Austin Journal of Proteomics, Bioinformatics & Genomics

Open Access

Austin Publishing Group

Research Article

Penncnv-Exomeseq: Genotype Improves Copy Number Variant Detection in Exome Sequencing

Joseph T Glessner[1]*, Jin Li[1], Yichuan Liu[1], Lifeng Tian[1], Kelly A Thomas[1], Ryan Golhar[1], Akshatha Desai[1], Bao-Li Chang[1], Xiao Chang[1] and Hakon Hakonarson[1,2]

[1]Center for Applied Genomics,The Children's Hospital of Philadelphia, Philadelphia, USA

[2]University of Pennsylvania, Perleman School of Medicine, Philadelphia, USA

***Corresponding author:** Joseph Glessner, Bioinformatics Specialist, Center for Applied Genomics, The Children's Hospital of Philadelphia, 3615 Civic Center Boulevard,Philadelphia, USA

Received: January 21, 2015; **Accepted:** March 11, 2015; **Published:** March 13, 2015

Abstract

Background: Sequencing has become a popular method for the generation of large-scale genomic data and with the inundation of such data source comes the necessity for accurate genotype calling of nucleotide bases (A/T/C/G) and copy number (0/1/2/3/4) variants (CNV). The use of SNP arrays as a point of reference for widely used assays for genomic variants, including the variability of different centers and algorithms impacting quality may bear fruit. PennCNV is the most popular method for CNV detection from SNP arrays. Therefore, we observe the unique features that set it apart: namely using both intensity and genotype in tandem to infer CNV states using an HMM and trio based recalling of CNVs to bring de novo rates to an acceptably low level.

Results: Sequencing offers features to assess CNVs intensity which has been leveraged by a number of algorithms, including XHMM, but the valuable feature of genotype for call accuracy has not been incorporated. Here we show derivation of genotype frequency from exome sequencing as a robust data to supplement intensity data in CNV detection.We detect more CNVs at a higher true positive rate than existing methods.

Conclusion: This application of BAF furthermore allows an arsenal of tools to be utilized including PennCNV and ParseCNV for sequencing data. PennCNV-ExomeSeq is freely available at http://penncnvexomeseq.sourceforge.net/.

Keywords: Copy number variant; Whole exome sequencing; Detection; Association

Abbreviations

ZPCARD: z-Score of Principal Components Analysis Normalized Read Depth; WES: Whole Exome Sequencing;BAF: B Allele Frequency; LRR: Log R Ratio; XHMM: Exome Hidden Markov Model; GATK: Genome Analysis Toolkit

Background

Motivated by the shift from SNP arrays to exome sequencing as the most commonly used genomic data, variant calling must be optimized for exome sequencing [1] and whole genome sequencing data sets [2, 3]. Multiple algorithms exists, including Conifer[4], ExomeCNV [5], ExomeCopy [6], ExomeDepth [7], Contra [8], XHMM [9], Excavator [10], Control-FREEC [11], and VarScan2 [12] that were designed for calling CNVs using exome sequencing data. Recently, XHMM has been published which uses exon based intensities normalized by PCA for exome capture biases. A HMM is then applied to segment genomic regions into deletion, diploid, and duplication states. While this has been informative, it's important to further differentiate copy number states into homozygous and hemizygous deletions and duplications, as well as copy neutral LOHs, a six state model. Accurately sizing large CNVs is possible through merging adjacent CNV call fragments into a single CNV call. Importantly, genotypes need to be assessed in tandem with relative intensity to boost CNV calling sensitivity and specificity. Genotype homozygosity is an important observation to correspond to the drop in intensity mode observed in deletions. Furthermore, genotype banding at 1/3 and 2/3 strongly indicates duplication and supports marginal intensity gain signals. Adding genotype to the CNV HMM algorithm improves confidence of CNV calls. Here, we leverage the most powerful softwarefor integration of their best features. XHMM [9], GATK [13], PennCNV [14]and ParseCNV [15] are the major components of thealgorithm cross-talk advanced here. XHMM provides zPCARD for each exon for each sample which is equivalent to Log R Ratio (LRR). GATK provides reference and alternative allele depths and total depth which can be divided to determine B allele frequency (BAF). Then point-wise BAF values are looked up in exomic segments of LRR for the corresponding sample creating the signal intensity file input for PennCNV.

