Bioinformatic Approaches to Detecting Copy Number Variation in Next Generation Sequencing Data for Clinical Diagnostics
Straver, R.

2018

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Straver, R. (2018). Bioinformatic Approaches to Detecting Copy Number Variation in Next Generation Sequencing Data for Clinical Diagnostics. [, Vrije Universiteit Amsterdam].

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address: vuresearchportal.ub@vu.nl

Download date: 21. Apr. 2024
Some genetic diseases are caused by changes in DNA that result in a different number of copies of a particular sequence of DNA, so called copy number variations (CNVs). Consequently, knowledge about CNVs in a patient enables diagnosing and providing the right type of healthcare for such diseases. While the effects of a CNV on a patient’s health vary strongly in severity, the effects of lengthy CNVs up to whole chromosomal copy number changes mostly have a critical impact such as dire intellectual disability and grave malformations, or even result in miscarriage because of early fetal demise. Early prenatal detection of such genetic disorders would greatly improve the ability to prepare for the right type of care, or, alternatively, enable prospective parents to make an informed decision on whether or not to carry a child to term. The introduction of next generation sequencing (NGS) provided us the ability to determine genetic variations from cell-free DNA (cfDNA), a mixture of fragmented DNA from cells across the whole body (including the placenta if pregnant) that can be extracted easily from a sample of blood. This thesis describes how we used NGS data for early detection of CNVs that cause genetic disorders.

For early detection, we have analyzed cfDNA in maternal blood. When sampling blood from a pregnant woman, the extracted cfDNA is partly of fetal origin. While this is a relatively small amount (5% to 20% at 10 to 14 weeks of gestation), it can be exploited to detect CNVs in the fetus, thereby opening doors for non-invasive prenatal testing (NIPT). For example, a fetus with trisomy 21 (causing Down syndrome) is reflected by a slight increase in the number of reads found from chromosome 21. Extending this approach to other chromosomes for non-invasive prenatal testing proved significantly harder. The distribution of DNA fragments obtained in NGS is not uniform across the genome as technical variations cause strongly deviating read counts over different regions. While chromosome 21 is relatively insensitive to these technical variations, other chromosomes were found to be more prone to technical and biological effects on the distribution of reads across the reference genome.

In this thesis, we introduce a within sample comparison method for the analysis of NIPT-data, which we implemented in WISECONDOR. This method inspects a large set of healthy reference samples to infer genomic regions that show similar read depth behavior over different samples. When testing a sample, WISECONDOR compares the read count of a region to several other (similarly behaving) regions within the same sample. In doing so it provides a significance score describing how likely a region is to be part of a CNV, while correcting for systematic variations in sequencing depth. This method proved reliable enough to not only enable detection of trisomies other than trisomy 21, it also showed promising results for sub-chromosomal fetal CNVs, down to approximately 10-15 megabase in size.

As extensive testing showed the method is very reliable, WISECONDOR is now used in several medical centers worldwide for non-invasive prenatal testing. In the Netherlands specifically, a pilot study was performed over several years that allowed women with in-

170
creased risk of having a baby with significant CNVs to test non-invasively. While multiple approaches were tested by different centers during this study, WISECONDOR showed the most reliable results and is used for all NIPT in the Netherlands since April 1st 2017.

An issue that was left open for diagnostic applications of NIPT was whether or not a sample had enough fetal DNA to determine any CNVs it may have. For male fetuses, this problem could be solved using the number (and distribution) of reads mapped to chromosome Y, as we also implemented in an approach called DEFRAG. However, this leaves half of the pregnancies, those with female fetuses, without any indication of the fetal fraction. While other methods aiming to solve this issue required additional testing or paired end sequencing, we developed an approach to estimate the fetal fraction of a sample directly from the data already obtained for NIPT. Our solution is based on a slight difference in the start positions of DNA fragments between maternal and fetal DNA, caused by the DNA being wrapped around histone complexes in so called nucleosomes. This method was implemented in SANEFALCON, a separate tool that we suggest using for quality control in diagnostics alongside any NIPT solution one chooses to implement.

Considering how successful the approach taken in WISECONDOR was in dealing with erratic read distributions (even changes in read depth of about 2.5% could be determined), we altered the within sample comparison to work on (non-fetal) whole exome sequencing (WES) data. For this type of data, the read distribution is heavily skewed toward genomic locations that are genetically coding. These targeted regions vary strongly in size and read depth behavior, rendering straightforward CNV detection methods unreliable. Instead of looking for lengthy CNVs with subtle changes in read effect size (as in NIPT), we aimed to detect strong CNVs across short regions. We created an implementation called WISExome that employs within sample comparison and we showed that it can correctly identify sub genic (and even sub exonic) CNVs in WES data, reducing the need for separate array analysis. This is especially useful when a CNV is suspected in a patient where SNP information obtained through WES could not provide a final conclusion.

With the work in this thesis we show that the importance of bioinformatics in diagnostics is imperative. Although nowadays lots of molecular data can be obtained, the data contains many experimental and technical variations. Drawing reliable conclusions from this data without carefully designed statistical models is impossible. We introduced WISECONDOR, a novel approach to normalizing data which fully relies on the data from the sample being measured. As a result, subtle deviating values in the data can be interpreted, improving the diagnostic value of the data significantly. Together, our work amplifies the importance of revolutions within molecular biology on being able to measure cellular material at an unprecedented scale. However, it also shows that true clinical impact can only be realized when a similar revolution in computational tools is realized to support the subsequent analysis and interpretation of the gigantic volumes of data being generated.