Chapter 16

An Overview of the Analysis of Next Generation Sequencing Data

Andreas Gogol-Döring and Wei Chen

Abstract

Next generation sequencing is a common and versatile tool for biological and medical research. We describe the basic steps for analyzing next generation sequencing data, including quality checking and mapping to a reference genome. We also explain the further data analysis for three common applications of next generation sequencing: variant detection, RNA-seq, and ChIP-seq.

Key words: Next generation sequencing, Read mapping, Variant detection, RNA-seq, ChIP-seq

1. Introduction

In the last decade, a new generation of sequencing technologies revolutionized DNA sequencing (1). Compared to conventional Sanger sequencing using capillary electrophoresis, the massively parallel sequencing platforms provide orders of magnitude more data at much lower recurring cost. To date, several so-called next generation sequencing platforms are available, such as the 454-FLX (Roche), the Genome Analyzer (Illumina/Solexa), and SOLiD (Applied Biosystems); each having its own specifics. Based on these novel technologies, a broad range of applications has been developed (see Fig. 1).

Next generation sequencing generates huge amounts of data, which poses a challenge both for data storage and analysis, and consequently often necessitates the use of powerful computing facilities and efficient algorithms. In this chapter, we describe the general procedures of next generation sequencing data analysis with a focus on sequencing applications that use a reference sequence to which the reads can be aligned. After describing how to check the sequencing quality, preprocess the sequenced reads, and map the sequenced reads to a reference, we briefly

Junbai Wang et al. (eds.), *Next Generation Microarray Bioinformatics: Methods and Protocols*,
Methods in Molecular Biology, vol. 802, DOI 10.1007/978-1-61779-400-1_16,

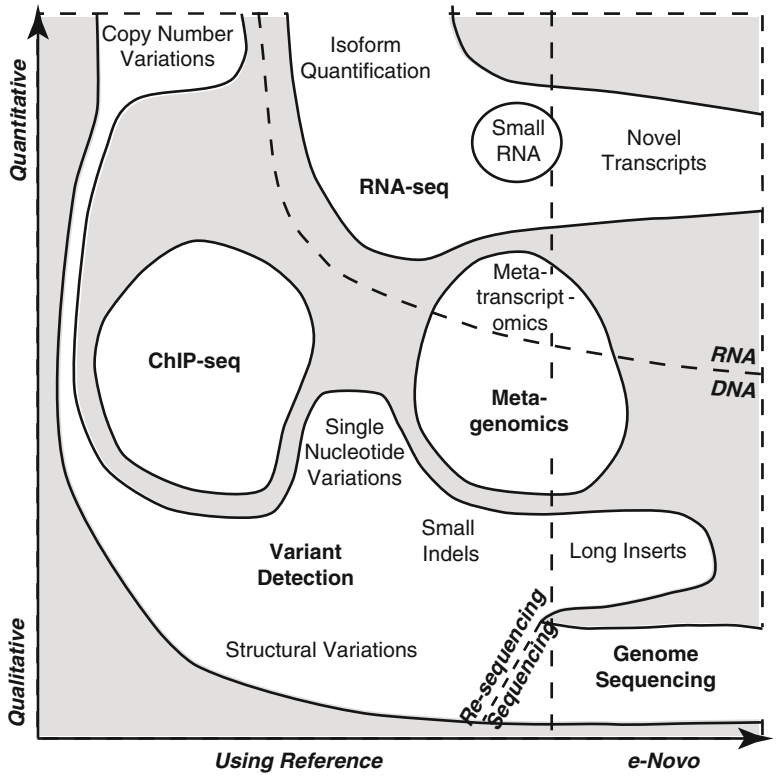

Fig. 1. Illustration of some common applications based on next generation sequencing. The decoding of new genomes is only one of various possibilities to use sequencing. Variant detection, ChIP-seq, and RNA-seq are discussed in this book. Metagenomics (16) is a method to study communities of microbial organisms by sequencing the whole genetic material gathered from environmental samples.

discuss three of the most common applications for next generation sequencing.

