Detection of Copy Number Variation from Exomes in the DDD and UK10K Projects

Parthiban Vijayarangakannan, Tomas Fitzgerald, Christopher Joyce, Shane McCarthy and Matthew Hurles, on behalf of the DDD and UK10K projects

Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA , UK

Abstract

Large-scale targeted-resequencing projects have become routine in studies that involve large sample sizes not only due to their cost and technological efficiency, but also to extend the analysis easily to a wider range of genetic abnormalities. The DDD (Deciphering Developmental Disorders) project and the UK10K project are two studies in the UK that aim to sequence thousands of individuals using whole-exome sequencing approaches. The DDD project aims to advance clinical genetic practice for children with developmental disorders through systematic development and application of latest microarray and next-gen sequencing methods. Starting from April 2011, the project aims to recruit 12,000 families in three years through the National Health Service Regional Genetics Services in the UK. The UK10K project aims to study diseases in 6000 patients caused by rare genetic changes in the human genome.

Copy Number Variation (CNV) in the human genome has been implicated in a range of rare genetic disorders. Here, we report CoNVex, a novel algorithm for the detection of CNV from targeted-resequencing data. CoNVex utilises the read depth information in probe regions, compares it to a reference (median depth across a set of samples) and detects copy number variable segments within the log2 ratio (of depth over median depth) using an error-weighted score and the Smith-Waterman algorithm. We have applied this to exome data from 393 DDD patient-parent trios. The DDD exome plus design includes probes that are used in the Agilent SureSelect 50Mb library in addition to custom probes. We evaluate the algorithm using samples for which exon-resolution aCGH and SNP array data are available for validation. We also compare our method to other exome-CNV calling algorithms. Our results show that CNV calling from exome data can have at least comparable resolution to single-chip aCGH for calling genic CNVs with high sensitivity and specificity.

Background

Copy Number Variation is widespread throughout the human genome, and can be highly polymorphic between individuals. They have long been studied using low-throughput molecular biology techniques and microarray-based genome scale studies[1,2] to explore the associations between human health and variation. Whole-exome sequencing makes use of the rapid sequencing capabilities of the massively parallel second and third generation sequencing technologies to identify rare variants in ~2% of the genome that codes for proteins[3].

Datasets

Dataset	Library	Samples
DDD trios	SureSelect 50Mb + Custom	393 trios
UK10K	SureSelect 50Mb	4060 samples

CNV Detection and Reporting Pipelines

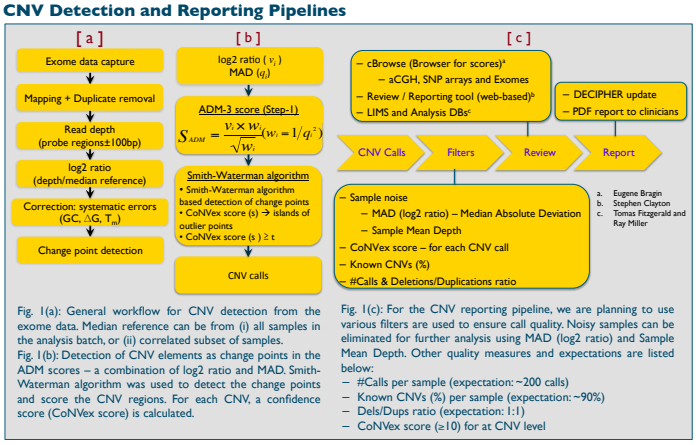


Fig. 1(a): General workflow for CNV detection from the exome data. Median reference can be from (i) all samples in the analysis batch, or (ii) correlated subset of samples.
Fig. 1(b): Detection of CNV elements as change points in the ADM scores – a combination of log2 ratio and MAD. Smith-Waterman algorithm was used to detect the change points and score the CNV regions. For each CNV, a confidence score (CoNVex score) is calculated.

Fig. 1(c): For the CNV reporting pipeline, we are planning to use various filters are used to ensure call quality. Noisy samples can be eliminated for further analysis using MAD (log2 ratio) and Sample Mean Depth. Other quality measures and expectations are listed below:
– #Calls per sample (expectation: ~200 calls)
– Known CNVs (%) per sample (expectation: ~90%)
– