A Novel Statistical Algorithm to Detection of Large-scale Deletions in PCR-enriched Target Sequencing Data

German Demidov

St Petersburg Academic University; Parseq Lab, St Petersburg, Russia, gdemidov@parseq.pro

Julia Vnuchkova

Parseq Lab, St Petersburg, Russia, jvnuchkova@parseq.pro

Anton Bragin

Parseq Lab, St Petersburg, Russia, abragin@parseq.pro

A multiplex PCR is a widespread technique for target enrichment in massively parallel sequencing protocols. One of the typical use cases is an accurate detection of SNVs and short rearrangements which made it routinely used for clinical diagnostics of genetic disorders. The obvious disadvantage of the method is the absence of a clear way to detect large (with size exceeding length of the amplicon) copy-number variations, however, this modifications can be the causes of genetic disorders too. Some methods for detection of large-scale deletions were proposed, but existing tools have a lot of restrictions and their accuracies are usually low especially for the comparatively small deletions (like the deletions that happened in one exon). The disadvantages of such approaches leaded us to develop our own. We have developed a method and implemented a tool that is easy to use and the assumptions used by this tool are suitable for the vast majority of the real datasets.

We imply the idea of clustering amplicons that have similar biochemical properties and consequently show some sort of relationship with each other and then preparing comparisons of coverages inside the clusters of similar amplicons. We assume that the changes in efficiencies depends on the structures of amplicons and primers so this changes are not random. Consequently if we have large population of amplicons some of them will react to such changes of the reaction's conditions in the similar way, which means that we can subdivide the problem on two different tasks: at first, find groups of amplicons that demonstrate similar behaviour, and then analyze amplicons inside this groups. Then the detection of deletions can be reformulated as detection of outliers using methods of robust statistical learning.

We have modified existing mathematical model (branching processes) for multiplex PCR and made conclusions based on this model. We have developed an algorithm for the detection of deletions, implemented a tool and tested it on a limited dataset of 4 runs (192 samples) of sequencing of three genes and found that performance of our approach is better than the performance of any existing tools.

1. Buysse,K., et al.. (2009) Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience, *European Journal of Medical Genetics*, **52**:398-403.
2. Lalam, N., et al.. (2004) Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency, *Advanced Applied Probability*, **36**:602-615.
3. Min Zhao, et al. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives, *BMC Bioinformatics*, **14**:1-16.
4. Takahiro, K. (2003) Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR), *Journal of Bioscience and Bioengineering*, **96, No. 4**, 317 - 323.
5. Sen, P.K. (1968) Estimates of the regression coefficient based on Kendall's tau, *Journal of the American Statistical Association*, **63**:1379-1389.
6. Pasman V. R., Shevlyakov G. L. (1987), Robust methods of estimating the correlation coefficient, *Avtomat. i Telemekh*., **Issue 3**:70–80.