Methods

Using the Genome Analysis Toolkit (GATK)variant call format (VCF) is a convenient procedure since these files are commonly generated for single nucleotide variant (SNV) detection and genotyping. Other tools require a separate software samtoolsrun with the option pileup which requires accessing the much larger binary alignment map (.bam)file and additional computational time.

The GATK VCF provides allele depth for the reference allele (ADRef) and allele depth for the alternate allele (ADAlt). To meet minimal quality control, (ADRef + ADAlt) must be greater than 0 and FILTER=".". Then genotype (GT)= heterozygous (0/1) and GT= homozygous alternate allele (1/1) are placed into 2 different files. B allele frequency (BAF)= ADAlt / (ADRef + ADAlt).

Citation: Glessner JT, Li J, Liu Y, Tian L, Thomas KA, et al. Penncnv-Exomeseq: Genotype Improves Copy Number Variant Detection in Exome Sequencing. Austin J Proteomics Bioinform & Genomics. 2015;2(1): 1009.

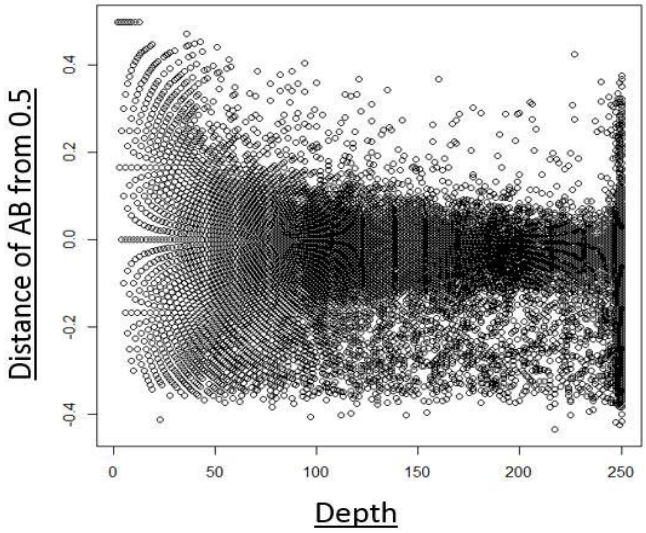

Figure 1: Sequencing read depth vs. distance of AB from 0.5. As sequencing depth increases, the standard deviation of the distribution of AB b allele frequency decreases, albeit at a modest rate with extreme outliers persisting.

Single sample VCFs are used so no homozygous reference(0/0) genotypes are present. The single sample BAFs are concatenated and sorted by position and chromosome (stable sort). The average BAF in the population of was calculated along with standard deviation and count samples contributing to the average. Assuming there are position specific biases in allele depths, we shifted each SNP population distribution mean to the expected values of 0.5 and 1.0 for heterozygotes and homozygotes, respectively.

We assessed the frequency of data at specific nucleotides across our population of samples.

250,533 SNPs >100 Count Occurrences AB only were found in a population of 2,190 VCFs.

184, 037 SNPs >100 Count Occurrences BB only were found in a population of 2,190 VCFs.

A total of 319,672 SNPs >100 count occurrence combined were therefore considered usable for PennCNV calling to make samples comparable in terms of genome resolution and were included in the PennCNV population frequency of b-allele (.pfb) file which sets the probeset to be used across the population of samples. Chromosome "GL" entries were removed.

We then applied clustering to center the expected value of heterozygotes to 0.5 and homozygotes to 1 in order to improve separation of distributions.

Heterozygous 0/1 genotype samples:

CenteredSampleBAF = sampleBAF – (averageBAF – 0.5)

Homomzygous 1/1 genotype samples:

CenteredSampleBAF = sampleBAF – (averageBAF – 1)

Then 0/1 and 1/1 normalized BAFs are concatenated for each sample.All samples were then concatenated again and sorted to calculate the PFB from the centered sample BAF values. Compile_pfb.pl failed due to different SNP sets and orders from GATK VCFs requiring a custom procedure and script.

