memo (2019) 12:111–114
https://doi.org/10.1007/s12254-019-0479-4

memo
magazine of european medical oncology

Educational no.5: next-generation gene panel sequencing

Wolfgang Kranewitter

Received: 4 February 2019 / Accepted: 6 February 2019 / Published online: 21 February 2019

Summary Next-generation sequencing allows the simultaneous interrogation of a large number of genomic targets and is therefore highly suitable for situations, in which mutations in many different genes may have a clinical impact. This short educational discusses the process from the sample to the reported mutations and the limits of the method.

Keywords Next-generation sequencing · Libarary preparation · Variant interpretation · Bioinformatic analysis · Hemato-oncology

Introduction

“Next-generation sequencing” (NGS, also called “massive parallel sequencing”) has changed the process in molecular diagnostics rapidly and is about to replace Sanger sequencing. Currently targeted sequencing methods are widely used in hemato-oncology, which focus on genes known to be involved in the pathophysiology of myeloid or lymphatic neoplasms.

NGS consists roughly of three steps: (1) library preparation, (2) sequencing, and (3) bioinformatic analysis. For a clinical report this is completed by the classification of the variants.

Library preparation

In a first step the regions of interest must be isolated from the rest of the genome. Generally, two different ways exist to accomplish this. For enrichment libraries the regions of interest are isolated from the randomly fragmented DNA by hybridization to complementary bait probes. Amplicon libraries are an alternative approach, which can be used for smaller panels. This basically involves the amplification of the target regions by a highly multiplexed PCR. Libraries can also be prepared from RNA after conversion to cDNA. This will allow the detection of gene fusions or aberrant mRNA splicing.

Sequencing

There are several companies offering a variety of sequencing machines, which use different techniques for the sequencing. Here, I refer mainly to the widely used Illumina sequencers (Illumina Inc, San Diego, CA, USA) and will add a short note on Ion Torrent sequencers (Thermo Fischer Scientific, Waltham, MA, USA).

The library is loaded on to a sequencing device (flow cell). This consists of a glass surface covered with a lawn of oligonucleotides. These oligonucleotides are complementary to the adapter sequence, which has been attached during library preparation and allow the immobilization of the library molecules on the flow cell (Fig. 1a, b). In a step called bridge amplification each immobilized DNA fragment of the library is amplified in a way that the amplification products are also immobilized in a cluster around the original fragment (Fig. 1c, d). Therefore, each of the millions of clusters represents one DNA fragment of the original library. A sequencing primer is extended by incorporation of the complementary, fluorescently labelled nucleotide analogs (Fig. 1e). The fluorescent signals of all clusters are recorded after each incorporation of a nucleotide (Fig. 1f). Usually, after 75–300 cycles the bridge amplification product is reversed and the complementary strands are sequenced from the other end.

W. Kranewitter (✉)
Labor für Molekularbiologie und Tumorzytogenetik,
Ordensklinikum Linz Barmherzige Schwestern,
Seilerstätte 4, 4010 Linz, Austria
wolfgang.kranewitter@ordensklinikum.at

