Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (B)

Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays

Eric L. Seiser[1] and Federico Innocenti[1,2]

[1]Center for Pharmacogenomics and Individualized Therapy, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]UNC Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

ABSTRACT: Somatic alterations in DNA copy number have been well studied in numerous malignancies, yet the role of germline DNA copy number variation in cancer is still emerging. Genotyping microarrays generate allele-specific signal intensities to determine genotype, but may also be used to infer DNA copy number using additional computational approaches. Numerous tools have been developed to analyze Illumina genotype microarray data for copy number variant (CNV) discovery, although commonly utilized algorithms freely available to the public employ approaches based upon the use of hidden Markov models (HMMs). QuantiSNP, PennCNV, and GenoCN utilize HMMs with six copy number states but vary in how transition and emission probabilities are calculated. Performance of these CNV detection algorithms has been shown to be variable between both genotyping platforms and data sets, although HMM approaches generally outperform other current methods. Low sensitivity is prevalent with HMM-based algorithms, suggesting the need for continued improvement in CNV detection methodologies.

KEYWORDS: copy number variation, genotyping microarray, hidden Markov model

SUPPLEMENT: Array Platform Modeling and Analysis (B)

CITATION: Seiser and Innocenti. Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays. *Cancer Informatics* 2014:13(S7) 77–83 doi: 10.4137/CIN.S16345.

RECEIVED: August 13, 2014. **RESUBMITTED:** November 18, 2014. **ACCEPTED FOR PUBLICATION:** November 21, 2014.

ACADEMIC EDITOR: J T Efird, Editor in Chief

TYPE: Review

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: seiser@email.unc.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

Somatic alterations in DNA copy number (CNAs, also commonly referred to as somatic copy number variants), ranging in size from kilobases to entire chromosomes, are a hallmark of cancers, and identification of recurrent gains and deletions is vital to improving the biological understanding of these complex diseases. Examples of recurrent CNAs include deletion of the *CDKN2A/B* locus at 9p21, associated with poor prognosis[1–3] in diffuse large B-cell lymphomas, and amplification of *HER2* at 17q21, used as a biomarker for outcome and to guide treatment[4–6] in breast cancer. In contrast to CNAs, germline DNA copy number variants (CNVs) represent gains or deletions of genomic sequence, either inherited or de novo, ranging in size from kilobases to megabases[7–10] and may impact the dosage of genes or regulatory elements. Copy number variants are part of normal variations within the human genome and provide genetic diversity, with over 100,000 recurrent CNVs currently identified in the Database of Genomic Variants.[11] In addition, copy number variation within tumor genomes is comprised of both germline CNVs retained in tumor cells and somatic copy number alterations. While CNVs have been linked with many complex neurological and developmental disorders,[12–14] the role of these variants in cancer is still emerging. Numerous studies have identified CNVs as cancer susceptibility loci in multiple cancer types,[15,16] including deletion of *UGT2B17* and decreased risk of colorectal cancer,[17] deletions in the *APOBEC3* gene cluster and increased risk of

breast cancer,[18,19] and deletion of *WWOX* and increased risk for lung cancer.[20] CNVs may also be relevant to cancer in a pharmacogenetic context, with CNVs in or around genes relevant for absorption, distribution, metabolism, and excretion of anticancer drugs potentially altering drug sensitivity and toxicity in patients.[21–23]