BAFs are used as scan_region.pl query and LRRs are used as scan_region.pl definition to match single base position BAFs to exon spanning LRRs (XHMM zPCARD). Some XHMM zPCARD entries with 2 values comma delimited where one value was expected were excluded.

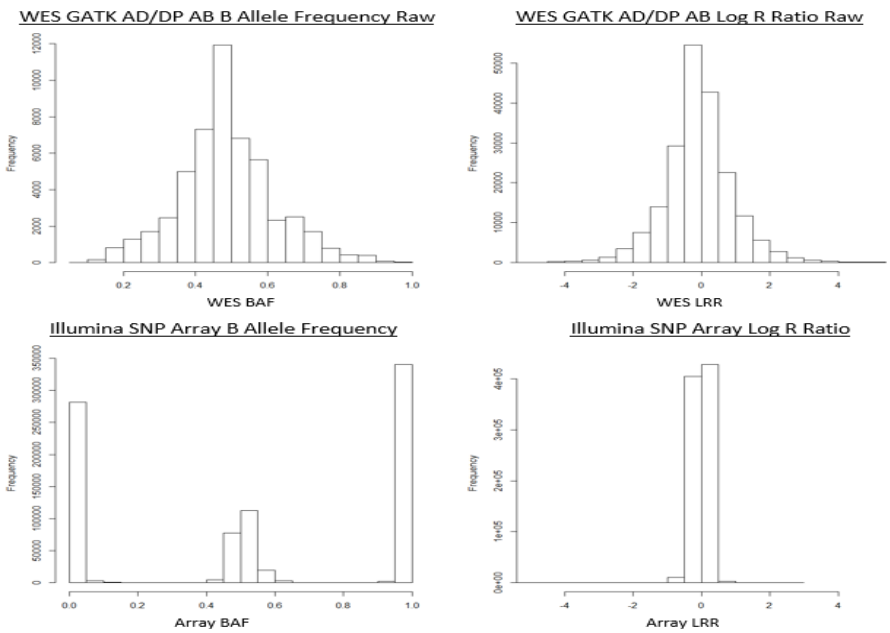

Figure 2: Distribution of genotype (BAF) and intensity (LRR) in exome sequencing and SNP array data. The exome is DNA content of maximal inclusion of bases on exons whereas the SNP array can test and select those with the lowest standard deviation from any number of SNPs in a linkage disequilibrium block. The low standard deviation is crucial for distinguishing copy number states from noise.

Results

Instead of using samtools mpileup which requires another lengthy iteration through all. Bam files, we use the GATK VCF, the most commonly generated file from whole exome sequencing studies. Unfiltered allele depth is provided for reference and alternative alleles at each heterozygous or homozygous alternative allele base position exome-wide to eliminate reporting of the frequent homozygous reference allele. The total depth is approximate and filtered for reads with root mean square of the mapping quality of the reads across all samples <255 or with bad mate pairs. The principal equation is quite simple and constitutes: BAF=ADAlt/(ADRef+ADAlt). We considered only those variants passing the quality filter of GATK. To segment the analysis, we separated the BAF values from AB (0/1) and BB (1/1) genotyped base positions. We investigated the effect of read depth on the clustering of BAF on the expected values 0.5 and 1 for AB and BB genotypes, respectively (Figure 1). We observed no strong bias of BAF at low values of depth other than effects of the fractional values constrained to be low values. We also observe lower read depth in the alternative allele than the reference allele at many AB positions. The BAF values exome-wide for a given sample were more widely distributed around 0.5 than those of a SNP array (Figure 2). Running PennCNV resulted in hundreds of erroneous calls and a high BAF standard deviation quality metric centered around 0.11 compared to aSNP array centered around 0.03.

To address this challenge, we applied the correction of clustering on the population at each SNP derived from the wisdom of SNP array efforts and the assumption that biases in allele depth would be reproducible at each SNP across the population. We clustered AB and BB states of each SNP and adjusted the population mean of each distribution to 0.5 and 1.0, respectively (See methods). Unfortunately, the standard deviation for allele depth in exome sequencing was much higher at 0.12 than SNP arrays at 0.03 (Figure 3).