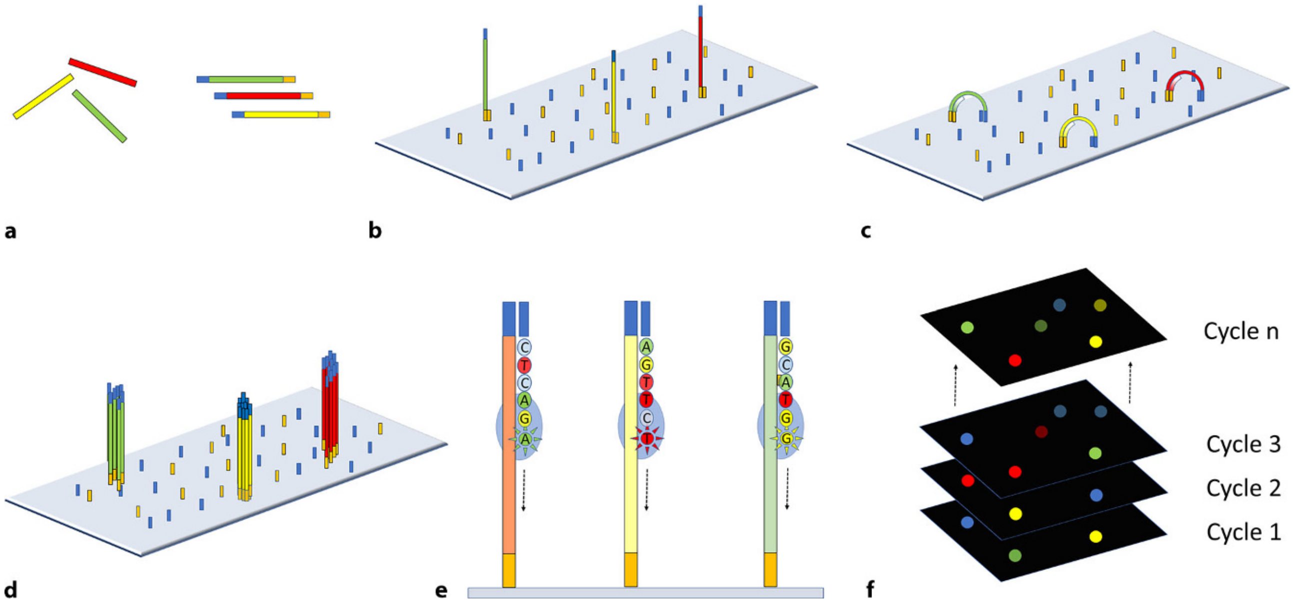

Fig. 1 Schematic of the sequencing process. **a** Adapters are attached to the library fragments. **b** The library DNA fragments are attached by hybridization of the adapter sequences to complementary oligonucleotides on the flow-cell. **c**, **d** The individual DNA fragments are amplified by bridge amplification to form clusters of (theoretically) identical molecules. **e**, **f** After primer extension with fluorescent oligonucleotide analogs (one nucleotide per cycle) a picture is taken from the signal of all the clusters

Ion Torrent sequencers use the H^+ ion which is released during the incorporation of a nucleotide for detection of the sequence. These are recorded on a sequencing matrix with millions of CMOS (complementary metal-oxide semiconductor) sensors, which can detect the small change in pH caused by the release of the H^+ ion. Ion Torrent sequencers can have a read length up to 400 bp.

Bioinformatic analysis

The data recorded by the sequencers are either image data (the images taken from the fluorescent clusters on the flow cells) or the data from the CMOS sensors from Ion Torrent sequencers. In the first step of the analysis—the base calling—these data are converted to the sequence information for each cluster. The sequence information generated from a single cluster is referred to as a "read". After trimming to remove adapter sequences and low-quality sections the reads are aligned to the human genome, i. e. for each read its position within the genome is determined. This is in many cases straight forward, however poses more of a challenge for larger indels, especially for duplications: a sequencing read must contain a sufficient number of bases on both sides flanking the duplication in order to be correctly aligned. Considering the limited read length (very often 150 bp) the larger the duplication the lower is the chance of this occurring. Repeat regions and genomic regions with a very high degree of sequence similarity to an intended target (like pseudogenes) pose another challenge for a correct alignment. It may be impossible for the alignment algorithm to decide whether a read belongs to a target or to an unintended pseudogene.

From the aligned reads (Figs. 2 and 3) the variants are called by algorithms, which are usually different for somatic and germline situations. Germline callers rely on the underlying assumption of a genotype being either normal, heterozygous or homozygous (and nothing in between). A somatic caller cannot rely on this as the tumour content and therefore the content of mutated cells can vary and in addition subclonal mutations can occur.

In order to reduce technical artefacts (false positive variants), which potentially can obscure the sequencing results, quality filters are applied. These filters consider the probability of a variant to be a real variant by using parameters like the sequencing quality of a single base, the number and the alignment quality of the reads containing the variant, or the presence of a variant in forward and reverse reads.

A variant is defined as any difference between a given reference sequence (the human reference genome) and the sequence information in the reads. Therefore, the list of variants generated contains not only the "mutations" relevant to the patient's disease but also common or private polymorphisms and eventually still artefacts.

Assessment of variants

The variants need to be classified according to their significance. For germline variants a 5-tier scheme ranging from benign to pathogenic (intermediate classifications being likely benign, unknown or likely

