A TOOL TO IDENTIFY SHARED IMBALANCES IN ARRAY HYB PARTNERS

ALED JONES

A dissertation submitted to the University of Manchester for the degree of Master of Science in the Faculty of Medical and Human Sciences;

2016

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CGH Comparative Genomic Hybridisation

CNV Copy Number Variation

CBS Circular Binary Segmentation

HMM Hidden Markov Model

something.

Dedicated to

1

ABSTRACT

Short summary of the contents...

DECLARATION

No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning;

Aled Jones

INTELLECTUAL PROPERTY

i The author of this dissertation (including any appendices and/or schedules to this dissertation) owns certain copyright or related rights in it (the "Copyright" and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made

iii The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the dissertation, for example graphs and tables ("Reproductions"), which may be described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iiii Further information on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/display.aspx?DocID=487), in any relevant Dissertation restriction declarations deposited in the Uni-

versity Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's Guidance for the Presentation of Dissertations.

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

ACKNOWLEDGMENTS

Put your acknowledgments here.

Many thanks to everybody who already sent me a postcard!

Regarding the typography and other help, many thanks go to Marco Kuhlmann, Philipp Lehman, Lothar Schlesier, Jim Young, Lorenzo Pantieri and Enrico Gregorio[1], Jörg Sommer, Joachim Köstler, Daniel Gottschlag, Denis Aydin, Paride Legovini, Steffen Prochnow, Nicolas Repp, Hinrich Harms, Roland Winkler, Jörg Weber, and the whole LaTeX-community for support, ideas and some great software.

Regarding LyX: The LyX port was intially done by *Nicholas Marietti* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and the contributions to the original style.

1 Members of GuIT (Gruppo Italiano Utilizzatori di TeX e LaTeX)

2

INTRODUCTION

An introduction to the project

- Diagnostic CNV detection in NHS
  - what CNVS are
  - Karyotyping
  - PCR/MLPA
  - arrayCGH
- What microarrays are
  - technology
- how microarrays are used at Guys
  - patient to patient
  - run sheets
  - studies to justify
- why this tool is needed
- what has been done before
- functional requirements

Array Comparative Genomic Hybridisation (CGH) is a commonly used diagnostic test in clinical genetics. The test utilises many probes (up to 60,000) attached to a glass slide. Each probe consists of many copies of an oligonucleotide 40-60 base pairs long, designed to target a specific region of the genome. Samples are hybridised to the array in a competitive reaction, usually a diagnostic sample in competition with a reference sample.

The diagnostic sample and reference samples are labelled with a fluorescent dye (patient Cy5 and reference Cy3). Once hybridised the array is scanned producing a high resolution image of the array (Ahn et al., 2010). This image undergoes feature extraction to calculate the signal intensities at each probe and associated quality scores.

This data can then be analysed for CNV by applying one of a number of available algorithms: The probe scores are normalised, split into segments of equal copy number and each segment assigned a copy number status (Hupe, 2004).

During the data normalisation the signal intensities are calculated and given quality scores. Dividing the signal intensity of Cy5 by that of Cy3 produces a ratio where equal copy number is 1. This ratio is then logged (base 2) to produce the log2 ratio where equal copy number is 0, loss of material is shown with a negative number and gain of material positive.

Copy number variation represents a deviance in copy number from a reference genome which typically contains 2 copies of a DNA segment (Roy and Motsinger Reif, 2013). CNVs can occur in recombination and replication events. CNVs can be benign polymorphisms or associated with Mendelian, sporadic and complex disease possibly through gene dosage, disruption, fusion or positional effects (Zhang et al., 2009).

In 2009 array CGH started replacing karyotyping as the method for detection of CNVs in NHS diagnostic genetic services. Array CGH offers a higher resolution, less reliance on analyst interpretation/skill and the advantage of a high throughput practical workflow, however few trusts actually have this test commissioned due to the higher cost. Guy's and St Thomas' NHS Trust adopted a patient to patient hybridisation approach (using 8x60K Agilent array design) which halves the cost of consumables by replacing the reference sample with another pa-

tient sample (Ahn et al., 2010).

As array CGH compares two samples the use of a normal reference sample infers any CNV detected is from the patient. Hybridising two patients removes this assumption, producing two challenges:

1. Is a CNV is a duplication in one patient, or a deletion in the second patient?

2. If both patients have the same CNV relatively no difference would be found, not detecting the CNV.

The first challenge can be overcome by comparing the signal intensities of each dye across the imbalance and across normal regions. One patient will have different signal intensity in this region and one will have no difference.

The second challenge requires "a careful consideration of patient referral information" to reduce the risk of this occurring. Hybridisation partners are mismatched on phenotype (Ahn et al., 2010). This assumes that patients with differing referral reasons eg heart vs. renal defects will not have the same underlying CNV.

The product of the increased resolution of arrays is a higher abnormality pick up rate than karyotyping.

The aim of this project is to create a tool which is run during data processing/analysis and uses signal intensity (as opposed to log signal ratio) to detect CNV independently of the hybridisation partner.
The Cy3 and Cy5 signal intensities can be compared to a reference set of signal intensities created from previous arrays. Results from each hybridisation partner can be compared to identify any shared CNV which may not be identified using the signal ratio.