Comparative genomic hybridization DNA microarrays (aCGH) based upon bacterial artificial chromosome (BAC) library sequences or synthesized oligonucleotides were developed specifically for genome-wide interrogation of DNA copy number, and currently have a resolution in the kilobase range.[24,25] Statistical issues associated with aCGH, including removal of technical variation and biological interpretation of array measurements, have led to the development of numerous algorithms for normalization, chromosome segmentation, and copy number calling.[26–30] More recently, genotyping arrays that were initially designed for genome-wide evaluation of single nucleotide polymorphisms (SNPs) have been adapted, through the use of additional processing of probe intensity data, to identify DNA copy number. In addition to the ability to capture both SNPs and copy number changes, genotype arrays can also provide information on copy number neutral loss of heterozygosity (LOH).[31] Although dependent on the platform used, genotype arrays typically have a higher overall resolution compared to comparative genomic hybridization (CGH) arrays, but variability in probe density in genotyping arrays may lead to areas of higher and lower coverage within the genome.[31,32] As with aCGH, computational approaches to handle raw data in the form of allele-specific probe signal intensity derived from the hybridization of a single DNA sample have been developed to process and interpret DNA copy number in genotype arrays. While DNA copy number calling is possible on both Illumina and Affymetrix genotyping platforms, this review provides a brief summary of algorithms freely available to academic users for Illumina genotyping array data, focusing on commonly utilized hidden Markov model (HMM)-based approaches for CNV detection.

HMM-Based Copy Number Detection and Algorithms for Illumina Genotype Data

HMMs are full probabilistic models that function to determine an unknown sequence of states based upon a sequence of observations. Markov models model stochastic processes in which known sequences are produced from a finite number of discrete states, where each new state of a sequence is only dependent upon the previous state. In the Markov model, a change from any one state to another is described by a matrix of transition probabilities. In contrast to the basic Markov model, the sequence of states is hidden in the HMM and can only be inferred through a sequence of observed random variables. In addition to a state transition probability distribution used in the Markov model, each state in an HMM has an emission probability distribution modeling the observed variable as a function of a particular hidden state. To identify the hidden sequence with the highest likelihood based upon the model, HMMs first optimize model parameters (including the emission and transition probability distributions) to best describe the observed sequence of variables. With the model parameters optimized, HMMs can then determine the most probable sequence of hidden states using a dynamic programming approach.

CNV detection from Illumina genotyping data involves utilizing observed signal intensity data from the microarray to determine the hidden copy number at each locus surveyed in the genome. The Illumina GenomeStudio software uses scanned genotype microarray files and platform information files to output genotype calls and other metrics, including those necessary for most copy number calling methodologies: the log *R* ratio (LRR) and the B allele frequency (BAF). The LRR is the logged ratio of the observed *R*, calculated as the sum of normalized signal intensity at each locus (the sum of allele-specific probe intensity at polymorphic loci or probe intensity at non-polymorphic loci), to the expected *R*, the total probe intensity computed by linear interpolation of the observed allelic intensity ratio (θ) with respect to three canonical genotype clusters (genotypes AA, AB, and BB) generated by Illumina using normal samples.[33] The BAF represents the frequency at which the B allele is called at a SNP, interpolated from the θ of three canonical genotype clusters and the sample θ.[33] In general, deviations of the LRR from 0 suggest higher or lower signal intensity than expected for balanced copy number and a BAF deviating from 0, 0.5, or 1 (alleles AA, AB, and BB, respectively) may suggest the presence of copy number imbalance. In HMM-based CNV detection, the observed probes LRR and BAF represent residues emitted from the hidden copy number state at each locus. Each possible copy number state has specific emission probabilities that model the LRR and BAF composition for that state. Additionally, copy number states are dependent on neighboring loci such that there is a probability associated with remaining in the current copy number state or transitioning to a new copy number state at the next locus. After the LRR and BAF emission probability and transition probability parameters are optimized for the observed sequence of LRRs and BAFs, a dynamic programing algorithm can use the optimized parameters in the HMM to determine the most likely sequence of hidden copy number states given the sequence of observed LRRs and BAFs.

Three HMM-based approaches to CNV detection using LRR and BAF data from Illumina genotyping microarrays are QuantiSNP,[34] PennCNV,[35] and GenoCN.[36] The statistical models underlying these software packages share similarities and differences with regard to the core elements of HMMs, including the number of states, emission probabilities of LRRs and BAFs, state transition probabilities, parameter estimation and optimization, and selection of the optimal sequence of hidden copy number states that can be used for biological interpretation.

