BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 8 2010, pages 1007–1014
doi:10.1093/bioinformatics/btq088

Genome analysis
Advanced Access publication February 25, 2010

Exploiting sequence similarity to validate the sensitivity of SNP arrays in detecting fine-scaled copy number variations

Gerard Wong[1,2,*], Christopher Leckie[1,2], Kylie L. Gorringe[4,5], Izhak Haviv[6,7], Ian G. Campbell[4,5] and Adam Kowalczyk[1,3]

[1]Victoria Research Laboratory, National ICT Australia, [2]Department of Computer Science and Software Engineering, University of Melbourne, [3]Department of Electrical and Electronic Engineering, University of Melbourne, [4]VBCRC Cancer Genetics Laboratory, Research Division, Peter MacCallum Cancer Centre, [5]Department of Pathology, University of Melbourne, [6]The Alfred Medical Research and Education Precinct, Baker Medical Research Institute, Epigenetics Group and [7]Department of Biochemistry and Molecular Biology, University of Melbourne, Australia

Associate Editor: John Quackenbush

ABSTRACT

Motivation: High-density single nucleotide polymorphism (SNP) genotyping arrays are efficient and cost effective platforms for the detection of copy number variation (CNV). To ensure accuracy in probe synthesis and to minimize production costs, short oligonucleotide probe sequences are used. The use of short probe sequences limits the specificity of binding targets in the human genome. The specificity of these short probeset sequences has yet to be fully analysed against a normal reference human genome. Sequence similarity can artificially elevate or suppress copy number measurements, and hence reduce the reliability of affected probe readings. For the purpose of detecting narrow CNVs reliably down to the width of a single probeset, sequence similarity is an important issue that needs to be addressed.

Results: We surveyed the Affymetrix Human Mapping SNP arrays for probeset sequence similarity against the reference human genome. Utilizing sequence similarity results, we identified a collection of fine-scaled putative CNVs between gender from autosomal probesets whose sequence matches various loci on the sex chromosomes. To detect these variations, we utilized our statistical approach, Dectecting REcurrent Copy number change using rank-order Statistics (*DRECS*), and showed that its performance was superior and more stable than the *t*-test in detecting CNVs. Through the application of *DRECS* on the HapMap population datasets with multi-matching probesets filtered, we identified biologically relevant SNPs in aberrant regions across populations with known association to physical traits, such as height, covered by the span of a single probe. This provided empirical confirmation of the existence of naturally occurring narrow CNVs as well as the sensitivity of the Affymetrix SNP array technology in detecting them.

Availability: The MATLAB implementation of *DRECS* is available at http://ww2.cs.mu.oz.au/~gwong/DRECS/index.html

Contact: gwong@csse.unimelb.edu.au

Supplementary information: Supplementary information is available at *Bioinformatics* online.

*To whom correspondence should be addressed.

Received on December 5, 2009; revised on February 21, 2010; accepted on February 22, 2010

1 INTRODUCTION

The importance of copy number variation (CNV) in influencing phenotypic differences and disease susceptibility has been well-established in a number of studies (Beckmann *et al.*, 2007; Frazer *et al.*, 2009; McCarroll and Altshuler, 2007; Nakamura, 2009; The Wellcome Trust Case Control Consortium, 2007). In determining CNVs that are more likely to have a causal association with phenotypes and disease predisposition, it is important to focus on recurrent copy number (CN) changes present in multiple samples. Recurrent CN changes are those that are consistent in an assumed homogeneous population, as well as those that stratify heterogeneous sub-populations. An ongoing challenge is the reliable and efficient detection of these changes, both variations and aberrations, down to the resolution provided by the microarray technology.

High-throughput microarrays based on single nucleotide polymorphisms (SNPs) are an effective platform for analysing CNV. In recent years, the resolution of SNP arrays has grown exponentially while the cost of experiments continued to fall, making these a feasible high-resolution platform to conduct CNV studies involving a large number of samples.

SNP arrays such as those by Affymetrix typically limit the length of probe sequences to 25mers to ensure accuracy in probe synthesis as well as to keep manufacturing costs down. A probeset consists of a collection of probes at a short distance offset from each other covering typically a span of 33mers. The use of short probeset sequences results in reduced specificity towards the targeted genomic locus and this in turn degrades the signal quality of the intended target in proportion to the number of unintended loci matched. However, the uniqueness or similarity of each probeset sequence on existing SNP arrays with respect to the reference human genome is not a fully explored aspect of SNP array quality. This is the focus of our first contribution in this article. We determined the sequence similarity for each and every probeset on all Affymetrix SNP array variants by sequence matching the two allelic versions of each flanking probeset sequence against the reference human genome assembly (NCBI 36). The result provides an interesting

insight in terms of target uniqueness, and can be used directly as a quality filter to remove multi-matching probesets prior to performing CN analysis.

