4	7.54
lfr1-5sfr1-5	8.14



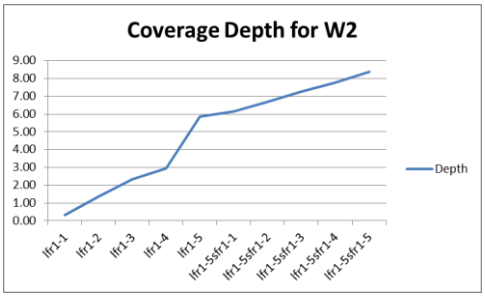
Run	Breadth
lfr1-1	0.09
lfr1-2	0.24
lfr1-3	0.32
lfr1-4	0.40
lfr1-5	0.51
lfr1-5sfr1-1	0.51
lfr1-5sfr1-2	0.52
lfr1-5sfr1-3	0.53
lfr1-5sfr1-4	0.54
lfr1-5sfr1-5	0.54



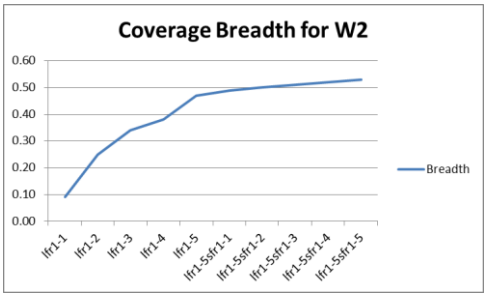
Run	Evenness
lfr1-1	13.05
lfr1-2	16.06
lfr1-3	16.51
lfr1-4	17.04
lfr1-5	16.95
lfr1-5sfr1-1	16.54
lfr1-5sfr1-2	15.98
lfr1-5sfr1-3	15.31
lfr1-5sfr1-4	14.84
lfr1-5sfr1-5	14.25



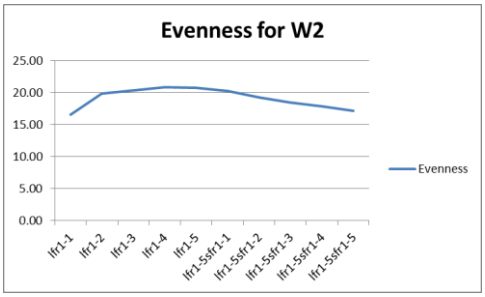
Run	Depth
lfr1-1	0.34
lfr1-2	1.37
lfr1-3	2.34
lfr1-4	2.96
lfr1-5	5.85
lfr1-5sfr1-1	6.13
lfr1-5sfr1-2	6.67
lfr1-5sfr1-3	7.25
lfr1-5sfr1-4	7.75
lfr1-5sfr1-5	8.39



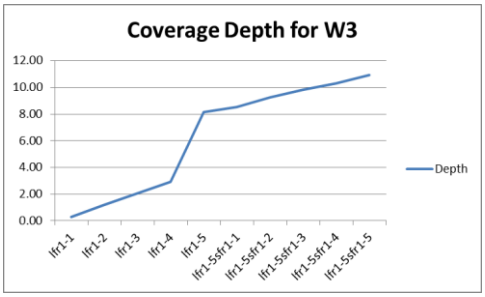
Run	Breadth
lfr1-1	0.09
lfr1-2	0.25
lfr1-3	0.34
lfr1-4	0.38
lfr1-5	0.47
lfr1-5sfr1-1	0.49
lfr1-5sfr1-2	0.50
lfr1-5sfr1-3	0.51
lfr1-5sfr1-4	0.52
lfr1-5sfr1-5	0.53



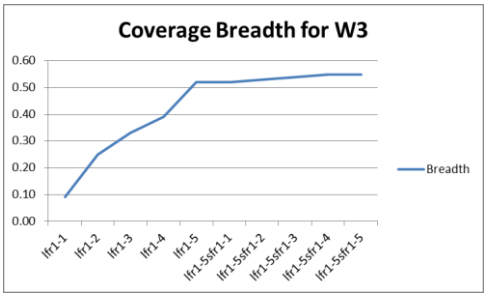
Run	Evenness
lfr1-1	16.59
lfr1-2	19.88
lfr1-3	20.37
lfr1-4	20.87
lfr1-5	20.77
lfr1-5sfr1-1	20.24
lfr1-5sfr1-2	19.31
lfr1-5sfr1-3	18.45
lfr1-5sfr1-4	17.84
lfr1-5sfr1-5	17.17



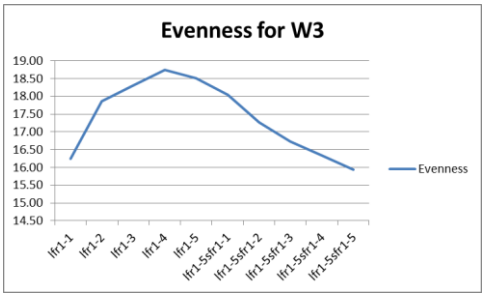
Run	Depth
lfr1-1	0.30
lfr1-2	1.22
lfr1-3	2.08
lfr1-4	2.90
lfr1-5	8.16
lfr1-5sfr1-1	8.52
lfr1-5sfr1-2	9.24
lfr1-5sfr1-3	9.81
lfr1-5sfr1-4	10.31
lfr1-5sfr1-5	10.93



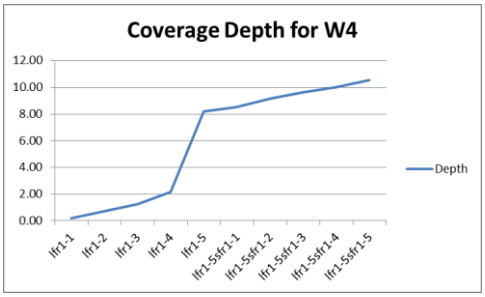
Run	Breadth
lfr1-1	0.09
lfr1-2	0.25
lfr1-3	0.33
lfr1-4	0.39
lfr1-5	0.52
lfr1-5sfr1-1	0.52
lfr1-5sfr1-2	0.53
lfr1-5sfr1-3	0.54
lfr1-5sfr1-4	0.55
lfr1-5sfr1-5	0.55



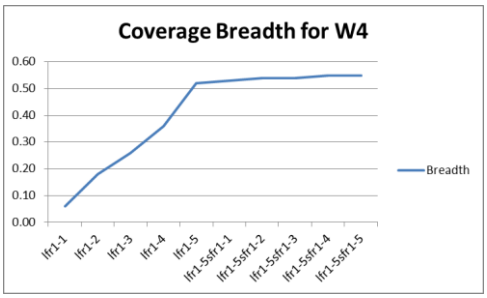
Run	Evenness
lfr1-1	16.25
lfr1-2	17.86
lfr1-3	18.31
lfr1-4	18.74
lfr1-5	18.50
lfr1-5sfr1-1	18.05
lfr1-5sfr1-2	17.26
lfr1-5sfr1-3	16.73
lfr1-5sfr1-4	16.33
lfr1-5sfr1-5	15.93



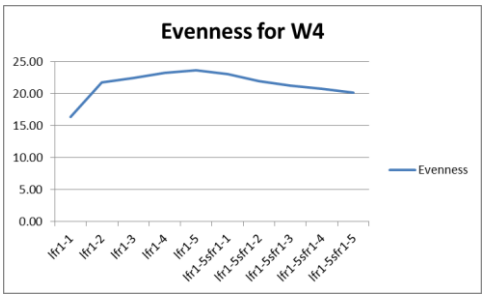
Run	Depth
lfr1-1	0.18
lfr1-2	0.73
lfr1-3	1.25
lfr1-4	2.15
lfr1-5	8.18
lfr1-5sfr1-1	8.51
lfr1-5sfr1-2	9.15
lfr1-5sfr1-3	9.62
lfr1-5sfr1-4	10.03
lfr1-5sfr1-5	10.54