2.1 WHAT HAS BEEN DONE PREVIOUSLY IN THIS AREA?

2.1.1 *Array design*

The arrays, reagents, equipment and software used to process and analyse are manufactured by Agilent Technologies (**?**). The probe design is stored on the online probe catalogue/custom array design tool eArray (**?**).

A request/suggestion for a tool similar to that which this project aims to create was sent and declined by Agilent.

2.1.2 *Raw array data*

The feature extraction produces a tab delimited text file of 10-15mb in size. This file is split into three sections: metadata about the array run, background measurements and then a number of measurements for each probe.

Each probe has a name, genomic location and physical position on the array and 35 measurements including the signal intensity, probe saturation and background readings for each dye.

2.1.3 *Use of algorithms in CNV calling*

If all goals of this project are met the tool produced will have much the same role as the aberration detection algorithms, taking feature extracted data and looking for CNV.

There are a number of algorithms which are in use for CNV calling from CGH data. These algorithms are recursive binary segmentation methods which break down chromosomes into segments of equal copy number. Examples include Z-scores, Circular Binary Segmentation (CBS), aberration detection method (ADM 2), nexus and Hidden Markov Model (HMM).

2.1.3.1 *Z-score algorithm*

Z-scores are a statistical measure of deviation from the mean for a normally distributed population. The Z-score algorithm starts off by giving a Z-score to signal log ratio for each probe. Any probes above a user defined Z-cut off can be classified as an outlier, or significantly away from mean.

The number of probes classified as above (R) and below this threshold (R') and total number of measurements (N) are recorded and used when calculating the moving average of small windows of the genome.

These 'windows' can be specified as a number of adjacent measurements or a fixed size eg every 1MB. Within each window the abundance of probes which log ratios which deviate from the mean is measured (r:r'). A Z-score is then calculated measuring the significance of the over-abundance of probes with a deviant score in this window (Agilent Technologies, 2012).

2.1.3.2 *Aberration Detection Method (ADM-1/ADM-2)*

These algorithms are designed by Agilent. These algorithms do not used fixed windows but segments the genome into intervals of equal copy number using signal log ratio scores from adjacent probes to best define interval breakpoints.

ADM-1

Firstly the data is normalised by subtracting the mean log ratio and dividing by the variance, creating a normally distributed population with a mean of 0. Each chromosome is then broken into intervals of equal copy number and intervals are assigned a score (S(I)) which denotes the difference from the mean. A user defined threshold is set and any intervals which are above this are called as an aberration. The interval with the highest score is selected and the same process is performed on this segment to further define breakpoints. This is repeated for all intervals with an S(I) above the threshold.

The ADM-2 algorithm builds on ADM-1 by including probe quality information to weight probe signal log ratios when assigning S(I).

2.1.3.3 *Circular binary segmentation (CBS)*

Circular binary segmentation (CBS) also uses adjacent probe log ratio scores to create intervals which, as opposed to ADM-1/2 (which classifies segments as aberrant or not), are grouped into intervals with equal copy number.

Each chromosome is made into a circle (Figure 1). This allows for two break points to be identified which increases the resolution of detection. When two breakpoints are detected the interval of potential different copy number is removed, and the flanking regions are joined to form a new circle. This allows a t-test to be performed between the mean log ratios on each interval.

Definition of an interval is determined using permutation testing, creating intervals using various breakpoints and looking for the most significant P value. If this P value is above a threshold an interval is created and the process is repeated for all intervals until no more changes are found. A number of checks are performed to ensure the correct end points have been found including edge effect correction, change point pruning and estimation of the log score distribution.

A

APPENDIX

BIBLIOGRAPHY

Joo Wook Ahn, Kathy Mann, Sally Walsh, Marwa Shehab, Sarah Hoang, Zoe Docherty, Shehla Mohammed, and Caroline Mackie Ogilvie. Validation and implementation of array comparative genomic hybridisation as a first line test in place of postnatal karyotyping for genome imbalance. *Molecular Cytogenetics*, 3:9, April 2010. ISSN 1755-8166. doi: 10.1186/1755-8166-3-9. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2885406/. PMID: 20398301 PMCID: PMC2885406.

Phillipe Hupe. GLAD-Introduction. http://bioinfo-out.curie.fr/projects/glad/node2.html, November 2004. URL http://bioinfo-out.curie.fr/projects/glad/node2.html.

Siddharth Roy and Alison Motsinger Reif. Evaluation of calling algorithms for array-CGH. *Frontiers in Genetics*, 4, October 2013. ISSN 1664-8021. doi: 10.3389/fgene.2013.00217. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3829466/. PMID: 24298279 PMCID: PMC3829466.

Feng Zhang, Wenli Gu, Matthew E. Hurles, and James R. Lupski. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, 10(1):451–481, 2009. doi: 10.1146/annurev.genom.9.081307.164217. URL http://dx.doi.org/10.1146/annurev.genom.9.081307.164217. PMID: 19715442.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`http://code.google.com/p/classicthesis/`

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

`http://postcards.miede.de/`