Bioinformatics analysis pipeline for exome sequencing data

Christos Hatzis

Background

Next generation sequencing (NGS), also referred to as high throughput DNA sequencing (HTS), is a versatile and reliable platform for large-scale DNA sequencing, allowing single labs to sequence at a reasonable coverage the entire genome of an individual within a couple of weeks and at a comparatively low cost (1). Due to its ability to sequence massive amounts of DNA, HTS has become an extremely useful and flexible tool, enabling researches to investigate a number of questions related to the genetic sequence, including identification of known and novel single nucleotide polymorphisms (SNPs), structural variants such as deletions, insertions or inversions, or even to examine gene expression levels by first converting mRNA to cDNA and then sequencing (RNA-seq). For these reasons, HTS has been considered the Swiss pocketknife of molecular biology (2).

A key application of NGS is detection of genetic variants associated with disease phenotypes. In these analyses, the DNA sequence of a carrier of the disease-related phenotype (e.g. a tumor) is compared to reference sequence from healthy individuals to determine genetic aberrations that might be "driving" the disease. Although protein coding genes constitute only about 1 percent of the genome, such regions typically harbor 85 percent of the mutations associated with diseases. This enrichment suggests that a more efficient strategy for identifying disease-related (functional) mutations is to sequence only the complete coding regions of genomes, called the "whole exome." Efficient methods to capture and sequence the whole exome have been developed and are already having an impact in clinical diagnosis (3, 4). Here, we review the key bioinformatics steps required to process, clean and assemble the DNA sequence from HTS whole exome sequencing data and, finally, to identify the functional mutations that might have important clinical implications in disease-specific prognosis and management.

The high throughput DNA sequencing pipeline

Compared to standard sequencing approaches, HTS data come with many peculiarities and biases that are a direct result of the workflows employed to enable massively parallel sequencing. A recent review by Altmann et al. (2) provides an informative yet accessible overview of the HTS process. Briefly, DNA is randomly fragmented to smaller segments allowing millions of sequencing reactions to be performed in parallel on each of these fragments (called reads), thus achieving extremely high sequencing throughput. The nucleotide sequence of each of the DNA reads is then determined by synthesis, during which a base-specific light signal is emitted when a nucleotide is incorporated in the complementary strand of the DNA fragment. Because incorporation of a single nucleotide does not emit a strong enough signal to be detected reliably, DNA fragments are first fixed on a substrate and amplified to form clusters or colonies of
Next Generation Sequencing: From Bench to Bedside

millions of identical DNA fragments. During a synthesis cycle, millions of identical nucleotides are incorporated in each of the clones in the cluster, allowing detection of the emitted amplified light signal and identification of the complementary nucleotide. The next synthesis step is then repeated to detect the next nucleotide in the sequence. In the early cycles of synthesis, the same nucleotide is incorporated in synchrony to all DNA clones within the read cluster. However this process can get partially out of synchrony after a number of cycles have been completed, resulting in a mixed light signal and an increased probability of calling the base erroneously. Thus, bases towards the end of a read have reduced reliability, which effectively limits the size of the DNA fragments that can be sequenced by HTS technologies.

The high throughput potential of NGS workflows comes with a high analytical burden of stitching together the small DNA fragments (reads) sequenced by these platforms to recover the original sequence, while accounting for the errors and potential biases introduced by the HTS platform. These tasks are not trivial and require advanced bioinformatics tools to recover the genome sequence and detect sequence variants that might be associated with the disease. The tools are used in a series of steps starting with raw sequence data and ending with a list of variants, and are organized in scripted workflows called pipelines. Set up and use of such tools usually requires significant programming and bioinformatics skills and a substantial computational infrastructure. Several excellent reviews cover available pipelines for NGS (2, 5–7). The basic steps in a typical NGS pipeline are described below.