In assessing the performance of various methods for the detection of recurrent CNV, it is often useful to have a set of ground truth labels that serve as an approximate gold standard. Our second contribution in this article is the derivation of a set of approximate ground truth labels from probeset sequence similarity for the detection of putative CNV in normal gender samples. The advantage of using this set of labels is that they are based on naturally occurring CN imbalances in gender and are not based on artificially implanted 'peaks or spikes' against simulated Gaussian noise as a background. Additionally, the labels of putative CNV cover both narrow as well as longer contiguous regions of CN change, which provides some form of natural diversity in the data that may not be captured as closely in synthetically simulated data. Finally, since these labels are based on DNA sequence similarity, the putative CNVs across gender will be recurrent in the majority of samples and not just a small subset of samples. This essentially is a perfect fit for the nature of CN change that we aim to detect.

Our third contribution is Dectecting REcurrent Copy number change using rank-order Statistics (*DRECS*), a statistical approach for the detection of recurrent CN variation/aberration in multiple samples. It is based on a normalized rank transformation of raw $\log_2$ ratios (i.e. replacing raw $\log_2$ ratios with their cumulative probability density) to reduce the impact of outlying measurements. This effectively transforms the distribution of the raw $\log_2$ ratios to a discrete uniform distribution for consistent estimation of variance and computation of statistical significance.

We have evaluated the effectiveness of *DRECS* in terms of:

- its sensitivity, specificity, accuracy and area under the ROC curve (AUC) in detecting putative CN difference across gender as described previously, using the *t*-test as a benchmark;
- its consistency in performance on both raw $\log_2$ and segmented CN as inputs;
- its scalability to sample size and array resolution; and
- its ability to identify fine-scaled regions of CNV with known biological significance.

In our final contribution, we show the existence of recurrent CNV in the HapMap populations, e.g. in single SNPs with known association to physical traits, such as height and freckles. This empirically validates the effectiveness of *DRECS* in identifying recurrent CNV that stratify populations down to the resolution of the microarray, and also highlights the sensitivity of the SNP microarray in detecting regions of narrow CNV.

In Section 2, we describe our approach in identifying the extent of sequence similarity to the human genome across every variant of the Affymetrix SNP array. In Section 3, we use the knowledge of sequence similarity to derive ground truth labels of putative CN change to help us assess the performance of our statistical approach, *DRECS* (Section 4), in detecting recurrent CNV. In Section 5, we evaluate the performance of *DRECS* under various experimental conditions with the *t*-test as a benchmark and apply *DRECS* to detect known CN polymorphisms. We conclude the section with a short discussion on two biologically relevant CNVs identified.

2 ASSESSING AFFYMETRIX SNP ARRAY PROBESET SEQUENCE SIMILARITY TO THE HUMAN GENOME

We first examine the extent to which probeset sequence similarity occurs in practice on the widely used Affymetrix SNP arrays. The probes on Affymetrix SNP arrays have a flanking sequence of 33mers (except for the Human Mapping50K Xba240 and Human Mapping50K Hind240 arrays with a slightly longer flanking sequence of 51mers). Since flanking sequences are short, we can reasonably expect that some probes will exhibit an over-representation of their complement sequence in the reference human genome. To profile sequence similarity on all Affymetrix SNP arrays, we perfectly matched (with zero mismatches) all flanking probeset sequences (using both allelic variants) against the Human Genome Assembly NCBI Build 36. Sequence matching was performed using BLAST (Altschul *et al.*, 1990) with the zero mismatch option. For non-polymorphic sequences, the 25mer probe sequences were used. For simplicity, the term 'probeset' in this article is taken to refer to both a collection of SNP probes targeting a particular locus over a flanking distance of 33mers as well as each individual non-polymorphic (CN) probe of length 25mers found in the Genome-Wide Human SNP 5.0 and Genome-Wide Human SNP 6.0 arrays.

The results gave us a landscape of the uniqueness of the ultimate genomic targets of each and every probeset in every variant of the Affymetrix SNP array. They indicate that between 0.43% and 1.03% of SNP probesets across all array variants have non-unique matching loci, with the most similar SNP probeset matching up to 1500 individual loci. While 6.11–16.77% of CN probesets have non-unique matching loci with the most similar probeset matching more than a remarkable 76 000 loci. These results are summarized in Figure 1 and Table 1. Detailed sequence similarity results for every Affymetrix SNP array variant are available at http://ww2.cs.mu.oz.au/∼gwong/DRECS/index.html.

This observed probeset similarity has important consequences for the detection of fine-scaled CNVs. In particular, spurious hybridization signals from multiple matching loci affect the reliability of any inferences made on these affected probesets. One approach to this reliability problem is to ignore any probesets that are known to exhibit multiple matches. However, rather than ignoring these probesets, we have found a novel, practical use for them. We have used these probesets to derive a set of 'ground truth labels', which can be used to evaluate the accuracy of methods that identify narrow regions of CN change. In Section 3, we describe our method for generating ground truth labels on the basis of the genomic loci matched by the probesets. This method is then used in the later sections to evaluate different techniques for detecting narrow regions of CN change.