Given the fact that the full set of 250,506 variant bases was unacceptably noisy, we filtered out SNPs with SD>0.1 to yield a set of 98,168 bases. Lowering the SD threshold down to 0.04, where the majority of SNP array bases reside, would have yielded only 1,793, an unacceptably low coverage. The BB was cut-off at SD>0.01 since the distribution of BB was tighter.

The clustering and high SNP SD filtering resulted in a much tighter distribution of BAF derived from exome sequencing at 0.5 and 1.0 (Figure 4).

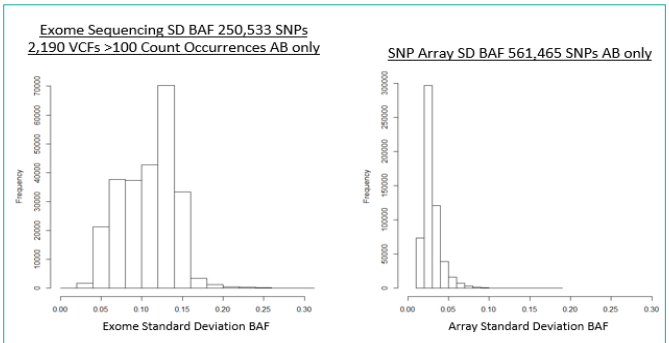

Figure 3: Standard deviation of AB BAF in whole exome sequencing and SNP array data. The standard deviations are quantified for the full panel of markers for each assay. The distribution of exome probes is clearly shifted to higher values, indicating a pronounced noise content hindering clustering ability.

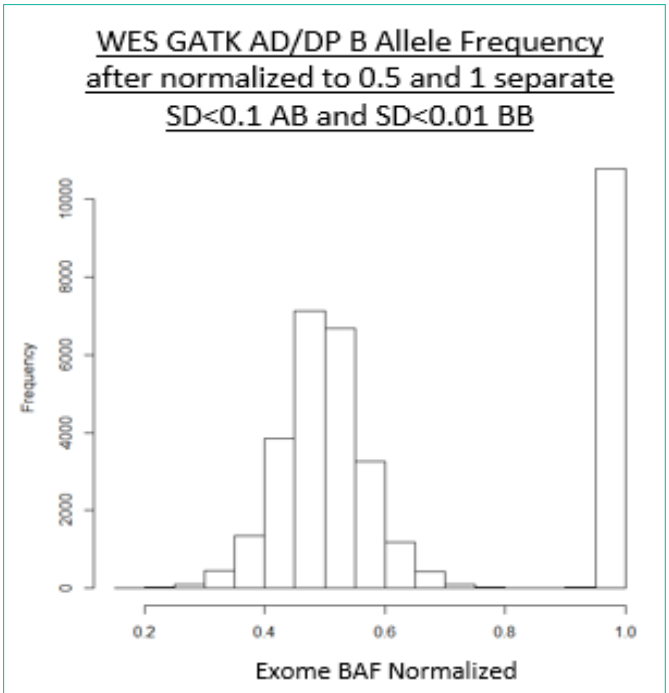

Figure 4: Normalized BAF from exome sequencing showing tighter distribution around 0.5 and 1.0 with distinct distributions. Using clustering and high standard deviation probe removal, the exome sequencing b allele frequency distribution is much improved to clarify copy number modes.

A reasonable BAF_SD of 0.7 was achieved and clear true positive CNVs were recovered (Figure 5) in a call set with mean of 36 calls per sample, more than XHMM which typically provides a mean set of 10 calls per sample. This broadens the realm of possibility of significant association in CNV by increasing the number of putative variants, which have additional confidence by referencing the BAF rather than LRR alone. We have also built in an algorithm to filter XHMM calls based on the frequency of heterozygous genotypes in deletion and duplication CNV calls (Figure 6).

Discussion

In addition to standard CNV calling, PennCNV also allows for family based recalling to be done, therebylowering the putative *de novo* rate which is important for prioritizing *de novo*CNVs and transmission disequilibrium testing in families [16]. Parental origin p-value for *de novo* further prioritizes *de novo*CNVs by measuring consistency of the inheritance model across the length of the *de novo* region in the child and tracing the genotype states back to the parents.

GC wave correction for intensity is now possible in exome sequencing data using the PennCNV gcmodel [17].