1. *Variant detection* (2) means to find genetic differences between the studied sample and the reference. These differences range from single nucleotide variants (SNVs) to large genomic deletions, insertions, or rearrangements.
2. *RNA-seq* (3) can be used to determine the expression level of annotated genes as well as to discover novel transcripts.
3. *ChIP-seq* (4) is a method for genome-wide screening protein–DNA interactions.

2. Methods

2.1. General Read Processing

Current next generation sequencing technologies based on photochemical reactions recorded on digital images, which are further processed to get sequences (reads) of nucleotides or, for SOLiD,

dinucleotide "colors" (5) (base/color calling). The sequencing data analysis starts from files containing DNA sequences and quality values for each base/color.

1. Check the overall success of the sequencing process by counting the raw reads, i.e., spots (clusters/beads) on the images, and the fraction of reads accepted after base calling (filtered reads). These counts could be looked up in a results file generated by the base calling software. A low number of filtered reads could be caused by various problems during the library preparation or sequencing procedure (see Note 1). Only the filtered reads should be used for further processing. For more ways to test the quality of the sequencing process see Notes 2 and 3.
2. Sequencing data are usually stored in proprietary file formats. Since some mapping software tools do not accept these formats as input, a script often has to be employed to convert the data into common file formats such as FASTA or FASTQ.
3. The sequenced DNA fragments are sometimes called "inserts" because they are wrapped by sequencing adapters. The adapters are partially sequenced if the inserts are shorter than the read length, for example, in small RNAs sequencing (see Subheading 2.4, step 5). In these occasions, it is necessary to remove the sequenced parts of the adapter from the reads, which could be achieved by removing all read suffixes that are adapter prefixes (see Note 4).

2.2. Mapping to a Reference

Many applications of next generation sequencing require a reference sequence to which the sequenced reads could be aligned. Read mapping means to find the position in the reference where the read matches with a minimum number of differences. This position is hence most likely the origin of the sequenced DNA fragment (see Note 5).

1. There are numerous tools available for read mapping (6). Select a tool that is appropriate for mapping reads of the given kind (see Note 6). Some applications may require special read mapping procedures that, for example, allow small insertions and deletions (indels) or account for splicing in RNA-seq.
2. Select an appropriate maximum number of allowed errors (see Note 7).
3. For most applications you only need uniquely mapped reads, i.e., reads matching to a single "best" genomic position. If nonuniquely mapped reads could also be useful, then consider to specify an upper bound for the number of reported mapping positions, because otherwise the result list is blown up by reads mapping to highly repetitive regions.

4. Most mapping tools create output files in proprietary formats, so we advice to convert the mapping output into a common file format such as BED, GFF, or SAM (7, 8).
5. Count the percentage of all reads which could be mapped to at least one position in the reference. A low amount of mappable reads could indicate a low sequencing quality (see Note 3) or a failed adapter removal (see Note 4).
6. Some pieces of DNA could be overamplified during library preparation (PCR artifacts) resulting in a stack of redundant reads that are mapped to the same genomic position and same strand. If it is necessary to get rid of such redundancy, discard all but one read mapped to the same position and on the same strand.
7. Transform SOLiD reads into nucleotide space after mapping.

2.3. Application 1: Variant Detection

The detection of different variation types requires different sequencing formats and analysis strategies. Tools are available for the detection of most variant types (2) (see Note 8).

1. For detecting SNVs, search the mapped reads for bases that are different from the reference sequence. Since there will probably be more sequencing errors than true SNVs, each SNV candidate must be supported by several independent reads. A sufficient coverage is therefore required (see Note 9). Note that some SNVs might be heterozygous, which means that they occur only in some of the reads spanning them.
2. Structural variants can be detected by sequencing both ends of DNA fragments (paired-end sequencing) (see Fig. 2) (9). After mapping the individual reads independently to the reference, estimate the distribution of fragment lengths. Then search for read pairs which were mapped to different chromosomes or have abnormal distance, ordering, or strand orientation. Search for a most parsimonious set of structural variants explaining all discordant read pairs. The more read pairs can be explained by the same variant, the more reliable this variant is and the more precise the break point(s) could be determined. If only one end of a DNA fragment could be mapped to the reference, the other end is possibly part of a (long) insertion. Given a suitable coverage, the sequence of the insertion can possibly be determined by assembling the unmapped reads.