3 GENERATING GROUND TRUTH LABELS FROM SEQUENCE SIMILARITY TO THE NON-AUTOSOMES

In this section, we present a novel method for generating ground truth labels corresponding to individual probesets with a known difference in CN measurements between male and female samples. Our method exploits the observation that sex chromosome CN differences between normal females and males results in a natural

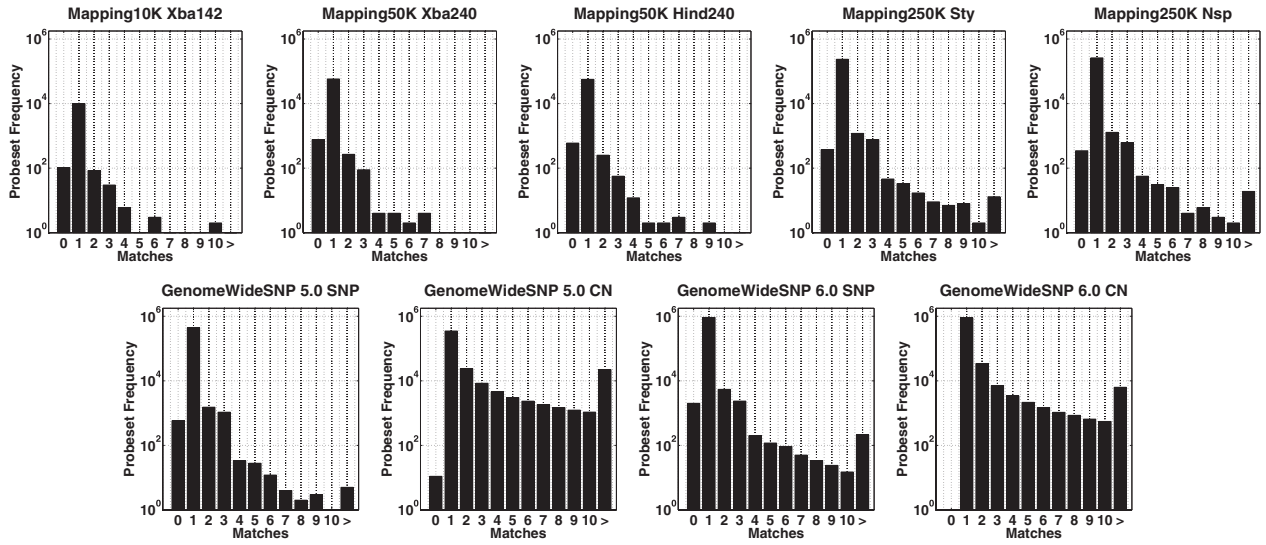

Fig. 1. Probeset sequence similarity profile of Affymetrix arrays. The frequency of matches of probeset sequences against the reference human genome assembly in March 2006 (NCBI 36) are shown. The plots are histograms of the number of loci matched by probeset sequences of each array variant. '0' indicates the probeset sequence was unmatched in the reference human genome (NCBI 36), '1' indicates a unique match and '>' indicates matches of >10 genomic loci. SNP flanking sequences were all 33mer except for the Mapping50K Xba240 and Mapping50K Hind240 variants that were 51mer. The probe sequences for the non-polymorphic CN probes were 25mer. The experiments exclude the quality control probes (AFFX) on the arrays.

Table 1. Affymetrix SNP array probeset sequence matching summary

Affymetrix array variant	Annotated probesets	Unique matching (%)	Non-matching (%)	Multiple matching (%)	Most sites matched by a probeset
Human Mapping10K Xba142	10 095	98.15	0.82	1.03	41
Human Mapping50K Xba240	58 625	98.87	1.27	0.47	8
Human Mapping50K Hind240	56 936	98.54	1.03	0.43	7
Human Mapping250K Sty	237 701	99.21	0.15	0.64	436
Human Mapping250K Nsp	261 563	99.35	0.13	0.52	311
Genome-Wide Human SNP 5.0 SNP	440 094	99.41	0.13	0.61	223
Genome-Wide Human SNP 5.0 CN	419 270	83.23	0.0007	16.77	76 060
Genome-Wide Human SNP 6.0 SNP	929 967	99.07	0.21	0.92	1500
Genome-Wide Human SNP 6.0 CN	945 806	93.89	0	6.11	15 146

The percentage of Affymetrix SNP array probesets that have unique matching loci against NCBI36 ranged from 83.23% to 99.41% across the variants. A small percentage of probeset sequences were unmatched against the reference human genome (NCBI 36), while the percentage with multiple matching genomic loci was variable between 0.43% and 16.77%. The less specific probesets were typically the non-polymorphic probesets with 25mer sequences. The percentage similarity of these probesets were significantly higher than the polymorphic SNP probesets. One particular CN probe sequence on the Genome-Wide Human SNP 5.0 array matched more than 76 000 loci, while one probeset sequence on the Genome-Wide Human SNP 6.0 array matched 1500 loci. The summary reported here includes only probesets annotated by Affymetrix (version na29) and exclude all 'AFFX' probesets across the arrays.