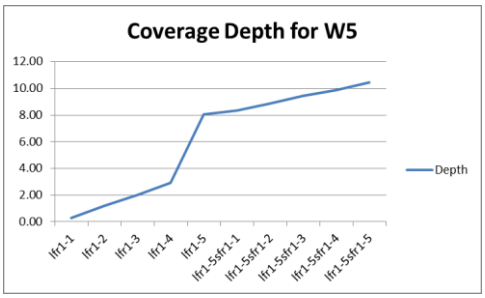
Run	Breadth
lfr1-1	0.06
lfr1-2	0.18
lfr1-3	0.26
lfr1-4	0.36
lfr1-5	0.52
lfr1-5sfr1-1	0.53
lfr1-5sfr1-2	0.54
lfr1-5sfr1-3	0.54
lfr1-5sfr1-4	0.55
lfr1-5sfr1-5	0.55



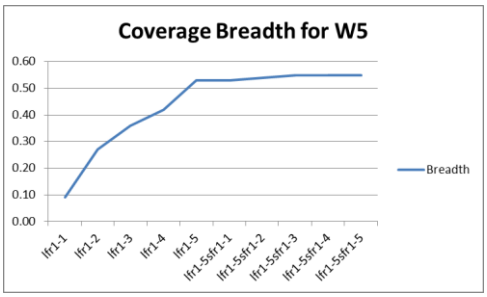
Run	Evenness
lfr1-1	16.34
lfr1-2	21.80
lfr1-3	22.45
lfr1-4	23.24
lfr1-5	23.67
lfr1-5sfr1-1	23.06
lfr1-5sfr1-2	22.01
lfr1-5sfr1-3	21.32
lfr1-5sfr1-4	20.77
lfr1-5sfr1-5	20.16



Run	Depth
lfr1-1	0.29
lfr1-2	1.19
lfr1-3	2.03
lfr1-4	2.94
lfr1-5	8.04
lfr1-5sfr1-1	8.32
lfr1-5sfr1-2	8.89
lfr1-5sfr1-3	9.42
lfr1-5sfr1-4	9.88
lfr1-5sfr1-5	10.45



Run	Breadth
lfr1-1	0.09
lfr1-2	0.27
lfr1-3	0.36
lfr1-4	0.42
lfr1-5	0.53
lfr1-5sfr1-1	0.53
lfr1-5sfr1-2	0.54
lfr1-5sfr1-3	0.55
lfr1-5sfr1-4	0.55
lfr1-5sfr1-5	0.55



Run	Evenness
lfr1-1	14.05
lfr1-2	18.10
lfr1-3	18.71
lfr1-4	19.18
lfr1-5	19.47
lfr1-5sfr1-1	19.03
lfr1-5sfr1-2	18.26
lfr1-5sfr1-3	17.62
lfr1-5sfr1-4	17.16
lfr1-5sfr1-5	16.67

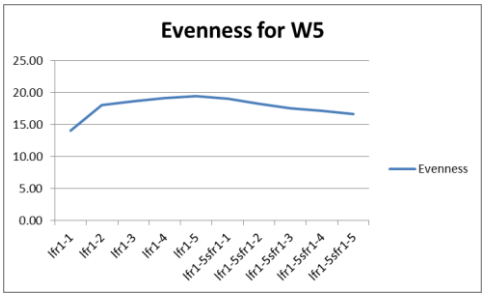


Figure 26. Distribution of coverage depth, coverage breadth, and evenness for population W3.6A.

Analysis of sequencing reruns

The ECT *D.pulex* gDNA library was sequenced twice without multiplexing and generating two paired-end sequencing datasets, ECT and ECT_rerun. These two datasets as well as their combined dataset was run through the SA_Run2Ref workflow, producing statistics presented in Table 13. Approximately 88% of the cleaned reads from the ECT or the ECT_rerun dataset were mapped to the referenced genome, covering 76% of the 5,191 scaffolds or 64% of the entire genome at a 9-fold depth. The combined dataset covered less than 1% or more scaffolds than individual datasets, and it also had similar genome coverage breadth and evenness as the two separate datasets, even though it doubled the genome coverage depth. The distribution of scaffold coverage breadth showed a very similar pattern with ca. 1200 scaffolds uncovered for all three datasets (Figure 27). In comparison with the two separate datasets, the combined dataset

covered 830 and 895 more scaffolds at > 4-fold depth or 700 and 774 more at > 10-fold depth than the ECT and the ECT_rerun datasets, respectively (Figure 27). The number of scaffolds with a coverage breadth of 50% or less in the two separate datasets was 188 (ECT) or 218 (ECT_rerun) more than that in the combined dataset. These results indicate that the additional sequencing run (ECT_rerun) did not improve much coverage breadth or evenness, and that the two runs covered almost the same scaffolds.

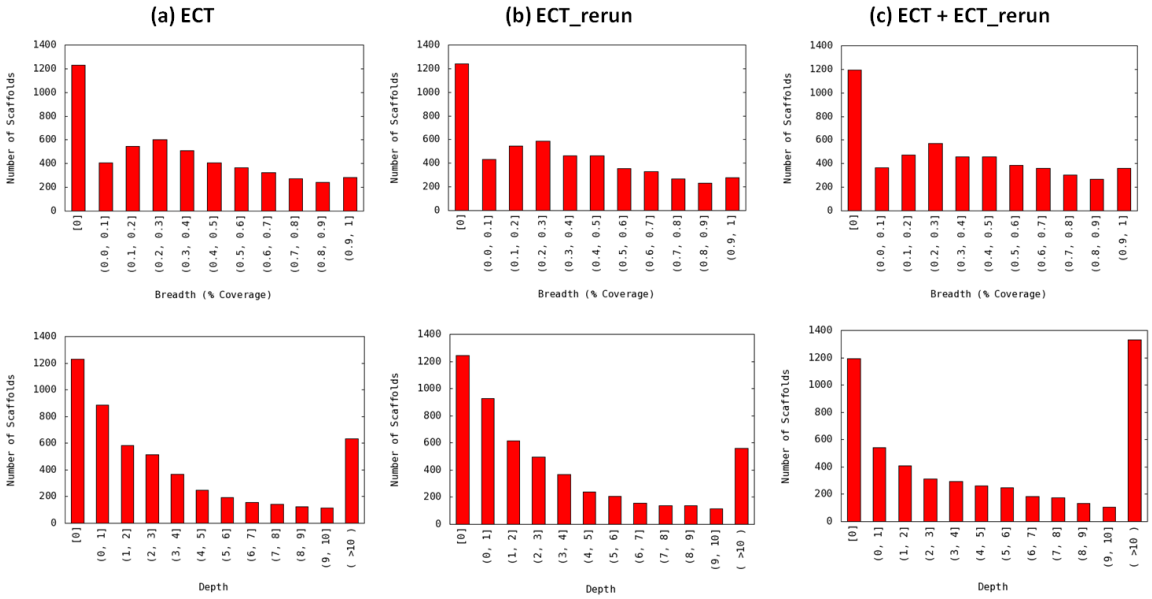


Figure 27. Distribution of scaffold coverage breadth and depth generated in the output files of the SA_Run2Ref workflow for two generated re-sequencing datasets produced for the same ECT gDNA library and their combination: (a) ECT, (b) ECT_rerun, and (c) ECT+ECT_rerun. See Table 13 for more information about the sequencing runs. Breadth and depth bins are open at the lower end and closed at the higher end, and breadth is expressed as percentage. For instance,
(0.3, 0.4] stands for 30%<breadth≤40%, and (0,1]stands for 0<depth≤1.