The sample quality metrics provided in the PennCNV log is critically important in measuring error and noise properties of samples and be able to exclude those that will bias association.

Having exome sequencing data interpretable by PennCNV allows for true integrative SNP array and exome sequencing CNV detection by combining these assays CNV data on the same samples generated

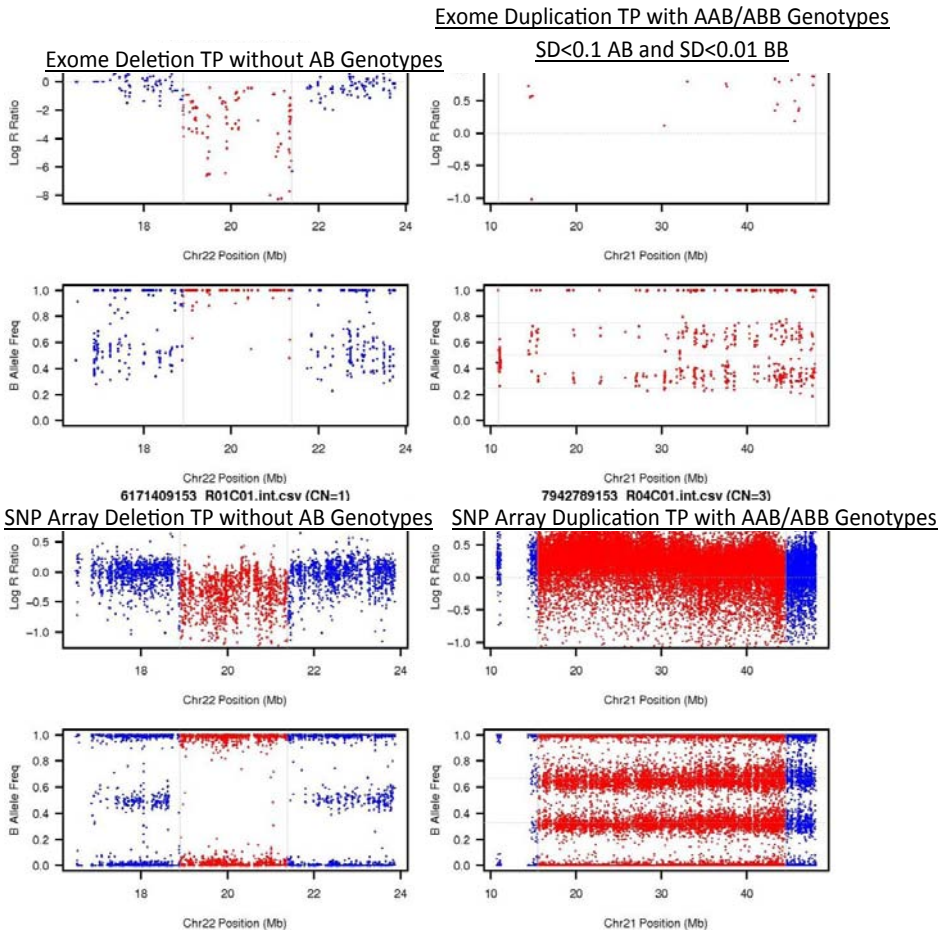

Figure 5: LRR and BAF plots of exome sequencing and SNP array deletion and duplication. Uniform processing of exome and SNP array data show similar patterns of LRR and BAF in tandem. Importantly, the loss of 0.5 values for BAF in exome data is evident in the putative deletion region indicated by the drop in LRR, a feature not considered by existing exome CNV detection software.

by these two methods. Large gaps between exons and between genes cause uncertainty in boundary determination HMM state transition from diploid to CNV states and this is helped by integrating the SNP array data.

PennCNV is by far the most widely used CNV detection software for SNP arrays. Here we adapt exome sequencing data into the proper format and normalized values to fully assess copy variants from WES data.

Mosaicism is another important case of CNV in a subset of cells or a diseased organ with important cytogenetic implications. Our derivation of BAF for sequencing allows visualization of mosaicisim and automated detection using RGADA-MAD [18]. Loss of heterozygosity (LOH) or run of homozygosity (ROH) can only be called by utilizing the allele content, another important feature for the CNV field otherwise missed.