CN difference in autosomal probesets that display some form of sequence similarity to the X and Y chromosomes. Effectively, known gender imbalances in the sex chromosomes plus sequence similarity in probesets allows us to derive an approximate set of ground truth labels for putative CN changes in gender at various probeset loci. These ground truth labels provide a basis for a standardized comparison of the performance of methods for detecting CN change in SNP array datasets. To the best of our knowledge, this is the first method of its kind published for this purpose.

To illustrate our method, consider Figure 2, which shows three example probesets (A, B and C) that display putative no change, gain and loss labels based on $\log_2$ ratios of female to male copies as identified by matches from our sequence similarity experiments. Most target probesets on the autosomes that exhibit sequence similarity to other loci on the autosomes will show no change in their putative CN difference, as shown by Example A. However, consider the case in Example B of a target probeset on the autosome that exhibits sequence similarity to a locus on chromosome X. In this case, the measurements from the female sample will indicate two extra matches corresponding to the two X chromosomes, while the male reference sample will exhibit only one extra match. This can be viewed as a ground truth label for a putative CN gain. Similarly, Example C provides an example of a ground truth label for a putative CN loss, due to a target probeset on an autosome that exhibits sequence similarity to a locus on the Y chromosome. In this case, the female sample experiences no increase in CN measurement,

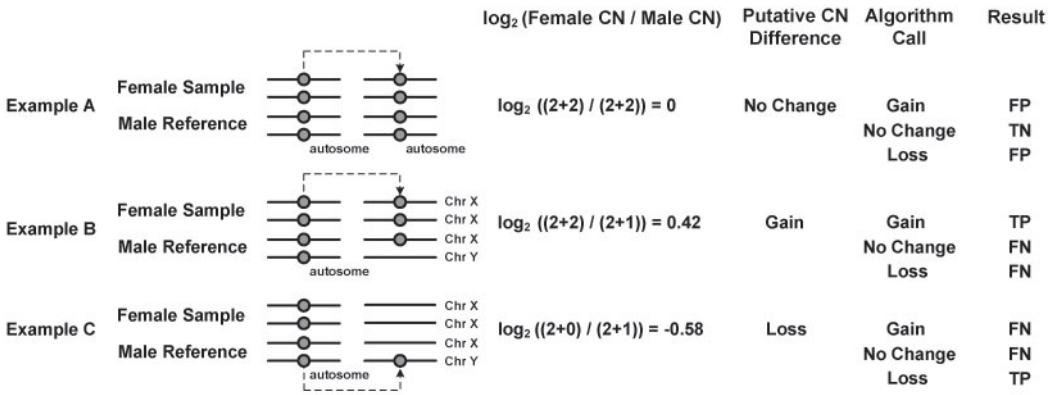

Fig. 2. Determination of true positive, false positive, true negative and false negative. The probesets show the three possible states for putative CN change (gain, loss and no change) as determined by their matches on the autosomes and non-autosomes. *DRECS* makes a call on each of the probesets in the entire array. The result is determined by the agreement of the ground truth label with the call computed by *DRECS*. For example, if *DRECS* predicts a gain and this matches a label of putative gain then we have a true positive. A gain for a probeset, as determined by *DRECS*, is a statistically significant positive deviation of the mean of normalized-transformed ranks from the mean of the null distribution. TP, FP, TN and FN are used subsequently to determine sensitivity, specificity, accuracy and AUC (Table 3).

Table 2. Ground truth label summary for normal female-male comparison based on intended target loci for probesets on the Human Mapping 500K array

Total probesets	No change (Chr. 1–22)	No change (Chr. X)	Gain (Chr. 1–22)	Gain (Chr. X)	Loss (Chr. 1–22)	Loss (Chr. X)	Unmatched (Chr. 1–22)	Unmatched (Chr. X)
500 568	489 186	11	84	10 522	47	0	715	3

This summary is based on the original annotation provided by Affymetrix for the Human Mapping 500K array. For our analysis, we used labels for all probesets excluding the unmatched ones against the reference human genome build NCBI 36. While these probeset sequences did not match the human genome build NCBI 36, some of them did match loci in the GRCh37 release of the human genome assembly by the Genome Reference Consortium. However, since these probesets are few in comparison to the rest of the labelled probesets, their inclusion would not change results significantly.

while the male reference has one extra match on the Y chromosome. Note that other combinations are possible, such as target probesets on the sex chromosomes with sequence similarity to the autosomes.

Table 2 summarizes the distribution of putative labels of no change, gain and loss for all probesets on the Human Mapping 500K array according to their intended target locus. Unmatched probesets are excluded from our analysis. Longer contiguous regions of CN difference between female and male samples occur in probesets with intended chromosome X targets, while narrower CNVs are more commonly identified in probesets with intended autosomal targets but also match loci on the sex chromosomes. These ground truth labels of putative CN difference generated for female–male comparisons are subsequently utilized in Section 5 for the assessment of the performance of *DRECS* in the detection of recurrent CN change.

4 METHOD

We now consider the problem of detecting recurrent CNVs across multiple samples. Existing approaches based on segmentation are not suited to this problem since they lack contextual information from multiple samples. Narrow CNVs are often 'smoothed out' (i.e. regarded as noise) as certain assumptions are made about the minimum widths that CNVs may assume. To improve the sensitivity in detecting CN variations/aberrations, we relax any assumptions about their widths and use the information derived from multiple samples to determine the boundaries of these variations/aberrations.