Hidden copy number states. All three tools employ a total of six discrete copy number states. Two states exist for deletion: full deletion (a null genotype and a copy number of 0) and single deletion (a genotype of allele A or B and a copy number of 1). The normal states have a genotype composition of heterozygotes (genotype AB) or homozygotes (genotype AA or BB) and a copy number of 2. While QuantiSNP has two separate normal states for heterozygotes and homozygotes, PennCNV and GenoCN combine normal heterozygotes and homozygotes into a single state and have an additional normal state representing copy number neutral LOH. Lastly, two states exist for gains, single copy gains (genotype AAA, AAB, ABB, or BBB and a copy number of 3), and double copy gains (genotype AAAA, AAAB, AABB, ABBB, or BBB and a copy number of 4).

Transition probabilities. In QuantiSNP, an exponential function involving the distance between two adjacent SNPs and a length inferred from the data is used to determine an a priori probability that copy number state change occurs between two adjacent loci. In the transition matrix, any state change has equal probability, whereas the probability of remaining in a copy number state differs between non-normal copy number regions, normal heterozygous regions, and normal homozygous regions. PennCNV determines transition probabilities by incorporating parameters estimated by the Baum–Welch algorithm[37] into an exponential function that uses the distance between two adjacent SNP loci and a constant distance, set to either 100 Mb for the copy number neutral LOH state or 100 kb for all other states. The transition matrix uses a common calculation for all state changes and another common calculation for remaining in a state. Transition probabilities in GenoCN, a continuous-time HMM, are determined using an intensity matrix to model the instantaneous transition rate as well as using the distance between two adjacent SNPs. These parameters are used in two exponential functions of the transition matrix, one to calculate copy number state change probabilities and another to calculate the probability of remaining in a copy number state.

Emission probabilities. Emission probabilities for the LRR are calculated in a similar fashion for QuantiSNP, PennCNV, and GenoCN, with the model using a mixture of a uniform distribution to model technical noise in the microarray and a normal distribution to model the LRR observations from a given state. BAF emission probabilities in QuantiSNP are also modeled using a mixture of uniform and normal distributions, but also include half-normal distributions for homozygous genotypes. In PennCNV, BAF emission probabilities are modeled as either a mixture of uniform and normal distributions (BAF between 0 and 1) or a mixture of point mass and truncated normal distributions (BAF is 0 or 1). GenoCN models BAF emission probabilities using a uniform distribution component and several truncated normal distribution components.

Parameter optimization. In QuantiSNP, an objective Bayes-based HMM, normal-gamma conjugate priors are used for emission model parameters, and hyperparameters are estimated from a reference data set using maximum marginal likelihood techniques. The model hyperparameters are then optimized using an expectation maximization algorithm.[38] In contrast, PennCNV estimates initial model parameters empirically using data from several large CNV regions in a large set of training samples. The Baum–Welch algorithm is then used to optimize the parameters by training the model to maximize the likelihood of LRR and BAF observations. Initial parameters of the model are estimated from the LRR and BAF data in GenoCN, and parameter optimization is performed using the Baum–Welch algorithm.

Optimal hidden state selection. Both QuantiSNP and PennCNV utilize the Viterbi algorithm to infer the most likely sequence of copy number states once all parameters within the model have been optimized. In contrast, GenoCN uses the optimized parameters to estimate the posterior probabilities for each SNP of a particular copy number state.

Performance Comparison and Conclusions

Performance reviews of CNV detection using Illumina genotyping data have focused predominantly on HMM-based approaches and commercially available software packages using different statistical models.[39–42] When applied to both real and simulated data, the number of CNVs detected, as well as the identification of the genomic boundaries of the CNVs, varies between algorithms.[39,40,43] Moreover, the ability to accurately detect CNVs may be dependent on microarray platform, copy number of the CNV, CNV size, and the number of array SNPs within a CNV region.[40,43] HMMs have shown a relatively low false positive rate but high false negative rate when calling CNVs, although individual algorithm performance can vary between data sets.[40,42,43] Other commercial CNV calling software packages that employ differing computational approaches, including cnvPartition within Illumina GenomeStudio and Nexus Copy Number (BioDiscovery), have typically demonstrated decreased specificity and sensitivity relative to HMMs in comparative studies.[40,42,43]