Collectively, we detected an average of 10 CNVs using XHMM and 36 CNVs using PennCNV-ExomeSeq on the same samples with a lower error rate and high confidence score. An average of 2.05 (21%) CNV calls per sample were the same between the two algorithms. This in part reflects calls being excluded by PennCNV-ExomeSeq due to heterozygous genotypes in putative deletions with low intensity waves and in putative duplications with high intensity waves.

There remains a challenge of noisy BAF and LRR derived from exome sequencing causing false positive CNV calls, even after clustering and excluding high standard deviation markers. Plotting the WES GATK UnifiedGenotyper vcf alt AD/DP (BAF) and WES XHMM zPCARD (LRR), we see wider distributions for both compared to SNP array. The main concern is the wide distribution around 0.5 for WES BAF. To tighten up this distribution we propose several new ideas for further investigation: PCA based normalization methods to remove high variance latent components?Filtered AD instead of unfiltered AD?There could be other filtering quality metrics, however minimum depth surprisingly showed modest improvement?Clustering?0/1 genotypes centeredPFB = PFB – (meanPFB – 0.5), including 1/1 for homozygosity, Samtools mpileup –q 15 –Q 20? Genotype base/exon, Intensity base/exon, closest match distance allowance filter, VarScan2 fpfilter. pl bam-readcount, specific heuristics VarScan2 proposed and annotated by bam-readcount were: read position 10-90, strandedness 1-99%, variant reads ≥4, variant frequency ≥5%, distance to 3' ≥20, Homopolymer <5, map quality difference <30, read length difference <25, and mismatch quality sum difference <100. Train hmm exome data iterative batches of 10, Square matrix with homozygote reference genotypes, filter for lower SD SNPs in pfb. We envision that the next

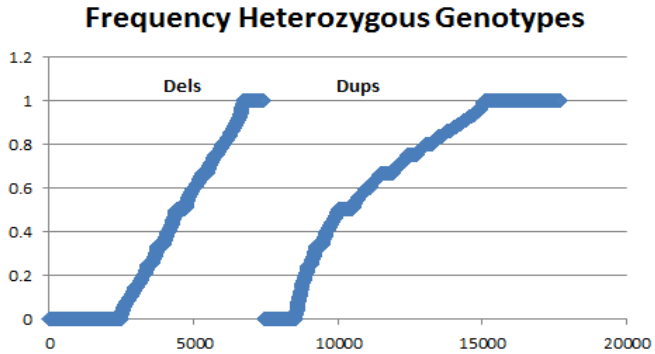

Figure 6: Frequency of heterozygous genotypes within XHMM CNV calls. XHMM deletion CNV calls have a distribution of heterozygous genotypes clearly indicating false positives in a large proportion of calls. XHMM duplication CNV calls have an increased frequency of heterozygous genotypes 0.5 to 1, indicating potential triallelic genotypes in support of the duplication call.

version of the PennCNV software we anticipate to release in about a year, will have vastly improved CNV detection capabilities for WES and other sequencing data.

Genomic variant detection from datasets available from large cohorts of diseased and healthy individuals lies at the heart of genomic association. In order for these associations to be confident and well powered, well validated tools must be continually improved and nuances of data quality considered. We see the application of PennCNV-ExomeSeq software to be highly impactful for elucidating variants completely and confidently.

Cell line immortalization may impact the copy number status of certain loci. Epstein-Barr virus nuclear antigen leader protein localizes to promoters and enhancers with cell transcription factors and EBNA2 [19]. Epstein-Barr virus nuclear antigen (EBNA) leader protein (LP) and EBNA2 (E2) up-regulation of virus and cell gene expression is important for human B-lymphocyte conversion to continuous, potentially malignant, lymphoblast cell lines. Although the molecular mechanismunderlying LP and E2 regulation of cell gene expression have been partially elucidated, LP ChIP sequencing studies have now revealed that LP and LP/E2 interact genome-wide with human B-cell transcription factors, mostly at or near prepatterned promoter sites, to increase cell transcription factor occupancies, increase activation-associated histone marks, and positively affect cell gene transcription. Epstein-Barr virus oncoprotein super-enhancers control B cell growth [20]. Resting B cells have enhancers primed with H3K27ac. Upon EBV infection, NFkB, STAT5, NFAT, and EBNA are recruited to these enhancers to yield immortalized lymphoblast cells with activated EBV super enhancers driving higher level expression of MYC, BCL2, and MIR155. Better understanding of the copy number states genome-wide along with epigenetic gene regulation in EBV cells will create a more complete picture of disease biology.