Table 13

Basic Statistics Produced by SA_Run2Ref for Two Sequencing Run Datasets. The two datasets of paired-end reads were generated using Illumina MiSeq by sequencing the same genomic DNA (gDNA) library prepared for a water flea (Daphnia pulex) from an ECT population. Library preparation involved shearing of extracted gDNA using a Covaris M220 focused-ultrasomicator (Woburn, MA). The average of library insert size distribution was 301 bp.

Illumina MiSeq runs (read length = 2 x 151 bp)	ECT	ECT_rerun	ECT + ECT_rerun
Total number of raw paired-end reads	7,575,822	7,064,035	14,639,857
Total number of cleaned reads	7,524,261	7,041,454	14,565,715
Total number of reads mapped to reference genome	6,573,572	6,193,164	12,766,736
Mapped/Cleaned reads (%)	87.37	87.95	87.65
Total number of scaffolds in reference genome	5,191	5,191	5,191
Number of covered reference scaffolds	3,960	3,948	3,998
Covered/Total scaffolds (%)	76.29	76.05	77.02
Genome coverage breadth (%)	64.48	64.32	66.12
Genome coverage depth	9.24	8.67	17.91
standard deviation of scaffold coverage depth	96.11	91.88	186.95
average scaffold coverage depth	16.27	15.41	31.33
Genome coverage evenness	6.79	6.86	6.82
Run time (min)	44.6	42.0	81.9

The TCO *D.pulex* library was split into two fractions: a large fraction (LF, insert size = 572 bp) and a small fraction (SF, 269 bp). Each fraction was sequenced five times along with 35 other indexed libraries in a multiplexing

fashion using Illumina Miseq, except for the fifth runs of LF (LF5) which was pooled with 5 other indexed libraries (Table 14). Hence, the quantity of reads in each LF or SF dataset was equivalent to 1/36 (or 1/6 for LF5) of a MiSeq run. As more datasets were pooled to form new reads collections as input to SA_Run2Ref, the ratio of mapped to cleaned reads remained stable at 82% to 85% (Table 14), and the scaffold coverage evenness had little change (Figure 28). Although the genome coverage depth steadily increased as more runs were added to the reads collection, the genome coverage breadth increased simultaneously until LF5 was added and then reached a plateau (Figure 28). The addition of 2.2 million SF reads raised coverage breadth by only 3% (Table 14 and Figure 28). The change in the distributions of scaffold coverage depth and breadth also supports this conclusion. Except the bin for non-covered scaffolds, the number of scaffolds in every bin increased continuously for both coverage breadth and depth from collection LF1 (Figure 29a) to LF1-5 (Figure 29b), but little difference was observed in the scaffold numbers for coverage breadth between LF1-5 and LF1-5SF1-5 collection (Figure 29c).

Table 14

Sequencing Datasets and Genome Mapping of the Daphnia pulex TCO Library. All of the NGS run datasets were generated by sequencing the TCO gDNA library which was split into two fractions: a large fraction (LF) with an average insert size of 572 bp and a small fraction (SF) with an average insert size of 269 bp. An Illumina MiSeq was used for sequencing, and both fractions were each sequenced five times in a 36× *or* 6×*multiplexing fashion, resulting in datasets LF1 to LF5 and SF1 to SF5. The reads collection were mapped to a D.pulex reference genome by running the SA_Run2Ref workflow.*

Reads collection	Sequencing runs/collection	Library fraction	Raw reads	Cleaned reads	Mapped reads	Mapped/ cleaned reads (%)	Run time (min)	Added run (multiplex, read length)
LF1	LF1	Large only	383,575	381,612	311,919	81.74	7.1	LF1 (36X,2X151)
LF1-2	LF1+ LF2	Large only	1,083,738	1,076,671	907,601	84.30	13.8	LF2 (36X,2X251)
LF1-3	LF1+ LF2+ LF3	Large only	1,782,006	1,743,523	1,478,140	84.78	21.7	LF3 (36X,2X251)

Table 14 (continued).

Reads collection	Sequencing runs/collection	Library fraction	Raw reads	Cleaned reads	Mapped reads	Mapped/ cleaned reads (%)	Run time (min)	Added run (multiplex, read length)
LF1-4	LF1+ LF2+ LF3+LF4	Large only	2,218,000	2,177,265	1,848,979	84.92	26.1	LF4 (36X,2X251)
LF1-5	LF1+ LF2+ LF3+LF4+LF5	Large only	4,242,048	4,178,856	3,524,528	84.34	45.9	LF5 (6X,2X251)
LF1-5SF1	LF1+ LF2+ LF3+LF4+LF5+ SF1	Large+ Small	4,542,917	4,478,675	3,766,787	84.10	48.1	SF1 (36X,2X151)
LF1-5 SF1-2	LF1+ LF2+ LF3+LF4+LF5_ SF1+SF2	Large+ Small	5,084,493	5,014,933	4,204,692	83.84	50.6	SF2 (36X,2X151)
LF1-5 SF1-3	LF1+ LF2+ LF3+LF4+LF5+ SF1+SF2+SF3	Large+ Small	5,530,560	5,457,878	4,561,648	83.58	52.7	SF3 (36X,2X151)

Table 14 (continued).

LF1-5 SF1-4	LF1+ LF2+ LF3+LF4+LF5+ SF1+SF2+SF3 +SF4	Large+ Small	5,920 ,185	5,845,8 27	4,872,8 85	83.36	54.8	SF4 (36X,2X151)
LF1-5 SF1-5	LF1+ LF2+ LF3+LF4+LF5+ SF1+SF2+SF3 +SF4+SF5	Large+ Small	6,411 ,123	6,333,0 54	5,270,6 16	83.22	56.5	SF5 (36X,2X151)

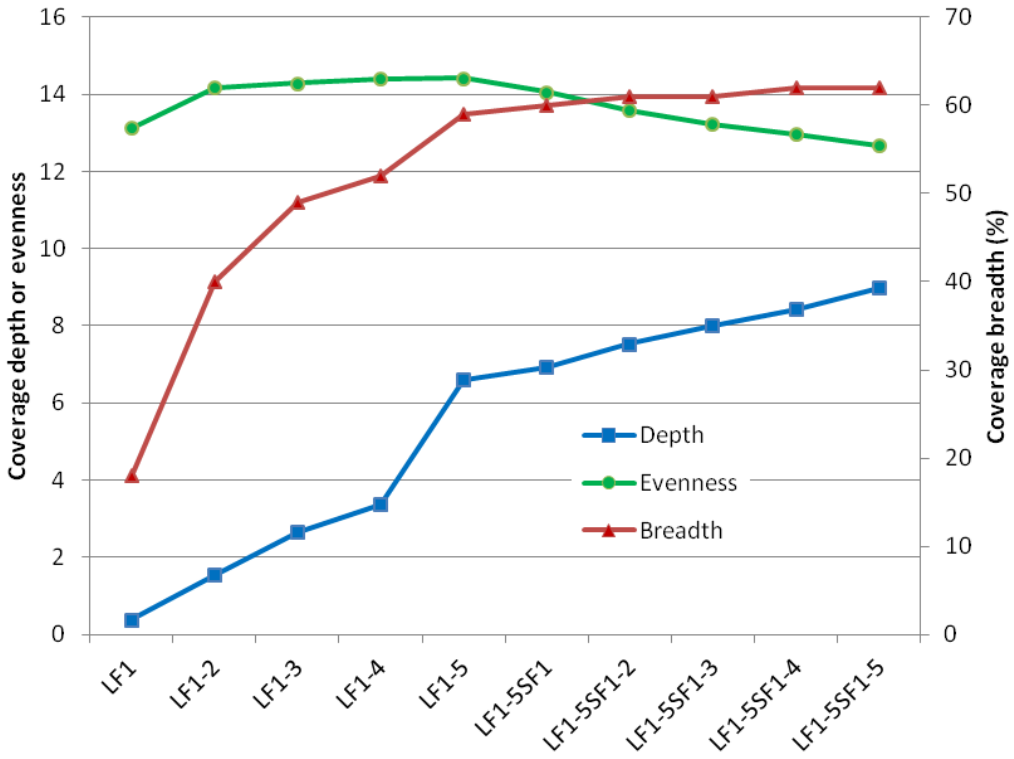


Figure 28. Change in genome coverage breadth, depth, and evenness as more sequencing runs for the same TCO library were pooled and used as the input of SA_Run2Ref. See Table 14 for the sequencing runs pooled to form reads collections.