Step 1: Base calling. Synthesis of new DNA strands from each template is captured as a series of fluorescence images, one from each synthesis cycle. Base-calling algorithms infer the nucleotide base incorporated in each step from the fluorescence intensity signal detected at each cluster of reads. A measure of uncertainly is assigned to each base call and is typically transformed to a standard Phred quality score. A Phred score of 30 corresponds to a 0.1 percent error rate in base calling. Platform-specific software is typically used for this step.

Step 2: Quality control of generated sequence data. Characteristics of the sequence reads are summarized and checked against expected values. Examples include average quality score per position in the read, over-represented sequences, deviation from expected GC content, or distribution of nucleotides per read position. Tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) or Qualimap (8) can be used.

Step 3: Alignment and assembly. Short reads are typically aligned against a reference genome to assemble the contiguous DNA sequence of the genome or the exome. The accuracy of alignment significantly impacts the effectiveness of variant detection, as incorrectly aligned reads could lead to errors in SNP calling. There are several alignment tools and the choice of the right tool often involves a compromise between speed and accuracy. The Stampy aligner provides the best tradeoff between accuracy and running time) (6).

Step 4: Alignment post processing. Post processing includes elimination of identical reads (starting at exactly the same position and having the same length), as these most likely represent artifacts of the PCR step used to amplify DNA fragments prior to base synthesis. Inclusion of these reads will lead to biased genotype assessment. Also reads that do not map to a unique location on the reference genome are removed, as the origin of these reads cannot be determined. The program Picard, which is part of SAMtools (9) is typically used for this step.
**Step 5: Quality score recalibration.** Base quality scores are critical inputs for the variant calling algorithms and the Phred-like quality scores provided by the sequencing platform often do not reflect true base error rate. Therefore, true base quality is reevaluated based on empirical mismatch rates on sites in the reference genome that do not have any reported SNPs. This recalibrated quality scores are used in subsequent steps. The GATK package offers this functionality (10).

**Step 6: Variant and genotype calling.** Deviations from reference sequence are called sequence variants. Uncertainty about a variant call is usually determined based on probabilistic approaches that use the concept of genotype likelihoods and can effectively integrate prior information about genotypes or variants at the same location extracted from existing databases such as dbSNP, base quality scores, read coverage and could even incorporate linkage disequilibrium information if variants are being called on multiple individuals at the same time. GATK, SAMtools, and VarScan 2 (11) are the tools most commonly used for somatic variant calling.

**Step 7: Filtering SNP candidates.** SNPs are filtered to remove potential false positive calls based on quality (recalibrated quality > 30) minimum depth, or quality per depth (>5). Other filtering steps include strand bias or unusual transition versus transversion ratios (typically 2 for non-synonymous and 5 for synonymous variants) (5).

**Discussion**

The plethora of tools available for NGS sequence data analysis (7) and the innumerable ways in which they can be combined makes selection of the right tool for each step of NGS sequence analysis a challenge, even for users experienced in bioinformatics. An alternative for beginners would be to use publicly available integrated pipelines such as HugeSeq (12), a fully integrated pipeline for NGS analysis and annotation of variants, SIMPLEX (13), and TREAT (14) which are both fully integrated and available on the Amazon cloud environment (EC2) or the WEP pipeline (15), which is available to non-IT expert users through a web interface. Alternatively, pipelines can be integrated using user-interface based workflow systems such as Galaxy (16) or Taverna (17).

**Future Directions**

A pipeline combines the steps required to call sequence variants from HTS sequence data. However the end goal of such efforts is to identify mutations that contribute to disease processes, such as oncogenesis, tumor metastasis or response to therapy. The critical step is, therefore, to link somatic variants with their predicted impact on the disease phenotype. This step is by far the most difficult and requires not only to predict the effect that a single mutation might have on the protein function but also whether a given functional mutation is irrelevant to a disease phenotype or whether it confers a selective advantage to the cells carrying it (driver mutation). This is currently a very active area of research (18) and one that will ultimately define the impact of NGS technologies on understanding complex genetic diseases such as cancer.
Next Generation Sequencing: From Bench to Bedside

References