To provide a performance evaluation between the three HMM-based CNV detection algorithms, data for three HapMap individuals of European ancestry (NA06985, NA06991, and NA06993) genotyped on the Illumina Human610-Quad BeadChip v1.0 were obtained from the NCBI GEO database[44] (GEO accession: GSE17205). CNV detection for these samples was performed using the default settings in Quanti SNP, PennCNV, and GenoCN. Total CNV regions were identified, along with their associated copy number state, and region overlap between methods was defined as regions having at least one shared probe and copy numbers of the same type (gain or deletion). QuantiSNP and GenoCN identify more CNV regions in these samples when compared to PennCNV (Table 1). Moreover, the majority of CNV regions identified

by PennCNV overlap with both QuantiSNP and GenoCN, and only a limited amount of unique CNV regions are identified by PennCNV. In contrast, the majority of CNV regions identified by QuantiSNP and GenoCN in the HapMap samples are unique to the algorithm. The distribution of CNV size for each sample, in terms of kilobases and also with regard to the number of probes on the genotyping array, was compared between detection methods (Fig. 1). For all three samples, PennCNV identified regions with the largest size (both in length and number of probes), whereas the CNV regions identified by GenoCN were predominantly the smallest (both in length and number of probes). Interestingly, instances were observed in which QuantiSNP and GenoCN identified multiple CNV regions within a single CNV region identified by another algorithm.

Performance of these three CNV detection approaches was further assessed by comparison with previously identified CNVs within these samples. Conrad and others[45] examined 40 HapMap individuals of European and African ancestry using an oligo-based CGH array with a median probe spacing of 53 bp to identify a catalog of putative CNV regions within the sample set. This catalog was subsequently used to generate a CGH-based CNV-typing array that was applied to a set of 450 HapMap individuals of varying ancestry. PCR validation of a subset of CNVs on the CNV-typing CGH array suggested that the false discovery rate in these data was ~15%; therefore, these data were selected as the gold standard for validation of the HMM-based CNV detection in the three aforementioned HapMap samples (NA06985, NA06991, and NA06993). A set of 2419 CNVs from the CNV-typing array contained at least one probe from the Illumina Human610-Quad BeadChip v1.0 and was used for comparison. To identify true positive and false positive CNVs, CNV calls from the HMM-based approaches were considered validated if at least one probe overlapped with the regions identified by Conrad et al and the CNV regions had a similar copy number type (gain or deletion). In addition, CNV regions from Conrad et al that had normal copy number were examined in the same manner to identify true negatives and false negatives. Results from these comparisons show that all three HMM-based detection approaches have a high specificity (>98%) when examining CNV overlap with the Conrad et al regions (Table 2). In contrast, the sensitivity of these tools was low (<15%), with PennCNV demonstrating the worst performance for all samples and GenoCN demonstrating the best performance in two of the three HapMap samples (Table 2). PennCNV had the lowest false discovery rate in two of the three HapMap samples and QuantiSNP the largest false discovery rate, although the converse was true in the third sample. To examine how performance changed relative to genotype microarray probe coverage in CNV regions, the list of gold standard CNVs was filtered to remove CNV regions that encompassed less than a minimum number of probes on genotype microarray. This filtering was performed in a stepwise fashion from a minimum of two probes to five probes. These data showed that as minimum probe coverage within CNV regions was increased, sensitivity also increased

Table 1. CNVs detected in three HapMap samples using QuantiSNP, PennCNV, and GenoCN. The HMM-based CNV detection tools were applied to Illumina Human610-Quad BeadChip v1.0 data from three HapMap samples of European ancestry (NA06985, NA06991, and NA06993). For each sample, the total number of CNVs detected using each algorithm is listed along with the percentage of regions unique to each algorithm and the percentage of regions that overlap results from the other CNV detection algorithms.