To achieve this, we propose a statistical algorithm, *DRECS*, for sensitive detection of recurrent CN changes (variations and aberrations). In this section, we summarize our approach in three phases and provide a more detailed description in Supplementary Material. The first phase involves the transformation of the raw $\log_2$ ratios into their ranks on a per sample basis, sorted in descending amplitude order and normalized by the total number of probes. This is equivalent to replacing the raw $\log_2$ ratios by their cumulative probability density and reduces the impact of extreme values on the derived statistics. The second phase of the approach is the estimation of the sample variance. Since raw $\log_2$ ratios are converted to normalized ranks, the variance of the transformed values is always known. The variance is that of a discrete uniform distribution over the unit interval of zero (not inclusive) to one. This simplifies the third phase, which is the computation of significance for each probe.

For the computation of significance for each probeset, we begin by defining the null hypothesis as there being no difference between the mean of the transformed probeset and the overall mean of the transformed dataset. Thus, the significance score for each probe is computed as the deviation of the sample probe mean from the population mean. Assuming probabilities are independent, the variance of the mean at each probe location is estimated by the variance of each sample divided by the number of samples. Therefore, regardless of the actual variance of the $\log_2$ ratios, which might be impacted by outlying raw measures, the eventual influence of outliers on the computation of significance is controlled by the normalized rank transformation procedure. We apply the Bonferroni-correction (Bonferroni, 1936) for multiple hypothesis testing to reduce the number of false positives flagged by our approach. We include the full details of our algorithm in the Supplementary Material. Extensions of our approach not detailed in this

article include ranking over chromosomes, chromosome arms as well as computing significance over a pre-specified window of more than one probe. For the purpose of detecting fine-scaled CNVs from SNP microarrays, we choose to employ genome-wide ranking and the computation of statistics over a window of size one, which is equivalent to our description above.

An important advantage of *DRECS* is its scalability. The *DRECS* algorithm is linear with respect to sample size and log-linear to the probe resolution, with an overall time complexity of $\mathbf{O}(NM\log M)$ where N is the sample size and M the number of probesets on the array. The log-linear complexity of *DRECS* is determined by the complexity of the sorting algorithm required to rank the probesets across the array, as it is the most expensive operation in the algorithm. The space complexity of *DRECS* is $\mathbf{O}(NM)$, which is the size of the matrix of input $\log_2$ ratios.

5 EVALUATION

In this section, we discuss the steps taken in preparing the test data for evaluating the accuracy of *DRECS* in detecting CNV. We then report the results of our evaluation using the *t*-test as a benchmark, and examine the impact of segmented CN as an input to *DRECS*. We conclude this section with a discussion on selected biologically relevant CNVs identified by *DRECS*.

5.1 Test data preparation

The required test data are estimates of raw CN, also known as raw $\log_2$ ratios of normal female samples normalized by male samples. The choice of gender comparison provides a natural imbalance in the expected number of copies of the sex chromosomes. Thus, for probesets targeting autosomes that show sequence similarity to the sex chromosomes, a CN imbalance is plausibly expected and this provides positive instances of CN gain or loss. The choice of normalizing female samples by male samples is arbitrary and the inverse may also be applied. To generate the required test data, we chose to use Copy Number Analyzer for Affymetrix GeneChip arrays (CNAG 3.0) (Yamamoto *et al.*, 2007) as it is publicly available.

Female HapMap samples were normalized by a global reference created by averaging multiple best fitting samples from pooled male references from all HapMap populations. The potential non-linear effects attributed to differences in GC content and PCR fragment length were considered and accounted for by producing AsCNAR-normalized $\log_2$ ratios and raw $\log_2$ ratios without AsCNAR normalization. AsCNAR is CNAG's (Yamamoto *et al.*, 2007) inbuilt compensation for GC content and PCR fragment length, which uses quadratic regression. This process was repeated for each of the three populations: European (CEU), African (YRI) and East Asian (CHB + JPT). The normalized female samples were then concatenated and analysed collectively as one large dataset of 127 samples.

5.2 Assessing performance in CN detection

In Figure 2, we examine the process of determining instances of TP, TN, FP and FN based on the predictions made by *DRECS*. In this section, we summarize the performance of *DRECS* by deriving measures of specificity, sensitivity, accuracy and AUC from previous results. We recall

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}, \tag{3}$$

$$\text{AUC}_{\text{weighted}} = \sum_{c} \text{AUC}(c) \times p(c). \tag{4}$$

AUC is computed as the weighted average AUC as defined in Equation (4), where each class, $c \in \{\text{Gain, Loss, No Change}\}$ is weighted according to its prevalence. Since there are >2 classes, the AUC for each class is computed using a one versus many approach. Class instance weightage prevents classes with fewer instance counts from having excessive influence on the results. This approach is similar to that adopted by (Hempstalk and Frank, 2008).