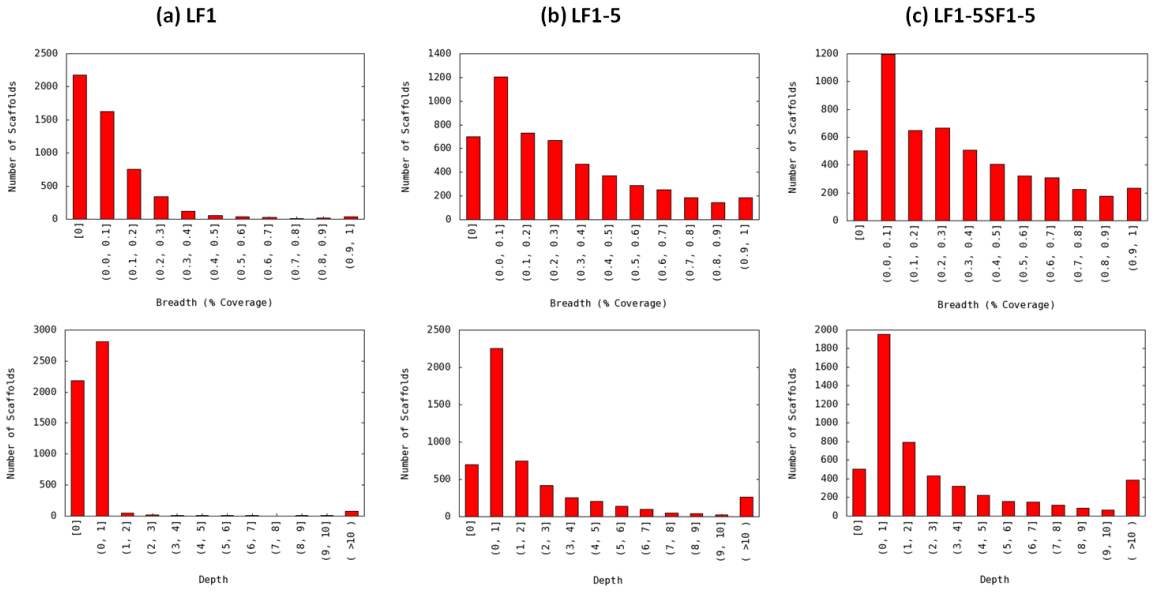


Figure 29. Change in the distribution of scaffold coverage breadth and depth as more sequencing runs for the same TCO library were pooled and used as the input of SA_Run2Ref. Shown are distributions for three reads collections: (a) LF1, (b) LF1-5, and (c) LF1-5SF1-5. See Table 14 for the sequencing runs pooled to form reads collections. Breadth and depth bins are open at the lower end and closed at the higher end, and breadth is expressed as percentage. For instance, (0.3, 0.4] stands for 30%<breadth≤40%, and (0,1]stands for 0<depth≤1.

MicroRNA Detection Using miRDisc

A microRNA is an endogenous small non-coding ribonucleic acid RNA that regulates gene expression at the post-transcriptional level. The different gene expression levels are shown in the form of genotype. Thus, microRNA is able to impact on the phenotype without the change of genotype. In this case study, the author explore microRNA for Drosophila melanogaster, Caenorhabditis elegans and earthworm *E. fetida* using miRExpress (Wang et al., 2009a), miRDeep2 (Friedlanender et al., 2012), sRNAbench (Hackenberg, 2013), and our developed software miRDisc.

Dataset Generation

Download dataset

Data for the Drosophila melanogaster and Caenorhabditis elegans are downloaded from the GEO (Gene Expression Omnibus) database:

- GSM322219: small RNAs were extracted from 2-4 days old Drosophila melanogaster (fruit fly) pupae and sequenced by the Illumina Genome Analyzer. Total number of unique reads is 385451; Maximum length is 36 nt; Minimum length is 18 nt; Average length is 22.22 nt; and Median length = 21 nt.
- GSM139137: small (~18-26 nt) RNAs were isolated using PAGE from total RNA extracted from mixed-stage, wild-type Caenorhabditis elegans (nematode worm, N2, 20 deg C) and sequenced by the 454 Genome Sequencer. Total number of unique reads is 181668; Maximum length is 85 nt; Minimum length is 1 nt; Average length is 22.25 nt; and Median length is 22 nt.

Experimental dataset

For earthworm *E. fetida*, two sequencing runs using a Solexa/Illumina Genome Analyzer I generated millions of short sequences from ERDC (U.S. Army Engineer Research and Development Center) and ECU (East Carolina University).

The following protocol in brief used by LC Sciences for the small RNA library preparation is based on the manufacturer's instructions: the

Illumina/Solexa's manual for preparing samples for analysis of small RNA by ERDC.

- Small RNA isolation by denaturing PAGE gel:

  For each sample, ~10 µg of RNA sample was size-fractionated on a 15% tris-borate-EDTA (TBE) urea polyacrylamide gel, and a 15-50 base pair fraction was excised. A small RNA fraction was eluted in 500 µL of 0.3 M NaC1 from the polyacrylamide gel slice. After elution, the small RNA fraction was precipitated by the addition of ethanol.

- Adapter ligation to the isolated small RNA:

  Based on the Illumina/Solexa's manual, the 5' RNA adapter and 3' RNA adapters were subsequently ligated to the precipitated RNA with T4 RNA ligase. Ligated RNA was size-fractionated on a 15% TBE urea polyacrylamide gel, and a 65-100 base pair fraction was excised and eluted and precipitated from the gel.

- Reverse transcript and PCR-amplification:

  The RNA was converted to single-stranded cDNA using M-MLV (Invitrogen) with the Illumina/Solexa's RT-primer. The cDNA was amplified with pfx DNA polymerase (Invitrogen) in 20 cycles PCR using Illumina/Solexa's small RNA primers set.

- Purification of amplified cDNA constructs:

  PCR products were purified on a 12% TBE polyacrylamide gel, and an 80-150 base pair fraction was excised. The excised fraction was eluted and precipitated from the gel. The purified PCR products were quantified on

the TBS-380 mini-fluorometer (Turner Biosystems) using Picogreen dsDNA quantitation reagent (Invitrogen) and diluted to 10 nM. For barcode samples, the amount of each sample was mixed equally. For example, 10 µL of 10 nM of sample A and 10 µL of 10 nM of sample B were mixed together and delivered for sequencing on the Illumina/Solexa G1 sequencer.