QuantiSNP	NA06985	NA06991	NA06993
total regions	103	120	113
% unique regions	33.0%	41.7%	38.1%
% PennCNV overlap	10.7%	14.2%	11.5%
% GenoCN overlap	19.4%	11.7%	15.9%
% PennCNV and GenoCN overlap	36.9%	32.5%	34.5%
PennCNV	**NA06985**	**NA06991**	**NA06993**
total regions	63	61	60
% unique regions	11.1%	8.2%	5.0%
% QuantiSNP overlap	14.3%	21.3%	21.7%
% GenoCN overlap	15.9%	6.6%	10.0%
% QuantiSNP and GenoCN overlap	58.7%	63.9%	63.3%
GenoCN	**NA06985**	**NA06991**	**NA06993**
total regions	169	108	132
% unique regions	48.5%	42.6%	50.8%
% QuantiSNP overlap	11.2%	13.0%	13.6%
% PennCNV overlap	6.5%	4.6%	4.5%
% QuantiSNP and PennCNV overlap	33.7%	39.8%	31.1%

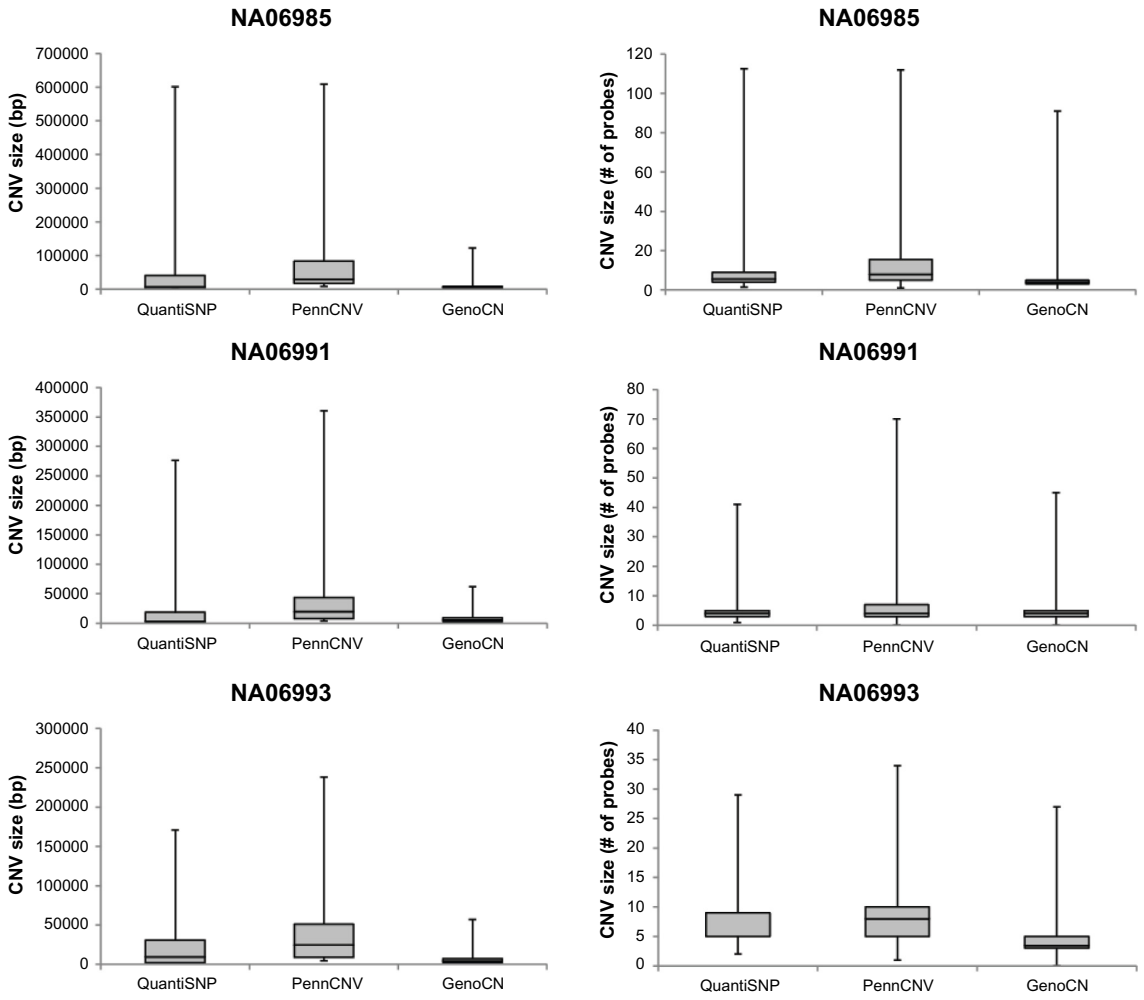

Figure 1. Size of CNVs detected in the HapMap samples using QuantiSNP, PennCNV, and GenoCN. The HMM-based CNV detection tools were applied to Illumina Human610-Quad BeadChip v1.0 data from three HapMap samples of European ancestry (NA06985, NA06991, and NA06993). For each sample, boxplots were generated for CNV sizes from each HMM-based detection method. Boxplots on the left are CNV sizes measured in genomic length (base pairs), and boxplots on the right are CNV sizes measured by the number of genotype microarray probes in the detected region.

because of a decrease in false negatives, although specificity decreased because of a decrease in true negatives (Table 2).

Although all three of the described CNV detection packages for Illumina genotyping microarrays are based upon HMMs, outcomes of performance comparisons presented here and found in other studies highlight how modification of core HMM elements can impact the results of these models. Specifically, variations in the computation of LRR and BAF emission probabilities, transition probabilities, parameter optimization, and hidden copy number state identification impact both the number and size of CNVs detected from genotyping microarray data. Yet these differences in core HMM elements do not lead to one methodology globally outperforming the other methodologies. Instead, variations of the HMMs produce improvements in certain aspects of