5.3 Benchmarking *DRECS*

In benchmarking the performance of *DRECS*, we choose the two-sided one-sample *t*-test as a benchmark. Our null hypothesis assumes that there are no significant differences between the population probeset mean and the sample probeset mean. Raw $\log_2$ ratios are used for the purpose of the *t*-test, since the rank transformation is specific to our approach. The Bonferroni-correction is applied to the obtained *P*-value for each probeset, and Bonferroni-corrected *P*-values are considered statistically significant if they are less than the value of $\alpha = 0.01$. This could be replaced by false discovery rate (FDR) (Benjamini and Hochberg, 1995) correction if it turns out to be too conservative.

5.4 CNV detection on segmented CN

Segmentation is often applied to raw CN at the individual sample level. In this section, we consider the potential impact of segmentation on the performance of *DRECS*. The segmentation algorithm we choose is HaarSeg (Ben-Yaacov and Eldar, 2008), one of the most recent segmentation algorithms for array comparative genomic hybridization (aCGH) data and reported as being sensitive to narrow CNVs. HaarSeg (Ben-Yaacov and Eldar, 2008) identifies statistically significant breakpoints in data using the maxima of the Haar wavelet transform and segments CN according to the identified breakpoints. The determined CN amplitude between breakpoints is computed as the average of all probes over that interval. HaarSeg (Ben-Yaacov and Eldar, 2008) is reportedly fast in execution and flexible in supporting different variants of microarray data through specific parameters. The parameter values were set to the default values suggested by the authors, while the parameters of quality of measurement and non-stationary variance were not specified since they did not apply to Affymetrix SNP arrays. In Table 3, we compare the performance of *DRECS* with the *t*-test for raw and segmented $\log_2$ ratios, as well as for AsCNAR-normalized $\log_2$ ratios. The performance of *DRECS* was found to be more stable than the *t*-test. In Supplementary Figures D and E, we demonstrate through two contrasting scenarios the impact that segmentation may have on the detection of recurrent CNVs. In Supplementary Figure D, we illustrate an artifactual extension of the width of fine-scaled CNVs in location 1q44 while in Supplementary Figure E we show the 'loss' of an expected narrow CNV in 22q13.31 from the application of Haarseg (Ben-Yaacov and Eldar, 2008) segmentation.

5.5 Biologically relevant CNVs detected

Our evaluation to this point has covered the validation of putative CN differences based on ground truth labels derived from the sequence

Table 3. Performance summary of *DRECS* versus *t*-test

Method	Input	Sensitivity	Specificity	Accuracy	AUC
DRECS	Raw CN	**0.9675**	0.9906	0.9901	**0.9839**
	HaarSeg-Segmented CN	0.9655	0.9981	0.9974	0.9833
	AsCNAR CN	0.9340	0.9525	0.9521	0.9642
	AsCNAR + HaarSeg-Segmented CN	0.9646	**0.9990**	**0.9982**	0.9823
t-test	Raw CN	0.9664	0.9966	0.9959	**0.9831**
	HaarSeg-Segmented CN	**0.9672**	0.4026	0.4146	0.9663
	AsCNAR CN	0.9223	0.9667	0.9658	0.9570
	AsCNAR + HaarSeg-Segmented CN	0.9631	**0.9999**	**0.9992**	0.9816

The performance of *DRECS* was stable to both segmented and raw $\log_2$ ratios. As segmentation tends to be conservative in its treatment of narrow CNVs (Supplementary Fig. D), more fine-scaled changes were detected using raw CN thus resulting in higher sensitivity. Overall, fewer false positives were detected with segmented CN, resulting in higher specificity although exceptions do exist (Supplementary Fig. E). This is the tradeoff between segmented and raw CN. The *t*-test did not demonstrate the same performance stability as *DRECS* and achieved much lower specificity (0.4026) on segmented CN. This occurs from a possible underestimation of sample variance in segmented CN resulting in more instances of CN change being identified as significant (as in Supplementary Fig. E). Normalization with AsCNAR (in CNAG) for GC-content and fragment length did not improve the performance of *DRECS* or the *t*-test over the use of raw $\log_2$ ratios. The values in bold indicate the best performance achieved according to the respective measures of sensitivity, specificity, accuracy and AUC for the various inputs to DRECS and the *t*-test. Overall, the best AUC was achieved by *DRECS* on raw CN.

similarity of probeset sequences. Our findings indicate the existence of narrow CNVs, some covered by the span of a single probe, and the ability of the Affymetrix SNP array technology in detecting them. This empirically validates the sensitivity of *DRECS* and the Affymetrix SNP array in detecting fine-scaled CNVs.

In this section, we proceed to discuss some biologically relevant CNVs detected in the HapMap populations. In conducting our analysis, probesets with more than one matching locus were filtered out prior to the application of *DRECS*, effectively only leaving probesets with unique genomic targets in the dataset. Through the application of *DRECS*, we were able to detect a selection of known CN polymorphisms that stratify populations as well as novel CNVs across SNPs with biological association to physical traits such as height and freckles.