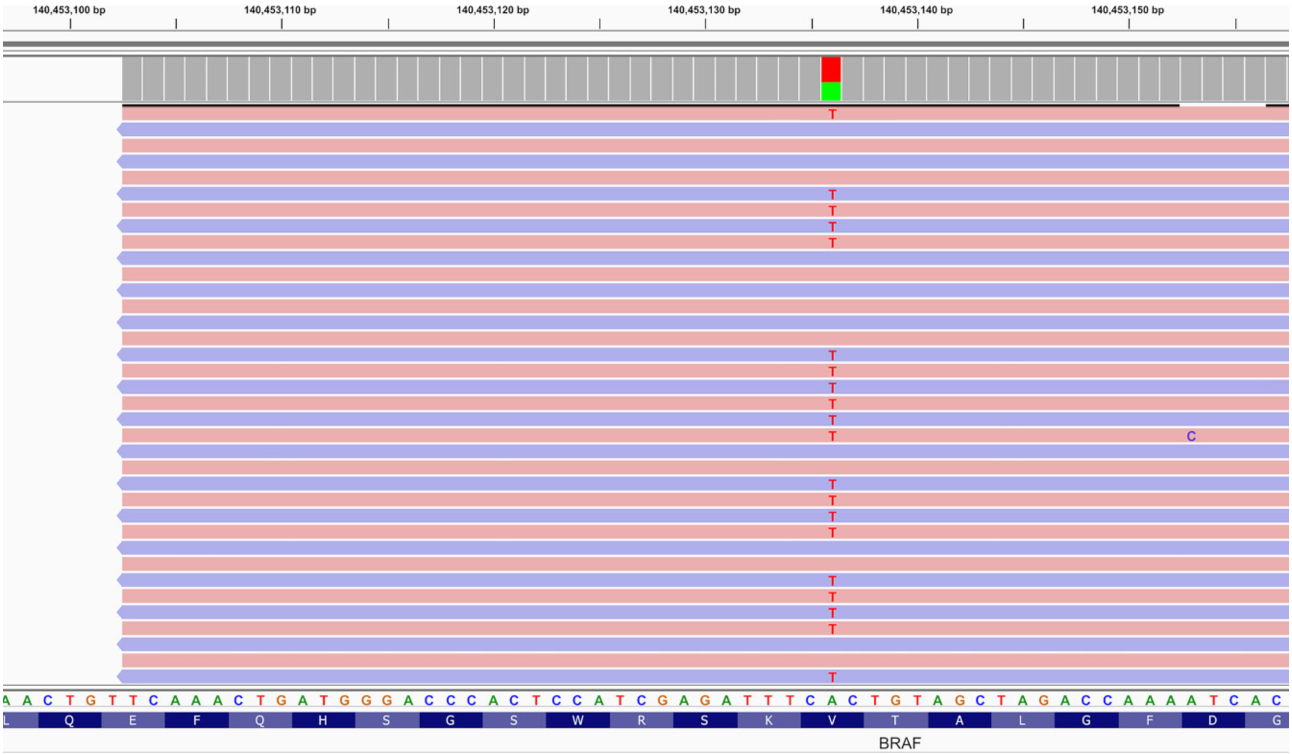
140,453,100 bp
140,453,110 bp
140,453,120 bp
140,453,130 bp
140,453,140 bp
140,453,150 bp
T
C
A A C T G T T C A A A C T G A T G G G A C C C A C T C C A T C G A G A T T T C A C T G T A G C T A G A C C A A A A T C A C
L Q E F Q H S G S W R S K V T A L G F D G
BRAF

Fig. 2 Visualization of aligned sequencing reads from an amplicon library. Individual reads are represented by *horizontal bars* in *pink* (forward direction) and *light-blue* (reverse). Differences to the reference sequence are displayed as *colored letters*. Reads show a BRAF p.V600E mutation. Note, that in an amplicon library, all the reads of a target region start at the same position

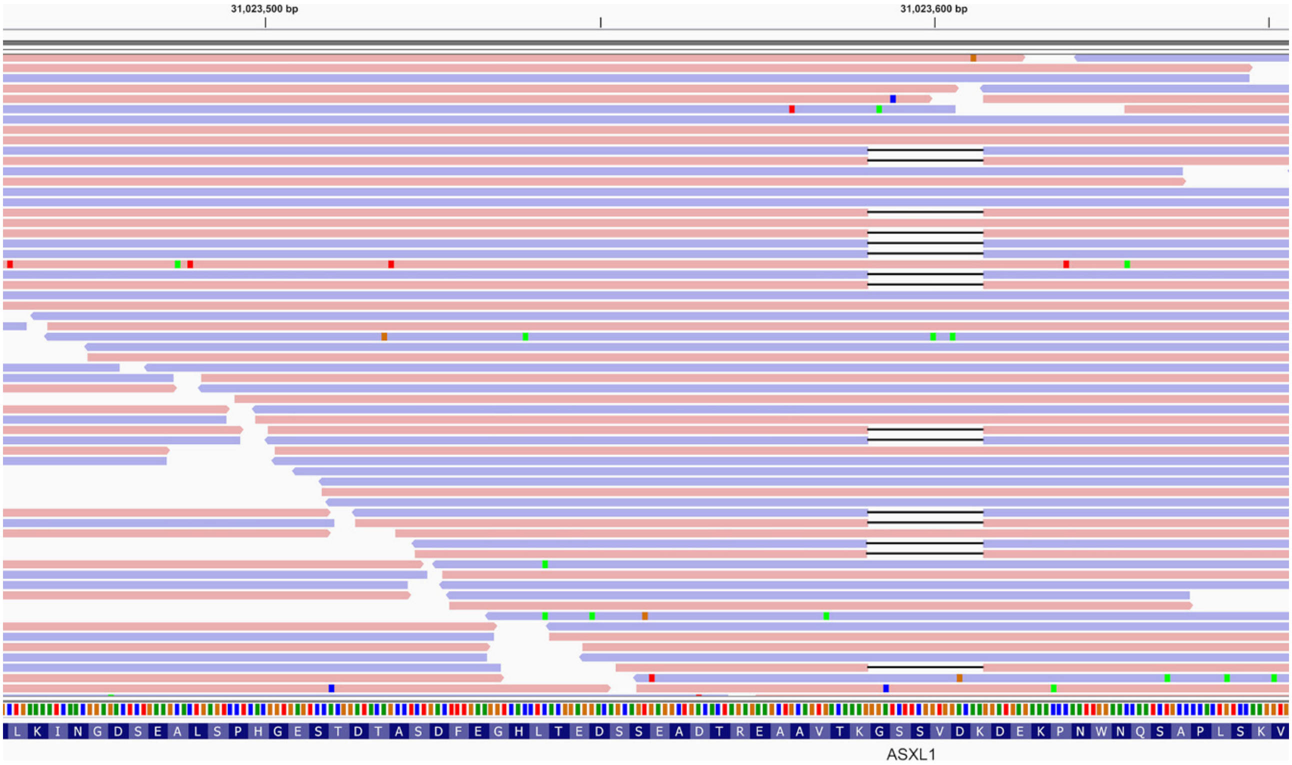
31,023,500 bp
31,023,600 bp
L K I N G D S E A L S P H G E S T D T A S D F E G H L T E D S S E A D T R E A A V T K G S S V D K D E K P N W N Q S A P L S K V
ASXL1

