A software pipeline for the discovery of variations in exome sequencing projects

E. Mattei, P.F. Gherardini, G. Ausiello, M. Helmer-Citterich✉
Department of Biology, University of Tor Vergata, Roma, Italy

Motivations

The recent advances in the technologies and strategies for DNA sequencing have dramatically facilitated the identification of novel human genes associated with rare and common diseases [1]. However novel methods are needed to identify high-quality variations among all the ones identified in a single experiment. The most successful approach to identify disease-causing mutations consists in using exome arrays [2] that allow the sequencing of only the coding regions in a genome. We developed a novel pipeline to identify high quality variations in the data produced by an exome sequencing experiment using the new 454 Roche sequencer [3].

Methods

The input data of our pipeline are the sequencer reads mapped on the reference genome and the variations already identified by the sequencer software along with their confidence score. The first step of the procedure consists in associating the confidence score to each variant nucleotide and then filtering out variations with a low score. Variations in the length of the reads lead to misalignments between the reads and the reference genome. The second step of the pipeline produces a more accurate local alignment that can be searched for variations. Since the aim of the procedure is to validate the original variations produced by the sequencer, and not to identify new ones, original variations are compared with the newly identified ones for confirmation. As expected the sequencer alignment error rate increases as the length of the reads decreases, causing nucleotide mismatching. Reads with a perfect alignment with the reference, are marked as FULL by the sequencer. However the sequencer software also reports variations identified on chimeric reads, i.e. reads for which different portions align to separate regions of the genome. The pipeline reports, for each variation, how many FULL reads support the variation and how many do not. Information about CHIMERIC reads are included as well and variations supported only by CHIMERIC reads are more likely to be incorrect. The next step is to flag single nucleotide polymorphisms as missense or synonymous and to use dbSNP [4] to discard the ones which are already known, and therefore unlikely to be associated with a disease. Subsequently, the propensity of each missense mutation to be deleterious for the function of a protein as opposed to neutral is calculated using CONDEL [5], a software that computes a weighted average of the scores of the SIFT [6] and POLYPHEN [7] methods. Moreover variations in dbSNP which are known to be associated with a disease are flagged. As a last step we prioritize mutations occurring in genes belonging to the same family of other genes known to be implicated in the pathology, if any. When SNP array [8] data of a patient genome are also available the procedure includes an additional test to identify which variations are more likely to be in a heterozygous site.

Results

We tested our pipeline on the sequenced exomes of two patients suffering from Noonan Syndrome [9] and having no mutations in any of the genes already known to be implicated in the disease. SNP array data was also available for one of the patients. The original number of variations identified by the Roche software was about 105,000. After filtering the original set using the Phred score, the number decreased to about 102,000. Using a statistical test based on comparing the sequencing and SNP array results we reduced the number to 22,000. The removal of known SNPs further reduced the number of newly identified variations to 15,000, only 1,400 of which were missense. CONDEL predicted 800 of them as deleterious and only 60 were found in genes likely implicated in the disease. Our filtering pipeline therefore reduced the initial number of variations by four orders of magnitude, resulting in a very limited number of variations that can be tested in follow-up experiments.

References

1. Roukos, D.H. (2010). Next-generation sequencing and epigenome technologies: potential medical applications. Expert Rev. Med. Devices 7, 723-726.
2. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. (2007). Genome-wide in situ exon capture for selective resequencing. Nat. Genet. 39, 1522-1527.
3. Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135-1145.
4. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308-311.
5. Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. 88, 440-449.
6. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affec protein function. Nucleic Acids Res. 31, 3812-3814.
7. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30, 3894-3900.
8. Mei, R., Galipeau, P.C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R.K. Chee, M.S., Reid, B.J., and Lockhart, D.J. (2000). Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Res. 10, 1126-1137.
9. Allanson, J.E., and Roberts, A.E. (1993). Noonan Syndrome. In GeneReviews, R.A. Pagon, T.D. Bird, C.R. Dolan, and K. Stephens, eds. (Seattle, WA).