5.5.1 Verification of known CN polymorphisms We verified three randomly selected CN polymorphisms in (McCarroll *et al.*, 2008) to validate the effectiveness of *DRECS* in detecting known recurrent CNVs between populations. The expected CN change for each of the three polymorphisms and the predicted outcomes are detailed in Supplementary Table 1 and Supplementary Figures F, G and H. All three polymorphisms were identified perfectly in terms of the expected CNV between paired comparisons performed across HapMap populations.

5.5.2 Height rs1042725 in the 3′ untranslated region of HMGA2 is causally associated with height (Bouatia-Naji *et al.*, 2009; Lettre *et al.*, 2008; Weedon *et al.*, 2007, see in Supplementary Figs I and J). Specifically, the C allele for rs1042725 was identified to be responsible for the height increase of an individual (Bouatia-Naji *et al.*, 2009). The genotype of the YRI samples are predominantly CC/CT while that of the CHB + JPT samples are CT/TT concurring with expected differences in height between the two populations. Our findings indicate a single copy gain of rs1042725 (mean $\log_2$ ratio difference of 0.32, Bonferroni-corrected P-value = 3.7×10^{-27}) in the YRI population over the CHB + JPT population. We postulate that the observed gain in copy of this SNP represents a novel CN polymorphism, previously undetected due to its small size, that upregulates the expression of the HMGA2 gene to confer a growth advantage.

5.5.3 Freckles Sulem *et al.* (2007) confirms the 'A' allele of rs1540771 to be strongly associated with increased likelihood of freckling. The genotype of the CEU population is predominantly AA/AG, while the YRI population is predominantly GG. Sulem *et al.* (2007) suggests the 'A' allele has been subject to possible positive selection in CEU populations owing to its effect on reduced skin pigmentation. At this stage, it is not known if the association is causal. In our experiments, we observed a single CN gain of rs1540771 in the CEU population over the YRI population (mean $\log_2$ ratio difference of 0.3, Bonferroni-corrected P-value = 5.3×10^{-27}). This SNP (rs1540771) exists as part of a wider CNV region reported in Redon *et al.* (2006) on chromosome six. The potential association of this observed CNV in rs1540771 with skin pigmentation and sensitivity to the sun is novel and should be further investigated beyond the scope of this article.

5.5.4 Verification of phenotype-associated CNVs on Genome-Wide Human SNP 6.0 Array To verify the presence of the observed CNVs reported in Sections 5.5.2 and 5.5.3, we acquired the corresponding Affymetrix Genome-Wide Human SNP 6.0 array dataset for the 270 HapMap samples. In analysing the Affymetrix Genome-Wide Human SNP 6.0 dataset, we applied similar normalization procedures as those used for the analysis of the Affymetrix Human Mapping 500K array. The reference population used to normalize the test population was the population not involved in the direct comparison. For example, in analysing the YRI versus CHB + JPT populations we used the CEU population as a reference. The normalization was done on a 64-bit version of the Affymetrix Genotyping Console version 4.0. We then analysed regions encompassing the height-associated SNP (rs1042725) and the freckles-associated SNP (rs1540771). First, we identified overlapping SNPs between the two SNP array platforms and subsequently excluded probesets with multiple matching loci. We then computed the difference in mean $\log_2$ ratios across samples within the two populations in direct comparison for all remaining SNPs in the region. We denote this difference by

the symbols ε_{500K} and ε_{SNP6} corresponding to their respective arrays. We compute Pearson's correlation coefficient to assess the correlation between the two quantities ε_{500K} and ε_{SNP6}. The Pearson's correlation coefficient corresponding to the region encompassing the height-associated SNP was 0.474 ($P = 6.49 \times 10^{-7}$) and while that of the region encompassing the freckles-associated SNP was 0.448 ($P = 8.34 \times 10^{-6}$). We did observe that certain SNPs were more variable than others in the two regions and found that the exclusion of 10% of the most variable SNPs improved Pearson's correlation coefficient to 0.627 ($P = 4.97 \times 10^{-11}$) and 0.663 ($P = 2.05 \times 10^{-11}$), respectively. The results provide strong empirical evidence in favour of the observed CNVs being genuine as they have been replicated in an independent experiment.

5.6 Impact of non-specific hybridization

Non-specific hybridization in the context of SNP arrays refers to the binding of DNA probe sequences to unintended targets that are not the exact complement of the DNA probe sequence. Whilst this is possible, the stringency of the hybridization and washing procedure and probe design of the Affymetrix Human Mapping SNP arrays is set to minimize this. This is the very basis that allows SNP alleles to bind differentially, since the designed probe sequences differ by a single base pair. It was reported in Zhang *et al.* (2007) that there is notable difference between DNA/DNA duplex formation and DNA/RNA duplex formation, where the former applies to the binding of DNA targets to DNA probe sequences as in the case of SNP arrays, while the latter refers to the binding of RNA targets to DNA probe sequences as in the case of expression arrays. The authors also report that the observed mismatch discrimination was much stronger in DNA/DNA duplexes than in DNA/RNA duplexes as inferred from the signals on the perfect matched (PM) probes as compared to those of the mismatched (MM) probes. Mismatch discrimination is inversely related to the degree of non-specific hybridization on the SNP array. Specifically, they found that only 0.5% of PM probe pairs had a lower signal than MM probe pairs on the 50K Xba240 arrays, which was in sharp contrast to the 30% observed on probe pairs from expression arrays. The above observation was subsequently reiterated in Binder *et al.* (2009) who rationalizes that the difference can be accounted for in terms of the smaller heterogeneity of genomic DNA copies (in terms of sequence and fragment length) and especially of the smaller range of CNV compared with the range of variation of mRNA transcript concentrations where the latter can cover several orders of magnitude while the former typically changes by less than a factor of 10. Thus, we believe that the impact of non-specific hybridization on SNP array probe sequences is minimal and is only likely to affect, if at all, a very small percentage of probes.

