Interpreting genomes and transcriptomes of patients with rare genetic diseases

Sergey Naumenko, Hernan Gonorazky, Arun Ramani, James Dowling, Michael Brudno

The Hospital for Sick Children, 686 Bay street, Toronto, ON, Canada, M5G 0A4

sergey.naumenko@yahoo.com,

The number of inherited monogenic rare diseases is estimated as 7,000, and cumulatively they affect millions of people [1]. A proper diagnosis remains the main challenge in rare diseases with many patients waiting for years to be diagnosed. During the last years genome, exome, and transcriptome sequencing analyses were introduced into the rare disease research.

Many cases of rare diseases can not be solved with regular gene and variant discovery methods using the standard WES analysis of mutations. These cases expand to small research projects gathering expression, WGS/WES, and phenotype data for a patient or a family, and supported by disease models. From the bioinformatics' standpoint, these projects need more diverse tools, and more relaxed thresholds, compared to the conventional gene discovery, to capture more subtle effects.

For example, Gonorazky et al [2] identified a novel exon in *DMD* gene of a patient with Duchenne muscular dystrophy. An intronic mutation created a novel splice site, which, in a pair with existing cryptic splice site, resulted in two smaller splicing events instead of a larger intron. The novel exon, a short insertion in the middle of the mature transcript, produced a stop codon, and the transcript was degraded by the nonsense mediate decay pathway. As a result, the dystrophin expression in skeletal muscle was reduced.

In this case, WES analysis could not help to identify the mutation, because it laid outside of the annotated exonic regions, which are covered by WES. RNA-seq or WGS analysis would be helpful, however, these analyses are non-standard in a clinical setting: a search for novel exons (splicing analysis), and an analysis of intronic mutations.

In the other example, for a recessive muscular disease, the current hypothesis is that *GMPPB* gene has two mutations: one allele carries a rare and harmful missense mutation, and the other allele carries a rare 5'UTR mutation which produces an alternative start codon for

translation. In the result, the two mutations constitute a compound heterozygote. Again, cases like this require non-standard analysis of 5'UTR mutations, which are usually ignored in conventional genetic testing.

Here we briefly observe the bioinformatics tools we use to investigate such a cases.

The typical workflow includes data quality control, small variants (mutation) calling using WES, RNA-seq, or WGS for individuals, families, or cohorts, variant prioritization, variant annotation, variant reporting, variant analysis using HGMD [3], gene expression analysis with comparison to protein atlas and GTEx controls [4], splicing analysis: exon and isoform usage.

We start with one of the pipelines of bcbio system [5] to quickly go through the routine analyses. Bcbio system includes germline variant calling, RNA-seq, and smallRNA-seq pipelines. It also provides analyses for somatic mutations calling (cancer tumor-normal), structural variant calling, single-cell RNA-seq, and ChIP-seq.

The bcbio system is an open source python framework, which includes wrappers for the various tools, and workflows. The development of bcbio is led by the researchers of the bioinformatics core of Harvard T.H. Chan School of Public Health. Bcbio uses bioconda package installation system [6], which is able to install the fresh versions of more than 1000 bioinformatics packages, and cloudbiolinux repository of biological data [7], which contains all the necessary references and databases. It is possible to run bcbio workflows transparently on a high-performance computer cluster, or in a cloud. Bcbio has a very active development and bug resolving process [5], well written documentation [8], and validation analytics [9]. Bcbio is extensively used in many laboratories around the world, which actively report bugs, request new features, and participate in development.

GEMINI framework [10,11] aggregates many existing annotations and provides more than 20 tools for variant analysis (like *de-novo* mutations detection and pathway analysis).

In germline variant calling analysis of bcbio it is possible to choose from several alignment programs with bwa [12] by default. Mutations could be called in a sample-wise, a family-wise, or a cohort-wise manner. The ensemble method of variant calling uses 4 variant calling tools: GATK [13,14], freebayes [15], samtools [16], and platypus [17], requiring at

least 2 algorithms to vote for a variant to be called. In bcbio GATK-haplotype algorithm follows GATK best practices with slightly relaxed filters to improve sensitivity. Other algorithms are followed by their respective filters, which are extensively tested [9]. Validation tests have shown that precision and sensitivity can be tuned simply by changing the threshold in the voting scheme from 1 algorithm to 7 available, and reach the level of precision/sensitivity close to that of clinical pipelines, out of box.

Variants called are then annotated with ensembl Variant Effect Predictor [18] and are loaded into GEMINI database, which adds a number of additional annotations. From the annotated database the small set of rare variants with low population frequency and high impact (missense mutations, splicing changes) is extracted, and annotated with additional fields like OMIM [19], Orphanet [20] gene descriptions, imprinting status, Exac gene scores [21]. The result is an excel report containing 45 annotation columns suitable for clinicians.

RNA-seq seq analysis in bcbio runs STAR alignment [22] and then quantifies gene expression counts. If WES data is absent in a project, RNA-seq data becomes an important source of knowledge about mutations. However, the precision of variant calling in RNA-seq is lower, compared to WES or WGS, and only expressed genes are covered.

The combined analysis of RNA-seq and WES/WGS data with automated bioinformatics pipelines allows to facilitate identification of genetic causes of rare diseases.

References:

1. C.L. Beaulieu et al. (2014) FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project, *Am J Hum Genet*, **94(6)**:809-17.

2. H. Gonorazky et al. (2016) RNAseq analysis for the diagnosis of muscular dystrophy, *Annals of Clinical and Translational Neurology,* **3(1)**:55-60.

3. http://www.hgmd.cf.ac.uk/ac/index.php

4. http://www.proteinatlas.org/

5. https://github.com/chapmanb/bcbio-nextgen

6. https://bioconda.github.io/

7. http://cloudbiolinux.org/

8. https://bcbio-nextgen.readthedocs.io/en/latest/index.html

9. http://bcb.io.

10. U. Paila et al. (2013) GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations, *PLoS Comput Biol*, **9(7)**: e1003153.

11. https://gemini.readthedocs.io/en/latest/

12. H. Li and R. Durbin. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform, *Bioinformatics*, 25:1754-60.

13. G.A. Van der Auwera et al (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, *Current Protocols in Bioinformatics*, **43**:11.10.1-11.10.33

14. https://software.broadinstitute.org/gatk/best-practices/

15. E. Garrison, G. Marth. (2012) Haplotype-based variant detection from short-read sequencing, *arXiv preprint arXiv:1207.3907 [q-bio.GN]*

16. H.A. Li. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics,* **27(21)**:2987-93.

17. A. Rimmer et al. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications, *Nature Genetics*, **46**:912-918.

18. http://useast.ensembl.org/info/docs/tools/vep/index.html

19. https://www.omim.org/

20. http://www.orpha.net/consor/cgi-bin/index.php

21. M. Lek et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans, *Nature*, **536(7616):**285-91.

22. A. Dobin et al. (2013) STAR: Ultrafast universal RNA-seq aligner, *Bioinformatics*, **29(1):** 15–21.