performance. PennCNV generally has a lower false discovery rate when compared to the other approaches, yet it also has a lower sensitivity and the regions identified predominantly overlap with regions detected using the other two methodologies. QuantiSNP and GenoCN appear to have higher sensitivity and higher false discovery rates than PennCNV, likely because of the increased number of CNV regions detected by these methodologies. Owing to the low sensitivity observed here for all three HMM-based approaches, these tools must be used with the understanding that they will not provide a comprehensive catalog of CNVs within a genome. Furthermore, the observed false discovery rates and sensitivities emphasize the need for additional experimental validation of CNVs identified from Illumina genotyping microarrays when conducting genome-wide association studies of disease risk and outcome. Given the mixed nature of the performance of the currently available HMM-based CNV detection algorithms for Illumina genotyping data, the need exists for continued development of approaches to model the core HMM elements so that sensitivity can be increased and the false discovery rate decreased. Alternatively, utilization of CNV calls from multiple HMM-based algorithms may provide a means to obtain an optimal balance

Table 2. Performance of the HMM-based CNV detection tools in three HapMap Samples. The copy number status of 2419 CNV regions, containing at least one probe on the Illumina Human610-Quad BeadChip v1.0 genotype array, was determined for three HapMap samples (NA06985, NA06991, NA06993) using results from QuantiSNP, PennCNV, and GenoCN. Comparison of the results from the HMM-based approaches to gold standard copy number data from the Conrad et al study allowed for the calculation of sensitivity, specificity, and the false discovery rate for each HMM-based method. These metrics were recalculated as the list of gold standard CNVs was filtered to remove CNV regions that encompassed less than a minimum number of probes on genotype microarray (stepwise from a minimum of two probes to five probes).