5.7 Assessing the impact of genotype frequency differences on observed CNVs

Binding strengths of different alleles in SNPs could result in marginal differences in the resultant binding intensities on SNP arrays, and hence affect the associated $\log_2$ ratios. In the Supplementary Material, we describe a statistical test based on z-statistics that we performed to determine whether genotype frequency differences between the two test populations accounted for the difference in the observed CNVs described in Sections 5.5.2 and 5.5.3. In summary, the test was statistically significant against the null hypothesis with P-values of $P = 9.04 \times 10^{-14}$ and $P = 8.88 \times 10^{-15}$ obtained, respectively, for the height-associated SNP (rs1042725, SNP_A-2216802) and the freckles-associated SNP (rs1570441, SNP_A-4269682). This provides strong evidence for us to conclude that the CNVs are not caused by genotype frequency differences.

6 DISCUSSION AND CONCLUSION

In summary, we have conducted a survey on sequence similarity across all variants of the Affymetrix SNP arrays to evaluate the specificity of probeset targets. We proposed the exclusion of probesets with high sequence similarity to the human genome prior to conducting CN analysis to enhance the reliability of any CN change detected. We also proposed the generation of ground truth labels of putative CNV from our results to serve as a basis for comparing the performance of methods designed for the detection of recurrent CN change.

To detect recurrent CNVs in multiple samples, we developed a statistical approach, *DRECS*, and demonstrated its ability to detect putative CNV in gender with high specificity, sensitivity, accuracy and AUC. By using rank calibration, *DRECS* reduces dependence on raw $\log_2$ ratio values, which are highly susceptible to experimental noise. This method controls the variance of the transformed values and suppresses the influence of outliers. The performance of *DRECS* was also found to be more stable than the t-test in detecting CNVs in both raw and segmented data.

In the application of *DRECS* to the HapMap dataset, biologically significant CNVs were detected across the HapMap populations. Specifically, we identified CNV SNPs with known association to physical traits such as height and freckles between the YRI and CHB + JPT populations and between the CEU and YRI populations, respectively.

We are confident that the application of *DRECS* can be extended directly to the analysis of disease-related high-throughput microarray data, such as that of cancer SNP array data to facilitate the efficient and reliable identification of novel disease-driving CN variations and aberrations down to the resolution provided by the microarray technology.

ACKNOWLEDGEMENTS

We would like to thank Dr Thomas Conway and Dr Bryan Beresford-Smith for running the sequence similarity experiments and for discussions. We also thank our anonymous reviewers for their valuable feedback, which has lead to various improvements in this manuscript.

Funding: This project is partially supported by NICTA. NICTA is funded by the Australian Government through the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Beckmann,J.S. *et al.* (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet.*, **8**, 639–646.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Ben-Yaacov,E. and Eldar,Y.C. (2008) A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, **24**, i139–i145.

Binder,H. *et al.* (2009) Mismatch and g-stack modulated probe signals on SNP microarrays. *PLoS ONE*, **4**, e7862.

Bonferroni,C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Bouatia-Naji,N. *et al.* (2009) Smallness for gestational age interacts with high mobility group A2 gene genetic variation to modulate height. *Eur. J. Endocrinol.*, **160**, 557–560.

Frazer,K.A. *et al.* (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.

Hempstalk,K. and Frank,E. (2008) Discriminating against new classes: one-class versus multi-class classification. In *AI '08: Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence*, Springer, Berlin, Heidelberg, pp. 325–336.

Lettre,G. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.*, **40**, 584–591.

McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39** (Suppl. 7).

McCarroll,S.A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.

Nakamura,Y. (2009) DNA variations in human and medical genetics: 25 years of my experience. *J. Hum. Genet.*, **54**1, 1–8.

Redon, R. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Sulem, P. *et al.* (2007) Genetic determinants of hair, eye and skin pigmentation in europeans. *Nat. Genet.*, **39**, 1443–1452.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Weedon,M.N. *et al.* (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat. Genet.*, **39**, 1245–1250.

Yamamoto,G. *et al.* (2007) Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of Affymetrix single-nucleotide-polymorphism genotyping microarrays. *Am. J. Hum. Genet.*, **81**, 114–126.

Zhang,L. *et al.* (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.*, **35**, e18.