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,

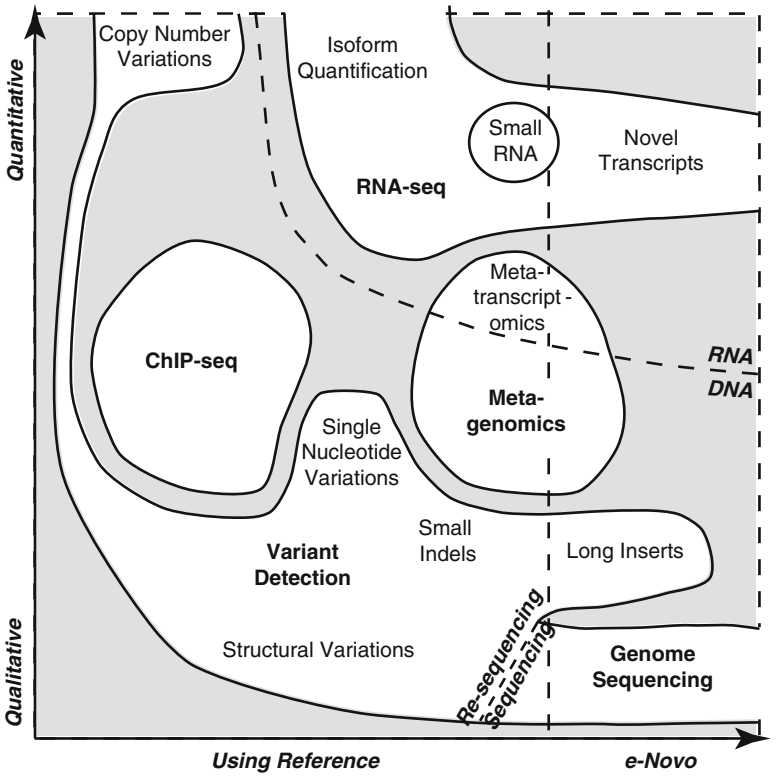

Fig. 1. Illustration of some common applications based on next generation sequencing. The decoding of new genomes is only one of various possibilities to use sequencing. Variant detection, ChIP-seq, and RNA-seq are discussed in this book. Metagenomics (16) is a method to study communities of microbial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

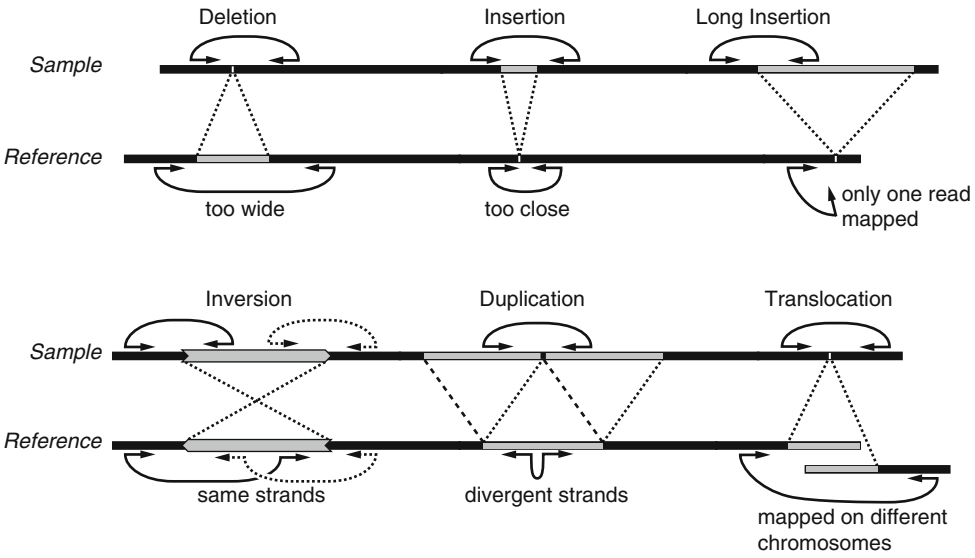

Fig. 2. Different variant types detected by paired-end sequencing (9). (1) Deletion: The reference contains a sequence that is not present in the sample. (2–3) Insertion and Long Insertion: The sample contains a sequence that does not exist in the reference. (4) Inversion: A part of the sample is reverse 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

contigs to the genome (11). In both cases, it is advisable to sequence long paired-end reads.

5. Small RNA-seq reads are first preprocessed to remove adapter sequences (see Subheading 2.1, step 3). To profile known miRNA, the reads could then be mapped either to the genome or to the known miRNA precursor sequences (12). Do not remove redundant reads (see Subheading 2.2, step 6) when analyzing this kind of data. The expression level of a specific miRNA could be estimated given the number of redundant sequencing reads mapped to its mature sequence (see Note 10). Normalize the raw read counts by the total number of mapped reads in the data set (see Subheading 2.4, step 2 and Note 11).
6. To discover novel miRNAs, use a tool such as miRDeep (13), which uses a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor.

2.5. Application 3: ChIP-seq

In ChIP-seq, chromatin immunoprecipitation uses antibodies to specifically select the proteins of interest together with any piece of randomly fragmented DNA bound to them. Then the precipitated DNA fragments are sequenced. Genomic regions binding to the proteins consequently feature an increased number of mapped sequencing reads.