Then the data is cleaned using the workflow shown in Figure 30. It consists of eight steps:

- Get unique seq family: Many sequenced sequences from Solexa are exactly identical. So the identical sequences were put together into a unique seq (family), including the index, sequence, and frequency (or count, or copy#) as follows:

  23->TGGAGTGTGACAATGGTGTTTGTCGTATGCCGTCTT->18560
- A, C, G, T composition filter: If sequences of 80% A, or C, or G, or T, or 3N (it is not necessary to be consecutive), the sequence will be filtered out.
- Filter sequence data using the Adapter (ADT) dimmer filter: ADT dimmer is 5’ ADT, 3’ ADT & 5’ ADT and 3’ ADT hooked together without insertion. Unique seqs are blasted against 3’ ADT and 3DIM (3’ ADT-3’ADT). The blast data is filtered using specified parameters. Then the unique seqs containing the 3ADT or 3DIM at the beginning of unique seqs were picked up. The 3ADT part at the non-beginning position was kept after removing the 3ADT part.

- Filter sequence data using the length filter: If the length of the remained part after removing 3ADT part is >=15, the unique sequence was kept.
- Junk filter: The sequence is filtered out with the following conditions:
    - If a sequence of 7 consecutive A, 8 consecutive C, 6 consecutive G ,or 7 consecutive T, the sequence is filtered out. The number of 7, 8, 6, or 7 is from the study of miRBase. All miR sequences at miRBase do not have homo stretch AAAAAAA (7A), but do have AAAAAA (6A); all miR sequences at miRBase don't have homo stretch CCCCCCCC (8C), but do have CCCCCCC (7C). Only one miR sequence (has-miR-1225-5p) at miRBase has homo stretch GGGGGG (6G). The tolerance is one; only one miR sequence (oan-miR-1422e) at miRBase has homo stretch TTTTTTT (7T).
    - If a sequence of 10 repeat of dimmer, 6 repeat of trimer, or 5 repeat of tetramer, the sequence is filtered out.
    - If a sequence contains only A & C without G&T, the sequence is filtered out and vice versa.
- Filter sequence data using the low-copy filter: If the copy number (frequency) of a unique seq is less than 3, it is filtered out.
- Filtered sequence data using the mRNA, RFam, & repbase filter: The remained unique seqs are blasted against mRNA, RFam, & repbase. If a unique seq hits any of the mRNA, or RFam & repbase with 1 error allowed, it is filtered out.

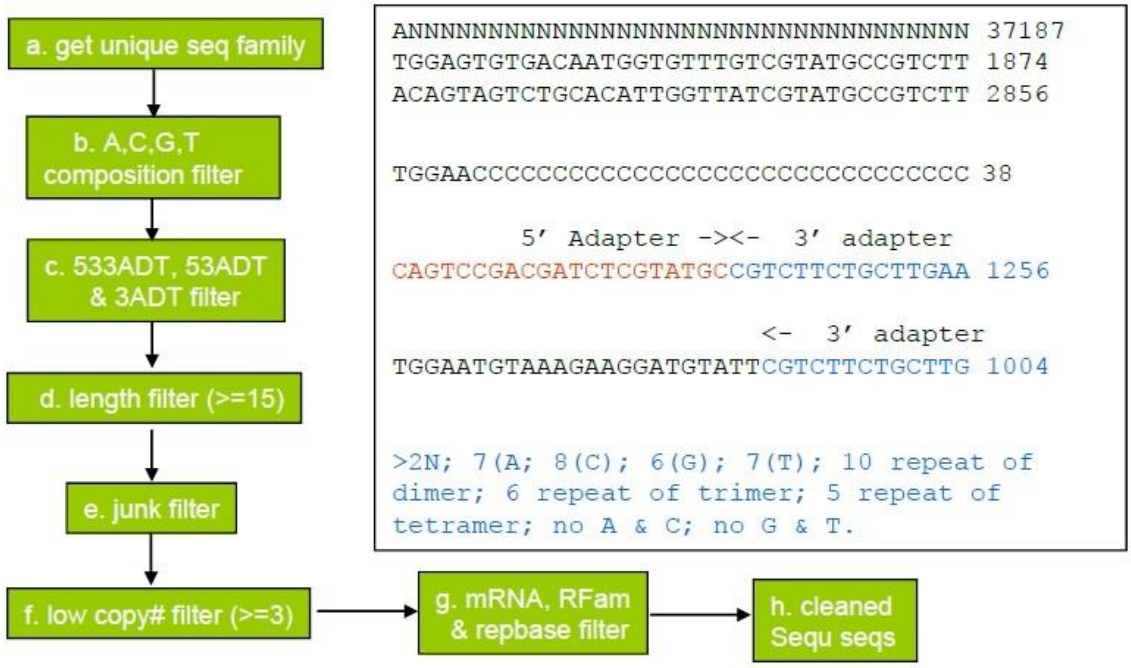


Figure 30. Clean solexa data.

After performing the above filters, the remained unique sequence data is cleaned data. For data from the ECU, similar processes for experimental design and data processing are provided. The only difference is that the sequences are filtered out with length of less than or equal to 26 in the data cleaning phase. After cleaning, there are 11591 sequences for the ERDC dataset and 96987 sequences for the ECU dataset. Then totally 108578 sequences are obtained. Maximum length is 36 nt; minimum length is 15 nt; average length is 21.8 nt; the median length is 20 nt.

Micro-RNA Identification

In this section, three datasets, including two downloaded datasets (Drosophila melanogaster and Caenorhabditis elegans) and earthworm *E. fetida*, are used to detect microRNA. Four microRNA identification software are used, including miRExpress, miRDeep2, sRNAbench and our developed software

miRDisc. Two different versions of miRBase, miRBase v14 (ftp://mirbase.org/pub/mirbase/14/), and miRBase v20 (ftp://mirbase.org/pub/mirbase/20/) are applied to all the dataset in order to compare the capability to detect known and conserved miRNA for different tools and also to measure the detection accuracy of novel miRNA for different tools. The results for three species, four softwares among two different versions of miRBase are shown below:

Caenorhabditis Elegans dataset

The original RNA sequences are filtered out by limiting copy number greater than or equal to 3, which means that any reads with a copy number less than 3 will be discarded, and reads with copy number greater than or equal to 3 will be kept for future use. The unique number of raw reads for C.elegans is 23,842, and the total number of raw reads (unique reads * count for each unique reads) is 674,456. Then the four methods mentioned above (miRExpress, miRDeep2, sRNAbench and miRDisc) are applied to this dataset. The standalone versions of tools are adopted in order to change the version of miRBase.

MiRDeep2 consists of three steps to identify known, conserved, and novel miRNAs. The first step is to obtain the index of the reference genome. In this project, the bowtie is used to index the reference genome. Then the second step is to process reads and map them to the reference genome. In the miRDeep2 package, ‘mapper.pl’ is used to finish this mapping task. Among all the parameters, parameter –c, -m, and –j are used besides other mandatory input

parameters. The last step is the miRNA detection phase, and the package uses ‘miRDeep2.pl’ to perform this function. In this step, all the C.elegans miRNAs in miRBase are used as the species database, and miRNAs other than C.elegans are used as the homolog database. After all three steps, the package generates a csv file and an html table containing all the known, conserved, and novel miRNAs. The strategy of miRExpress is to align the raw RNA sequence to the miRNA sequence in miRBase without a reference mapping procedure. It consists of four steps. First, the raw input is converted into the miRExpress file format, which contains two columns, count number, and a corresponding RNA sequence with each line separated by tab. The second step is the adaptor trimming, which can be skipped if the input file is already cleaned. The next two steps are alignment and miRNA detection. Both steps use the default parameter settings except the miRBase database. The entire miRBase is used, including all the species in the database rather than only the C.elegans. sRNAbench is a java-based package, which has a lot of functions. It is a replacement for miRanalyzer (Hackenberg, Rodriguez-Ezpeleta, & Aransay, 2011). Here,only the microRAN detection part is used. The first step is database preparation. Bowtie-build is used to get the reference index files, and makeSeqObi.jar in the package is used to compress the input short sequence file. All of the obtained files are put into the default database folder. The C.elegans miRNAs are used as the species database, and all other miRNAs are used as the homolog database. The Mature microRNA mismatch (matureMM) is set into 1 other than the default 0. The novel

microRNA detection function is activated. All of the other parameters use the default value.