Fig. 3 Visualization of aligned sequencing reads from an enrichment library. A 17 bp deletion in ASXL1 is shown as *horizontal black lines*. The reads start at random positions, typical for enrichment libraries

pathogenic) is well established [1]. For the somatic situation in cancer a 4-tier scheme has been proposed, classifying variants as variants of strong clinical significance (tier I), potential clinical significance (tier II), unknown clinical significance (tier III), and benign or likely benign variants (tier IV) [2]. Significance in this context covers either therapeutic (predictive), prognostic or diagnostic significance. Tier I and II may even be differentiated further according to the level of evidence available for the significance.

The task of classification of variants is usually still done by humans and may be a bottle-neck in the diagnostic process, as larger gene panels are used more frequently. To illustrate this: in our 34 gene panel for hemato-oncologic diseases, the unfiltered list of variants usually contains about 200–300 variants. After algorithmic removal of most-likely artificial variants approximately about 30–60 variants still remain to be assessed. Most of them are benign germline variants and only a handful (in most cases 1–5) turn out to be of clinical importance. Quite often also a conclusive classification of the significance cannot be reached due to the lack of functional information about this variant.

Limits of the method

NGS works straight forward for the detection of small variants (single nucleotide changes, deletions or insertions of a view bases). For larger indels, especially when their size gets larger than the read length, the sensitivity becomes lower. Deletions or duplications of complete exons can only be discovered by a copy number analysis, in which the number of reads covering a targeted region is compared to copy number normal samples. In gene panel sequencing structural variants (large chromosomal changes) can usually only be discovered with RNA panels, in which the resulting gene fusions can be detected.

The limit of detection (LOD, i.e. the percentage of DNA molecules in a sample carrying a specific variant) depends among other factors on (1) the read depth (the number of reads covering a position, also termed as "coverage") and (2) the ability of the analysis pipeline to distinguish real low frequency variants from technical artefacts. For somatic mutations a LOD of 1–5% can usually be reached. By means of sophisticated library preparation methods and specialized bioinformatic tools, a LOD down to 0.1% (or even lower) may be reached [3].

In this short educational only the surface of this topic can be touched. The complex interplay of technical, biological, bioinformatic and clinical aspects can make the generation of a meaningful sequencing report a demanding task. For deeper insights the interested reader is referred to references [4, 5].

Conflict of interest W. Kranewitter declares that he has no competing interests.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24.
2. Li MM, Datto M, Duncavage EJ, et al. Standards and guidelines for the interpretation and reporting of sequence variants in Cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017;19(1):4–23.
3. Ståhlberg A, Krzyzanowski PM, Egyud M, Filges S, Stein L, Godfrey TE. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. Nat Protoc. 2017;12(4):664–82.
4. Le Gallo M, Lozy F, Bell DW. Next-generation sequencing. Adv Exp Med Biol. 2017;943:119–48.
5. Yohe S, Thyagarajan B. Review of clinical next-generation sequencing. Arch Pathol Lab Med. 2017;141(11):1544–57.