2.4. Application 2: RNA-seq

The experimental sequencing protocols and hence the data analysis procedures are usually different for longer RNA molecules such as mRNA (Subheading 2.4 steps 2 and 3) and small RNA such as miRNA (Subheading 2.4 steps 5 and 6).

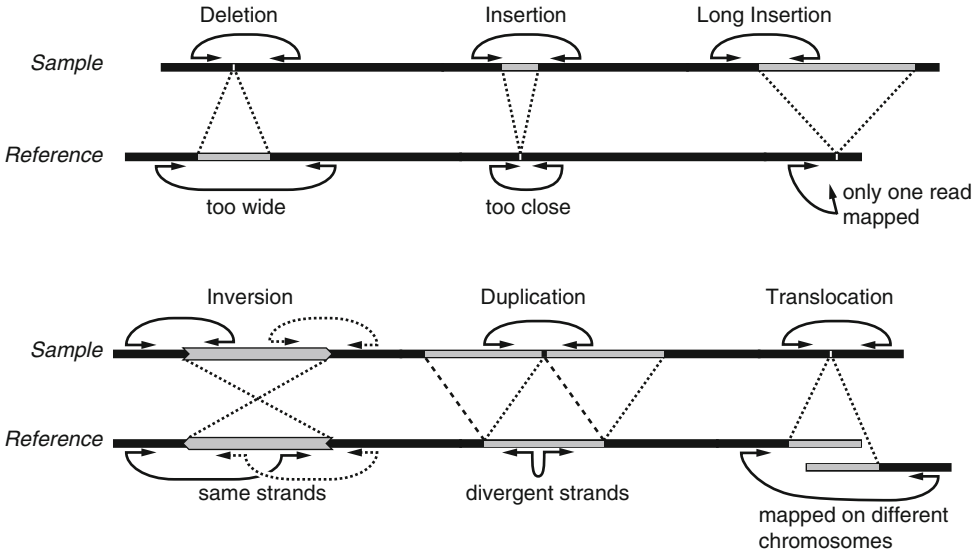

Fig. 2. Different variant types detected by paired-end sequencing (9). (1) Deletion: The reference contains a sequence that is not present in the sample. (2–3) Insertion and Long Insertion: The sample contains a sequence that does not exist in the reference. (4) Inversion: A part of the sample is reverse compared to the reference. (5) Duplication: A part of the reference occurs twice in the sample (tandem repeat). (6) Translocation: The sample is a combination of sequences coming from different chromosomes in the reference. Note that the pattern for concordant reads varies depending on the sequencing technologies and the library preparation protocol.

1. Check the data quality. Classify the mapped reads on the basis of available genome annotation into different functional groups such as exons, introns, rRNA, intergenic, etc. For example, in the case of sequencing polyA-RNA, only a small fraction of reads should be mapped to rRNA.
2. Determine the expression level of annotated genes by counting the reads mapped to the corresponding exons, and then divide these counts by the cumulated exon lengths (in kb) and the total number of mapped reads (in millions). The resulting RPKM ("reads per kilobase of transcript per million mapped reads") can be used for comparing expression levels of genes in different data sets (10).
3. To quantify different splicing isoforms, select reads belonging exclusively to certain isoforms, for example, reads mapping to exons or crossing splicing junctions present only in a single isoform. From the amounts of these reads infer a maximum likelihood estimation of the isoform expression levels.
4. To discover novel transcripts or splicing junctions, use a spliced alignment procedure to map the RNA-seq reads to a reference genome. Then find a most parsimonious set of transcripts that explains the data. Alternatively, you could first assemble the sequencing reads and then align the assembled

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.