After obtaining the candidates from different tools, two validation steps are applied to the known and conserved candidates, and one step is applied to the novel candidates. For some tools, they generate the known and conserved microRNA candidates based on an analysis of multiple short RNA reads, so the final candidate microRNA sequence may change a little from the input short RNA reads. Then, the first validation step for known and conserved microRNA is to map the candidates sequence back to the input short RNA reads to check whether the candidate sequence actually exists in the short RNA reads. If the candidate sequences do exist, they are kept, and they are discarded if the candidate sequences do not exist in short RNA sequence. Blastn is used in the mapping phase. The second validation step is to remove the dead microRNA from the results. The dead microRNAs are microRNAs that are detected as miRNAs in the old version of miRBase; however these are proven as false miRNAs afterward. The results are mapped to the latest version of dead miRNAs and the dead miRNAs are removed. For novel microRNA candidates, in order to validate them, the novel candidates with miRBase v14 are compared to the miRBase v20 to check how many of them exist in miRBase v20, in other words, to check how many of them are true microRNAs.

Tables 15 and 16 list all of the results from miRDeep2, miRExpress, sRNAbench, and miRDisc for miRBase v14 and v20, respectively. They contain the output directly from the tools and also the results after validation steps.

Table 15

Result of miRDeep2, miRExpress, sRNAbench, and miRDisc for C.elegans with miRBase v14: (a) known microRNA, (b) conserved microRNA, and (c) novel microRNA.

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1	number of candidates after validation step2
miRDeep2	385489	133	114	113	113
miRExpress	312968	474	122	120	120
miRDisc	41462	105	46	46	46
sRNAbench	350543	2703	121	120	120

(a)

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1	number of candidates after validation step2
miRDeep2	0	0	0	0	0
miRExpress	204014	261	378	373	373
miRDisc	77990	93	63	52	52
sRNAbench	213848	1581	282	261	261

(b)

Table 15 (continued).

	total number of aligned reads (unique*count)	unique number of aligned reads	# of candidate aligned to miRBase 20	prediction accuracy
miRDeep2	353	8	1	12.50%
miRExpress	X	X	X	X
miRDisc	2298	32	13	40.63%
sRNAbench	0	0	0	0

(c)

Table 16

Result of miRDeep2, miRExpress, sRNAbench, and miRDisc for C.elegans with miRBase v20: (a) known microRNA, (b) conserved microRNA, and (c) novel microRNA.

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1
mirdeep2	385509	135	131	93
mirexpress	318225	624	191	173
miRDisc	38740	143	78	69
sRNAbench	354143	2914	192	173

(a)

Table 16 (continued).

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1
mirdeep2	0	0	0	0
mirexpress	241586	384	697	693
miRDisc	83211	135	128	101
sRNAbench	243708	1842	576	535

(b)

	total number of aligned reads (unique*count)	unique number of aligned reads
mirdeep2	428	14
mirexpress	X	X
miRDisc	1420	19
sRNAbench		23

(c)

In miRBase, the miRNAs that share the same seed region are grouped together into families. For known and conserved candidates the detected microRNAs are grouped into families to check how many families are detected by each method. However, not every miRNA belongs to a certain family. So for those do not belong to any family, their precursors are used to stand for their temperate family name. The grouped results are shown in Tables 17 and Table 18 for miRBase v14 and v20, respectively.

Table 17

Grouped Results of Candidate microRNA for C.elegans with miRBase v14: (a) known microRNA and (b) conserved microRNA.

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	114	82	73	32	32
miRExpress	122	87	74	35	34
miRDisc	46	32	26	14	14
sRNAbench	121	87	74	34	34

(a)

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	0	0	0	0	0
miRExpress	378	373	49	5	5
miRDisc	63	54	26	9	9
sRNAbench	282	269	52	13	13

(b)

Table 18

Grouped Results of Candidate microRNA for C.elegans with miRBase v20: (a) known microRNA and (b) conserved microRNA.

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	131	106	84	25	23
miRExpress	191	162	86	29	25
miRDisc	78	64	36	14	14
sRNAbench	192	162	86	30	25

(a)

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	0	0	0	0	0
miRExpress	697	682	67	15	15
miRDisc	128	109	36	19	19
sRNAbench	576	549	69	27	27

(b)

Drosophila melanogaster dataset

The basic steps and parameter settings for the Drosophila dataset are similar to the process of C.elegans dataset. The unique number of raw reads for C.elegans is 54,078, and the total number of raw reads (unique reads * count for

each unique reads) is 1,750,122. Original results and grouped results are shown in Tables19 - 22.

Table 19

Result of miRDeep2, miRExpress, sRNAbench, and miRDisc for Drosophila with miRBase v14: (a) known microRNA, (b) conserved microRNA, and (c) novel microRNA.

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1	number of candidates after validation step2
miRDeep2	172908	51	47	45	45
miRExpress	305034	605	131	115	115
miRDisc	86606	54	19	19	19
sRNAbench	160574	1331	60	55	55

(a)

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1	number of candidates after validation step2
miRDeep2	0	0	0	0	0
miRExpress	296803	620	1240	1158	1158
miRDisc	89908	87	313	292	292
sRNAbench	125665	989	600	561	561

(b)

Table 19 (continued).

	total number of aligned reads (unique*count)	unique number of aligned reads	# of candidate aligned to miRBase 20	prediction accuracy
miRDeep2	1436	4	0	0
miRExpress	X	X	X	X
miRDisc	6452	48	20	41.67%
sRNAbench		14		

(c)

Table 20

Result of miRDeep2, miRExpress, sRNAbench, and miRDisc for Drosophila with miRBase v20: (a) known microRNA, (b) conserved microRNA, and (c) novel microRNA.

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1
miRDeep2	172913	52	50	42
miRExpress	351694	857	215	119
miRDisc	90370	77	31	30
sRNAbench	172705	1573	105	95

(a)

Table 20 (continued).

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA	number of candidates after validation step1
miRDeep2	0	0	0	0
miRExpress	335602	930	1907	1862
miRDisc	93017	116	503	470
sRNAbench	150831	1307	1023	945

(b)

	total number of aligned reads (unique*count)	unique number of aligned reads
miRDeep2	253	7
miRExpress	X	X
miRDisc	1257	27
sRNAbench		10

(c)

Table 21

Grouped Results of Candidate microRNA for Drosophila with miRBase v14: (a) known microRNA and (b) conserved microRNA.

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	47	32	26	15	15
miRExpress	131	84	63	47	45

Table 21 (continued).

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDisc	19	11	10	8	8
sRNAbench	60	37	29	23	22

(a)

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	0	0	0	0	0
miRExpress	1240	1221	70	19	19
miRDisc	313	301	20	12	12
sRNAbench	600	584	27	16	16

(b)

Table 22

Grouped Results of Candidate microRNA for Drosophila with miRBase v20: (a) known microRNA and (b) conserved microRNA.

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	50	50	40	0	0
miRExpress	215	197	96	18	16

Table 22 (continued).

