Structural Variations in Genetic Sequences

Mark Louis*

Department of Molecular Cell Biology, Katholieke Universiteit Leuven, Belgium

Introduction

While crucial for genomic interpretation, SV identification has historically proven to be a challenge. Numerous SVs have been found thanks to detection techniques that use ensemble algorithms and cutting-edge sequencing technology that overcome short-read constraints. This has revealed details about their prevalence, connections to diseases and potential impacts on biological functions. Multiplatform discovery is required to resolve the entire spectrum of variation due to the heterogeneity in SV type and size as well as specific detection biases of new genomic platforms. Single nucleotide variants (SNVs), tiny insertions and deletions (indels; 50bp) and structural variations (SVs) are frequently found as differences between individuals, according to widespread application of whole-genome high throughput sequencing (HTS) for the identification of genetic variants.

Description

SVs alter more of the genome per nucleotide change than any other kind of sequence variant due to their tremendous diversity in type and size, ranging from 50 bp to well over megabases of sequence. They consist of numerous subclasses that include balanced rearrangements like inversions and inter- and intrachromosomal translocations as well as unbalanced copy number variants (CNVs) such deletions, duplications and insertions of genetic material. Additionally, SVs include segmental duplications, multi-allelic CNVs with highly variable copy numbers, mobile element insertions and complicated arrangements that combine some of the aforementioned events. Every human genome contains SVs, which have an impact on transcriptional apparatus, molecular and cellular functioning, regulatory mechanisms and 3D structure. Therefore, to understand the genetics of physiological and pathological processes, it is important to increase our understanding of SV structure and prevalence. Short-read signatures are used by many of the common tools and algorithms to detect SVs to infer their presence when compared to a reference genome. The limited sequence and insert sizes of conventional short-read HTS prevent SV detection from fully overcoming the resolution power of short-read methods for SNVs. Due to technical challenges in identifying the precise structures of SVs given their significant variability and close closeness to repeated regions, there are still significant restrictions on what can be accomplished in SV analysis. Due to their smaller size than SVs, SNVs discovered by short-reads can be sequence-resolved during the discovery step, whereas most SVs would require computational inference after the fact. As a result, modern genomics has examined SNVs relative to SVs to a substantially greater extent. For example, extensive functional data from genome-wide association studies, reliable detection systems, high-quality reference sets and defined best practices are all available for SNV research. Progress in SV analysis, in contrast, has lagged far behind since detection is insufficient and reference sets are shallow, diverse and deficient in sample size.

Platforms that produce reads several orders of magnitude longer than those produced by short-read HTS have emerged as a result of a significant increase in the development and accessibility of novel sequencing technologies that utilize, among other things, protein pores, advanced microfluidics and specialized flow cells. This allows for the direct detection of numerous SVs. We utilize data obtained from other genomic platforms in addition to short-read SV callers as a way to fully detect the wide range of SVs. We highlight the individual methodologies, their applications and new findings because each approach has different merits. The majority of sequencing-based SV identification relies on signs that come from mismatches in mapping between a sample read and the reference genome: Split-read (SR) approaches use alignments that map over breakpoints; read-pair (RP) approaches evaluate the orientation and distance of paired-ends; read-depth (RD) approaches identify deletions or duplications based on divergences in mapping depth; and alternatively, de novo or local assembly (AS) reassembles contigs before pairwise comparison to a reference.

However, recent technological and methodological advancements have allowed for significant progress. It is now possible to produce readings of several thousand base pairs thanks to long-read sequencing technology, in particular thanks to Pacific Biosciences (PacBio) and Oxford Nanopore technologies (ONT), which can even produce reads as long as two million base pairs (MBP). However, the longer reads and higher mistake rates of modern long-read technologies can present additional methodological difficulties. A major advancement in addition to long reads has been the use of transcriptomics (RNA-Seq) to find SVs, specifically rearrangements. In fact, it is conceivable to concentrate on SVs with potential functional implications by locating apparent RNA fusions, which are then intrinsically transcribed. Last but not least, recent advancements in benchmarking have substantially improved our comprehension of the advantages and disadvantages of each strategy. Such de novo-assembled sequences can be aligned to a reference assembly or another assembly and the differences between the two can be systematically identified to detect SVs. The identification of all types of variations should be possible by comparing each position in one genome to its corresponding position in the other genome. Different patterns result from discontinuities caused by specific kinds of SVs during a whole-genome alignment. Although conceptually straightforward, genome alignment is far from being a simple computational task [1-5].

Conclusion

The identification of SVs based on a genomic alignment has received numerous proposals. Whether they create an assembly graph or work directly on the assembled sequences, these can be separated from one another. Although they are often slower, methods that build the assembly graph can offer more insights since they use the read data directly. One of these techniques, called Cortex, can construct many genomes at once using short-read sequencing data. Merging reads or sequences calls for a perfect match, which raises the assembly quality. SGVar has been demonstrated to perform better for insertion and deletion identification than other techniques, such as Cortex, using both simulated and actual data (chromosome six of the human genome).

*Address for Correspondence: Mark Louis, Department of Molecular Cell Biology, Katholieke Universiteit Leuven, Belgium, E-mail: lkmark@nic.gen.in

References

1. 1000 Genomes Project Consortium. "A global reference for human genetic variation." *Nature* 526 (2015).

2. Voight, Benjamin F., Hyun Min Kang, Jun Ding and Cameron D. Palmer, et al. "The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular and anthropometric traits." (2012).

3. Trynka, Gosia, Karen A. Hunt, Nicholas A. Bockett and Jihane Romanos, et al. "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease." *Nat Genet* 43 (2011): 1193-1201.

4. Howie, Bryan, Christian Fuchsberger, Matthew Stephens and Jonathan Marchini et al. "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." *Nat Genet* 44 (2012): 955-959.

5. Simons, Yuval B., Michael C. Turchin, Jonathan K. Pritchard and Guy Sella. "The deleterious mutation load is insensitive to recent population history." *Nat Genet* 46 (2014): 220-224.

How to cite this article: Louis, Mark. "Structural Variations in Genetic Sequences." J Mol Genet Med 16 (2022): 557.