	NA06985			NA06991			NA06993		
	QuantiSNP	PennCNV	GenoCN	QuantiSNP	PennCNV	GenoCN	QuantiSNP	PennCNV	GenoCN
1 or more probes									
sensitivity	7.85%	5.39%	7.50%	12.15%	8.68%	14.24%	10.70%	8.72%	11.74%
specificity	98.99%	99.04%	98.81%	99.29%	99.62%	99.29%	99.40%	99.72%	99.45%
false discovery rate	53.66%	61.76%	59.09%	30.00%	24.24%	26.79%	28.89%	18.75%	25.53%
2 or more probes									
sensitivity	11.95%	8.23%	11.46%	17.71%	13.02%	20.31%	14.29%	12.38%	17.33%
specificity	98.57%	98.50%	98.05%	98.98%	99.38%	98.98%	99.24%	99.54%	99.08%
false discovery rate	50.00%	60.61%	59.09%	27.66%	24.24%	25.00%	25.64%	19.35%	25.53%
3 or more probes									
sensitivity	11.38%	10.66%	14.05%	21.38%	17.24%	24.83%	17.95%	16.13%	20.65%
specificity	98.03%	97.92%	97.49%	98.53%	99.10%	98.53%	98.89%	99.33%	98.78%
false discovery rate	56.25%	59.38%	57.50%	29.55%	24.24%	26.53%	26.32%	19.35%	25.58%
4 or more probes									
sensitivity	12.50%	12.63%	15.96%	24.55%	20.91%	26.36%	20.66%	19.17%	22.50%
specificity	97.38%	97.39%	96.63%	97.95%	98.74%	98.10%	98.74%	99.22%	98.59%
false discovery rate	58.62%	58.62%	59.46%	32.50%	25.81%	29.27%	24.24%	17.86%	25.00%
5 or more probes									
sensitivity	14.29%	14.46%	17.07%	23.53%	18.82%	25.88%	20.88%	18.89%	21.11%
specificity	96.76%	96.77%	96.02%	97.82%	98.61%	98.02%	98.63%	99.22%	98.83%
false discovery rate	58.62%	58.62%	60.00%	35.48%	30.43%	31.25%	26.92%	19.05%	24.00%

between sensitivity and false discovery rate using currently available HMM-based CNV detection tools.

Author Contributions

Conceived and designed the experiments: ELS. Analyzed the data: ELS. Wrote the first draft of the manuscript: ELS. Contributed to the writing of the manuscript: FI. Agree with manuscript results and conclusions: ELS, FI. Jointly developed the structure and arguments for the paper: ELS. Made critical revisions and approved final version: ELS. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Lenz G, Wright GW, Emre NC, et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc Natl Acad Sci USA*. 2008;105(36):13520–5.
2. Lee E-S, Kim L-H, Abdullah WA, Peh S-C. Expression and alteration of p16 in diffuse large B cell lymphoma. *Pathobiology*. 2010;77(2):96–105.
3. Jardin F, Jais JP, Molina TJ, et al. Diffuse large B-cell lymphomas with CDKN2 A deletion have a distinct gene expression signature and a poor prognosis under R-CHOP treatment: a GELA study. *Blood*. 2010;116(7):1092–104.
4. Owens MA, Horten BC, Da Silva MM. HER2 amplification ratios by fluorescence in situ hybridization and correlation with immunohistochemistry in a cohort of 6556 breast cancer tissues. *Clin Breast Cancer*. 2004;5(1):63–9.
5. Park JW, Neve RM, Szollosi J, Benz CC. Unraveling the biologic and clinical complexities of HER2. *Clin Breast Cancer*. 2008;8(5):392–401.
6. Li SG, Li L. Targeted therapy in HER2-positive breast cancer. *Biomed Rep*. 2013;1(4):499–505.
7. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet*. 2004;36(9):949–51.
8. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004;305(5683):525–8.
9. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444–54.
10. Abecasis GR, Altshuler D, Auton A, et al; 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061–73.
11. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986–92.
12. Tam GWC, Redon R, Carter NP, Grant SGN. The role of DNA copy number variation in schizophrenia. *Biol Psychiatry*. 2009;66(11):1005–12.
13. Coe BP, Girirajan S, Eichler EE. The genetic variability and commonality of neurodevelopmental disease. *Am J Med Genet C Semin Med Genet*. 2012;160C(2):118–29.
14. Southard AE, Edelmann LJ, Gelb BD. Role of copy number variants in structural birth defects. *Pediatrics*. 2012;129(4):755–63.
15. Krepischi ACV, Pearson PL, Rosenberg C. Germline copy number variations and cancer predisposition. *Future Oncol*. 2012;8(4):441–50.
16. Kuiper RP, Ligtenberg MJL, Hoogerbrugge N, Geurts van Kessel A. Germline copy number variation and cancer risk. *Curr Opin Genet Dev*. 2010;20(3):282–9.
17. Angstadt AY, Berg A, Zhu J, et al. The effect of copy number variation in the phase II detoxification genes UGT2B17 and UGT2B28 on colorectal cancer risk. *Cancer*. 2013;119(13):2477–85.