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDisc	31	30	19	1	1
sRNAbench	105	99	45	6	5

(a)

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	0	0	0	0	0
miRExpress	1907	1856	101	51	43
miRDisc	503	488	29	15	15
sRNAbench	1023	999	45	24	24

(b)

Earthworm dataset

The basic steps and parameter settings for the earthworm dataset are similar to the process of the above two datasets. The unique number of raw reads for the earthworm is 40,696, and the total number of raw reads (unique reads * count for each unique reads) is 1,809,040. The difference is the miRBase database. Since earthworm does not exist in miRBase, the species database for earthworm should be set as none or empty, and the homolog database is the entire miRBase. Thus, the results only contain conserved microRNA and novel

microRNA candidates. Original results and grouped results are shown in Tables 23-26.

Table 23

Result of miRDeep2, miRExpress, sRNAbench, and miRDisc for earthworm with miRBase v14: (a) conserved microRNA and (b) novel microRNA.

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA
miRDeep2	0	0	0
miRExpress	43535	298	741
miRDisc	539	30	10
sRNAbench	985607	1419	220

(a)

	total number of aligned reads (unique*count)	unique number of aligned reads	# of candidate aligned to miRBase 20	prediction accuracy
miRDeep2	238415	13	3	23.08%
miRExpress	X	X	X	X
miRDisc	56608	35	0	0
sRNAbench		5		

(b)

Table 24

Result of miRDeep2, miRExpress, sRNAbench, and miRDisc for earthworm with miRBase v20: (a) conserved microRNA and (b) novel microRNA.

	total number of aligned reads (unique*count)	unique number of aligned reads	number of identified microRNA
miRDeep2	0	0	0
miRExpress	122836	364	1166
miRDisc	2890	25	11
sRNAbench	990041	1504	324

(a)

	total number of aligned reads (unique*count)	unique number of aligned reads
miRDeep2	238942	17
miRExpress	X	X
miRDisc	54509	34
sRNAbench		6

(b)

Table 25

Grouped Results of Candidate Conserved microRNA for earthwrom with miRBase v14.

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	0	0	0	0	0

Table 25 (continued).

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRExpress	741	725	40	16	16
miRDisc	10	7	6	3	3
sRNAbench	220	210	10	10	10

Table 26

Grouped Results of Candidate Conserved microRNA for earthwrom with miRBase v20.

	number of identified microRNA	number of identified microRNA has family	number of identified family	number of identified microRNA without family	number of precursor for microRNA without family
miRDeep2	0	0	0	0	0
miRExpress	1166	1149	40	17	17
miRDisc	11	6	5	5	5
sRNAbench	324	311	14	13	13

Summary

The miRDsic shows pretty good performance on novel miRNA detection, not good on known and conserved miRNA detection for all the three species. There are several reasons that cause this situation. The most important one is

that the design for miRDisc is based on the biosynthesis principle, while other existing methods are mostly based on sequence comparison. Such logic is good for novel miRNA discovery. However, the strict filtering conditions decrease the number of candidates for known and conserved miRNA. Furthermore, mapping procedure and folding process are two significant steps in the pipeline. Then the accuracy for mapping algorithm and Unafold greatly affects the results.

CHAPTER VI

CONCLUSIONS

Summary and Conclusions

High-throughput next-generation sequencing (NGS) technologies are capable of generating massive amounts of data in the form of paired-end or single-end reads with either fixed or variable lengths. This prompts the development of analysis software or tools for next-generation sequencing data. Here, the author has developed SeqAssist, SVDisc, and miRDisc to analyze next-generation DNA/RNA sequencing data.

SeqAssist is a useful and informative tool that can serve as a valuable "assistant" to a broad range of investigators who conduct genome re-sequencing, RNA-Seq, or *de novo* genome sequencing and assembly experiments. It consists of three separate workflows: (1) the SA_RunState workflow generates basic statistics about an NGS dataset, including numbers of raw, cleaned, redundant and unique reads, redundancy rate, and a list of unique sequences with length and read count; (2) the SA_Run2Ref workflow estimates the breadth, depth, and evenness of genome-wide coverage of the NGS dataset at ta nucleotide resolution; and (3) the SA_Run2Run workflow compares two NGS datasets to determine the redundancy between the two NGS runs.

SVDisc is a novel and integrative SV discovery pipeline that provides an all-in-one toolkit for investigators who are interested in identifying SVs in their studied species from genome re-sequencing data. The novelty of SVDisc lies in the fact that there is no similar pipeline or infrastructure available in the SV

research community. It can detect all of the common types of SVs with user-defined sizes, including insertion, deletion, duplications, inversion, intra-chromosomal, and inter-chromosomal translocations.

miRDisc was developed as a novel method to predict known, conserved, and novel miRNAs, especially to predict the miRNAs in transcriptome enriching species.

Future Work

All of the three tools described in this dissertation are very useful for analyzing biological dataset and provide important information, which will help researchers with further analysis. However, these tools can be improved for better performance. Visualization features can be added to the output of SeqAssist, such as the distribution figure of depth. This figure uses different colors to present different depth levels and marks the position of depth. With this figure, users are able to read the results more easily. Then for miRDisc, in order to improve the number of known and conserved candidates, the logic of the right pipeline can be replaced by simplifying the strategy with sequence comparisons of mature and precursor sequence.

REFERENCES

(n.d.). Retrieved from http://www.lifetechnologies.com/us/en/website-overview/ab-welcome.html

(n.d.). Retrieved from https://code.google.com/p/mosaik-aligner/

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*, 1061-1073.

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*, 56-65.

Abel, H. J., Duncavage, E. J., Becker, N., Armstrong, J. R., Magrini, V. J., & Pfeifer, J. D. (2010). SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. *Bioinformatics*, *26*, 2684-2688.

Abyzon, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011a). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*, 974-984.

Abyzov, A., & Gerstein, M. (2011b). AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics*, *27*, 595-603.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, *12*, 363-376.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*, 403-410.

Ambros, V. (2004). The functions of animal microRNAs. *Nature*, *431*, 350-355.

An, J., Lai, J., Lehman, M. L., & Nelson, C. C. (2013). miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Research*, *41*, 727-737.

Bartel, D. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, *116*, 281-297.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., . . . Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*, *37*, 766-770.

Berezikov, E., Guryev, V., van de Belt, J., Wienhold, E., Plasterk, R. H., & Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, *120*, 21-24.

Bernardi, G., Wiley, E. O., Mansour, H., Miller, M. R., Orti, G., Haussler, D., . . . Venkatesh, B. (2012). The fishes of Genome 10K. *Marine Genomics*, *7*, 3-6.

Boisvert, S., Laviolette, F., & Corbeil, J. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, *17*, 1519-1533.

Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the reads length matter? *Genome Research*, *19*, 336-346.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., . . . Mardis, E. R. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*, 677-681.

Chevreus, B., Pfisterer, T., Drescher, B., Drisesl, A. J., Muller, W. E., Wetter, T., & Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, *14*, 1147-1159.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., . . . Lander, E. S. (2009). High-resolution mapping of copy-number alternations with massively parallel sequencing. *Nature Methods*, *6*, 99-103.

Cock, P. J., Field, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality score, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*, 1767-1771.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*, 133-138.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*, 85-97.

Friedlaender, M. R., Mackowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). miRDeep2 accurately identifies know and hundreads of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, *40*, 37-52.

Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., & Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*, *26*, 407-415.

Griffiths-Jones, S. (2004). The microRNA registry. *Nucleic Acids Research*, *32*, D109-D111.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, *34*, D140-D144.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research*, *36*, D154-D158.

Hackenberg, M. (2013). *sRNAbench*. Retrieved from http://arn.ugr.es/srnabench/

Hackenberg, M., Rodriguez-Ezpeleta, N., & Aransay, A. M. (2011). miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research*, *39*, W132-W138.

Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J. M., & Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research*, *37*, W68-W76.

Handsaker, R. E., Korn, J. M., Nemesh, J., & McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics*, *43*, 269-276.

He, L., & Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, *5*, 522-531.

Hendrix, D., Levine, M., & Shi, W. (2010). miRTRAP: a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biology*, *11*, R39.

Homer, N., Merriman, B., & Nelson, S. F. (2009). BFAST: An alignment tool for large scale genome resequencing. *PLoS ONE*, *4*, e7767.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., . . . Sahinalp, S. C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, *26*, i350-i357.

i5K Consortium. (2013). The i5K Initiative: advancing arthoropod genomics for knowledge, human health, agriculture, and the environment. *The Journal of Heredity*, *104*, 595-600.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, *35*, W339-W344.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., . . . Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research*, *37*, D98-D104.

Korbel, J. O., Abyzon, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., & Gerstein, M. B. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, *10*, R23.

Lai, E. C., Tomancak, P., Williams, R. W., & Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. *Genome Biology*, *4*, R42.

Langmead, B., Trapnell, C., Pop, M., & Salzbeng, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*, R25.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., . . . Church, D. M. (2013). DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Research*, *41*, D936-D941.

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C.elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*, 843-854.

Levine, R. (2011). i5K: The 5,000 Insect Genome Project. *The America Entomological Society*, *72*, 110-113.

Li, H., & Drubin, R. (2009a). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*, 2078-2079.

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics*, *24*, 713-714.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., & Wang, J. (2010). De novo assebly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*, 265-272.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., . . . Fan, W. (2012). Comparison of the two major classed of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Functional Genomics*, *11*, 25-37.

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., & Bartel, D. P. (2003a). Vertebrate microRNA genes. *Science*, *299*, 1540.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., . . . Bartel, D. P. (2003b). The microRNAs of Caenorhabditis elegans. *Genes & Development*, *17*, 991-1008.

Liu, L., Li, Y., Li, S., Hu, N. He, Y., Pong, R., . . . Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, *25*, 1364.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387-402.

Mardis, E. R. (2013). Next-generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, *6*, 287-303.

Margulies, M., Egholm, M., & Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., . . . Rothberg, J. M. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature*, *437*, 376-380.

Maroney, P. A., Chamnongpol, S., Souret, F., & Nilsen, T. W. (2007). A rapid, quantitative assay for direct detection of microRNAs and other small RNAs using splinted ligation. *RNA*, *13*, 930-936.

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*, 671-682.

Mathelier, A., & Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, *26*, 2226-2234.

Mayer, P., Farinelli, L., & Kawashima, E.H. (2013). Patent: Method of nucleic acid amplification. Retrieved from http://www.google.com.tw/patents/US8476044

Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, *6*, S13-S20.

Metzker, M. L. (2010). Sequencing technologies-the next generation. *Nature Reviews Genetics*, *11*, 31-46.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, *95*, 315-327.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., & Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*, 59-65.

Nam, J. W., Kim, J., Kim, S. K., & Zhang, B. T. (2006). ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Research*, *34*, W455-W458.

Needleman, S. B., & Wunsch, C. D. (1970). A general method application to the search for similarities in the amimo acid sequence of two proteins. *Journal of Molecular Biology*, *48*, 443-453.

Ning, Z., Cox, A. J., & Mullikin, J. C. (2001). SSAHA: a fast search method for large DNA database. *Genome Research*, *11*, 1725-1729.

Osherovich, L. (2010). *2 chasing 1*. Available from http://www.nature.com/scibx/journal/v3/n11/full/scibx.2010.331.html

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., . . . Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, *408*, 86-89.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*, i333-i339.

Ritchie, W., Theodule, F. X., & Gautheret, D. (2008). Mireval: a web tool for simple microRNA prediction in genome sequences. *Bioinformatics*, *24*, 1394-1396.

Sanger, F., & Coulson, A. R. (1975). A rapid method for determing sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*, 441-448.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*, 5463-5467.

Scientists 10K Community of Scientists. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity*, *100*, 659-674.

Sharp, A. J., Mefford, H. C., Li, K., Baker, C., Skinner, C., Stevenson, R. E., . . . Eichler, E. E. (2008). A recurrent 15p13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*, *40*, 322-328.

Shaw-Smith, C., Pittman, A. M., Willatt, L., Martin, H., Rickman, L., Gribble, S., . . . Carter, N. P. (2006). Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nature Genetics*, *38*, 1032-1037.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, *19*, 1117-1123.

Sindi, S. S., Onal, S., Peng, L. C., Wu, H. T., & Raphael, B. J. (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biology*, *13*, R22.

Smith, T., & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*, 195-7.

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., . . . Stratton, M. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature*, *486*, 400-404.

Stratton, M. (2008). Genome resequencing and genetic variation. *Nature Biotechnology*, *26*, 65-66.

Suzuki, S., Yasuda, T., Shiraishi, Y., Miyano, S., & Nagasaki, M. (2011). ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics*, *12*, S7.

Taylor, E., & Gant, T. (2008). Emerging fundamental roles for non-coding RNA species in toxicology. *Toxicology*, *246*, 34-39.

Wang, W. C., Lin, F. M., Chang, W. C., Lin, K. Y., Huang, H. D., & Lin, N. S. (2009). miRExpress: Analyzing high-throughput sequencing data for profilling microRNA expression. *BMC Bioinformatics*, *10*, 328.

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., & Li, Y. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, *21*, 3610-3614.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*, 57-63.

Weinberg, M. S., & Wood, M. J. (2009). Short non-coding RNA biology and neurodegenerative disorders: novel disease targets and therapeutics. *Human Molecular Genetics*, *18*, R27-R39.

Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., . . . Daly, M. J. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *The New England Journal of Medicine*, *358*, 667-675.

Williams, R., Peisajovich, S. G., Miller, O. J., Magdassi, S., Tawfik, D. S., & Griffiths, A. D. (2006). Amplification of complex gene libraries by emulsion PCR. *Nature*, *3*, 545-550.

Wong, P., Wiley, E., Johnson, W., Ryder, O., O'brien, S., & Haussler, D. (2012). Tissue sampling methods and standards for vertebrate genomics. *Gigascience*, *1,* 8.

Wu, Y., Wei, B., Liu, H., Li, T., & Rayner, S. (2011). MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics*, *12*, 107.

Xie, F., Xiao, P., Chen, D., Xu, L., & Zhang, B. (2012). miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Molecular Biology*, *80*, 75-84.

Xue, C., Li, F., He, T., Liu, G. P., Li, Y., & Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local strucure-sequene features and support vector machine. *BMC Bioinformatics*, *6*, 310.

Yang, J. H., Shao, P., Zhou, H., Chen, Y. Q., & Qu, L. H. (2010). deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Research*, *38*, D123-D130.

Yang, X., & Li, L. (2011). miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics*, *27*, 2614-2615.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertion from paired-end short reads. *Bioinformatics*, *25*, 2865-2871.

Yoon, S., Xuan, Z., Makaron, V., Ye, K., & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, *19*, 1586-1592.

Yousef, M., Showe, L., & Showe, M. (2009). A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification. *The FEBS Journal*, *276*, 2150-2156.

Zhai, J., Jeong, D. H., De Paoli, E., Park, S., Rosen, B. D., Li, Y., . . . Meyers, B. C. (2011). MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes & Development*, *25*, 2540-2553.

Zhu, E., Zhao, F., Xu, G., Hou, H., Zhou, L., Li, X., . . . Wu, J. (2010). mirTools: microRNA profiling and discovery based on high-throughput sequencing. Nucleic Acids Research, *38*, W392-W397.