Conclusion

Here we advance a novel feature to be assessed in whole exome sequencing CNV detection: built in genotype B allele frequency for automated read of sequencing data. We find the standard deviation to be much higher in whole exome sequencing derived BAF than SNP array BAF and address the disparity by population clustering and probe standard deviation quality metric filtering. The integration of BAF in whole exome sequencing CNV detection led to more putative CNV calls to add power to association studies using ParseCNV and allows for interpretation of mosaicism and LOH/ROH regions. We leveraged this collective wisdom of the SNP array era and the tried and true popular tool PennCNV to inform sequencing CNV detection and association through ParseCNV [15].

Availability and Requirements

Project name: Penncnv-ExomeSeq

Project home page:http://penncnvexomeseq.sourceforge.net/

Operating system(s): Platform independent

Programming language: Perl

Other requirements: None

License: GNU GPL

Any restrictions to use by non-academics: Licenseneeded

Competing Interest

The authors declare that they have no competing interests.

Authors' Contribution

JTG conceptualized and carried out the study and drafted the manuscript. JL, YL, LT, KAT, RG, AD, BC, and XC provided useful discussions on software features and statistical model. HH conceptualized the study and drafted the manuscript.

Acknowledgement

The authors thank Kai Wang for excellent mentorship and continuing advice.

References

1. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat 2014;35:899-907.
2. Zhao M,Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinformatics 2013; 14: S1.
3. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. Nature 2011;470:59-65.
4. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res 2012;22:1525-1532.
5. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 2011; 27:2648-2654.
6. Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, et al. Modeling Read Counts for CNV Detection in Exome Sequencing Data. Stat Appl Genet MolBiol 2011;10.
7. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 2012; 28:2747-2754.
8. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics 2012;28:1307-1313.
9. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, et al. Discovery

and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. Am J Hum Genet 2012; 91:597-607.

10. Magi A1, Tattini L, Cifola I, D'Aurizio R, Benelli M, et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. Genome Biol 2013;14:R120.

11. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28:423-425.

12. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568-576.

13. DePristo MA, Banks E, Poplin R, Garimella K, Maguire J, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 2011; 43:491-498.

14. Wang K, Li M, Hadley D, Liu R, Glessner J, 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;17:1665-1674.

15. Glessner JT, Li J, Hakonarson H. ParseCNV integrative copy number variation association software with quality tracking. Nucleic Acids Res 2013; 41:e64.

16. Wang K, Chen Z, Tadesse MG, Glessner J, Grant SF, et al. Modeling genetic inheritance of copy number variations. Nucleic Acids Res 2008; 36:e138.

17. Diskin SJ, Li M, Hou C, Yang S, Glessner J, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 2008;36:e126.

18. González JR, Rodríguez-Santiago B, Cáceres A, Pique-Regi R, Rothman N, et al. A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. BMC Bioinformatics 2011;12:166.

19. Portal D, Zhou H, Zhao B, Kharchenko PV, Lowry E, et al. Epstein-Barr virus nuclear antigen leader protein localizes to promoters and enhancers with cell transcription factors and EBNA2. ProcNatlAcadSci U S A 2013;110:18537-18542.

20. Zhou H, Schmidt SC, Jiang S, Willox B, Bernhardt K, et al. Epstein-Barr Virus Oncoprotein Super-enhancers Control B Cell Growth. Cell Host Microbe 2015;17:205-216.

Austin J Proteomics Bioinform & Genomics - Volume 2 Issue 1 - 2015
ISSN : 2471-0423 | www.austinpublishinggroup.com

Citation: Glessner JT, Li J, Liu Y, Tian L, Thomas KA, et al. Penncnv-Exomeseq: Genotype Improves Copy Number Variant Detection in Exome Sequencing. Austin J Proteomics Bioinform & Genomics. 2015;2(1): 1009.