18. Long J, Delahanty RJ, Li G, et al. A common deletion in the APOBEC3 genes and breast cancer risk. *J Natl Cancer Inst*. 2013;105(8):573–9.
19. Xuan D, Li G, Cai Q, et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. *Carcinogenesis*. 2013;34(10):2240–3.
20. Yang L, Liu B, Huang B, et al. A functional copy number variation in the WWOX gene is associated with lung cancer risk in Chinese. *Hum Mol Genet*. 2013;22(9):1886–94.
21. Gamazon ER, Huang RS, Dolan ME, Cox NJ. Copy number polymorphisms and anticancer pharmacogenomics. *Genome Biol*. 2011;12(5):R46.
22. Kalari KR, Hebbring SJ, Chai HS, et al. Copy number variation and cytidine analogue cytotoxicity: a genome-wide association approach. *BMC Genomics*. 2010;11:357.
23. He Y, Hoskins JM, McLeod HL. Copy number variants in pharmacogenetic genes. *Trends Mol Med*. 2011;17(5):244–51.
24. Coe BP, Ylstra B, Carvalho B, Meijer GA, Macaulay C, Lam WL. Resolving the resolution of array CGH. *Genomics*. 2007;89(5):647–53.
25. Hester SD, Reid L, Nowak N, et al. Comparison of comparative genomic hybridization technologies across microarray platforms. *J Biomol Tech*. 2009;20(2): 135–51.
26. Wineinger NE, Kennedy RE, Erickson SW, Wojczynski MK, Bruder CE, Tiwari HK. Statistical issues in the analysis of DNA copy number variations. *Int J Comput Biol Drug Des*. 2008;1(4):368–95.
27. Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*. 2005;21(19):3763–70.
28. Karimpour-Fard A, Dumas L, Phang T, Sikela JM, Hunter LE. A survey of analysis software for array-comparative genomic hybridisation studies to detect copy number variation. *Hum Genomics*. 2010;4(6):421–7.
29. Roy S, Motsinger Reif A. Evaluation of calling algorithms for array-CGH. *Front Genet*. 2013;4:217.
30. Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*. 2005;21(22):4084–91.
31. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*. 2007;39(7 suppl):S16–21.
32. Coughlin CR, Scharer GH, Shaikh TH. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome Med*. 2012;4(10):80.
33. Peiffer DA, Le JM, Steemers FJ, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res*. 2006;16(9):1136–48.
34. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35(6):2013–25.
35. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007;17(11):1665–74.
36. Sun W, Wright FA, Tang Z, et al. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res*. 2009;37(16):5365–77.
37. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat*. 1970;41(1):164–71.
38. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B*. 1977;39(1):1–38.
39. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic*. 2009;8(5):353–66.
40. Dellinger AE, Saw S-M, Goh LK, Seielstad M, Young TL, Li Y-J. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res*. 2010;38(9):e105.
41. Tsuang DW, Millard SP, Ely B, et al. The effect of algorithms on copy number variant detection. *PLoS One*. 2010;5(12):e14456.
42. Marenne G, Rodríguez-Santiago B, Closas MG, et al. Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish bladder cancer/EPICURO study. *Hum Mutat*. 2011;32(2):240–8.
43. Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*. 2011;29(6):512–20.
44. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41(D1):D991–5.
45. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464(7289):704–12.