1. Use a "peak calling" tool to search for enriched regions in the ChIP-seq data (10) (see Note 12). ChIP-seq data should be evaluated relative to a control data set obtained either by sequencing the input DNA without ChIP or by using an antibody with unspecific binding such as IgG (see Note 9).
2. An alternative way to analyze the data that is especially suited for profiling histone modifications is to determine the normalized read density (RPKM) of certain genomic areas such as genes or promoter regions. This method is similar to the analysis of RNA-seq data (see Subheading 2.4, step 2).

3. Notes

1. In some cases, the sequencing results could be improved by manually restarting the base calling using nondefault parameters. For example, choosing a better control lane when starting the Illumina offline base caller could boost up the number of successfully called sequencing reads. Candidates for good control lanes feature a nearly uniform base

distribution (see Note 2). Note that for this reason a flow cell should never be filled completely with, e.g., small RNA libraries, since these are not expected to produce uniform base distributions.

2. Check the base/color distribution over the whole read length. If the sequenced DNA fragments are randomly sampled from the genome – for example, sequencing genomic DNA, ChIP-seq, or (long) RNA-seq libraries – then the bases should be nearly uniformly distributed for all sequencing cycles. The software suite provided by the instrument vendors usually creates all relevant plots.
3. The base caller annotates each base with a value reflecting its putative quality. These values could be used to determine the number of high/low quality bases for each cycle. The overall quality of sequenced bases normally declines slowly toward the end of the read. A drop of quality for a single cycle could be a hint for a temporary problem during the sequencing.
4. Since the sequenced adapter could contain errors, it is reasonable to allow some mismatches during the adapter search. Note that there is a trade-off between the sensitivity and the specificity of this search.
5. In order to avoid wrongly mapped reads, it is important to use a reference as accurate and complete as possible. All possible sources of reads should be present in the reference.
6. Not all tools can handle SOLiD reads in dinucleotide color space; Roche 454 reads may contain typical indels in homopolymer runs. When mapping the relatively short reads created by or Illumina Genome Analyzer or SOLiD, it is usually sufficient to consider only mismatches, unless it is planned to detect small indels.
7. We recommend to choose a mapping strategy that guarantees accurate mappings rather than to maximize the mere number of mapped reads. Next generation sequencing usually generates huge quantities of reads, so a negligible loss of reads is certainly affordable. Consequently, most mapping tools are optimized to allow only a small number of mismatches. Higher error numbers are only necessary if the reads are long or if we are especially interested in variations between reads and reference.
8. Check the success of your experiment by comparing your results to already known variants deposited in public data bases such as dbSNP (14) and the Database of Genomic Variants (15).

9. Sequencing reads are *never* uniformly distributed throughout the genome, and any statistical analysis assuming this is inaccurate. Some parts of the genome usually are covered by much more reads than expected, whereas some other parts are not sequenced at all. The experimenter should be aware of this fact, for example, when planning the required read coverage for variant detection. Moreover, this effect certainly impacts quantitative measurements such as ChIP-seq or RNA-seq. ChIP-seq assays, for example, should always include a control library (see Subheading 2.5, step 1), and in a RNA-seq experiment, it is easier to compare expression levels of the same gene in different circumstances rather than the expression level of different genes in the same sample.
10. Note that the actual sequenced mature miRNA could be shifted by some nucleotides compared to the annotation in the miRNA databases.
11. One problem of this normalization method is that sometimes few miRNAs get very high read counts, which means that any change of their expression level could affect the read counts of all other miRNAs. In some cases, a more elaborated normalization method could therefore be necessary.
12. Most tools for analyzing ChIP-seq data focus on finding punctuate binding sites (peaks) typical for transcription factors. For ChIP-seq experiments targeting broader binding proteins, like polymerases or histone marks such as H3K36me3, use a tool that can also find larger enriched regions. In order to precisely identify protein binding sites, it is often necessary to determine the average length of the sequenced fragments. Some ChIP-seq data analysis tools estimate the fragment length from the sequencing data. Keep in mind that this is not trivial, because ChIP-seq data usually consist of single-end sequencing reads. Therefore, always check whether the estimated length is plausible according to the experimental design.

References

1. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature Biotechnology 26:1135–1145
2. Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nature Methods 6:S13-S20
3. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621–628
4. Johnson DS, Mortazavi A, Myers RM et al (2007) Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316 (5830):1497–1502
5. Fu Y, Peckham HE, McLaughlin SF et al. SOLiD Sequencing and 2-Base Encoding. http://appliedbiosystems.com
6. Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nature Methods 6:S6-S12
7. UCSC Genome Bioinformatics. Frequently Asked Questions: Data File Formats. http://genome.ucsc.edu/FAQ/FAQformat.html
8. Sequence Alignment/Map (SAM) Format. http://samtools.sourceforge.net/SAM1.pdf
9. Korbel JO, Urban AE, Affourtit JP et al. (2007) Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. Science 318 (5849):420–426
10. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nature Methods 6:S22-S32
11. Haas BJ, Zody MC (2010) Advancing RNA-seq analysis. Nature Biotechnology 28:421–423
12. Griffiths-Jones S, Grocock RJ, van Dongen S et al (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research 34:D140-D144. http://microrna.sanger.ac.uk
13. Friedländer MR, Chen W, Adamidi C et al (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnology 26:407–415
14. dbSNP. http://www.ncbi.nlm.nih.gov/projects/SNP
15. Database of Genomic Variants. http://projects.tcag.ca/variation
16. Handelsman J, Rondon MR, Brady SF et al (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5:245–249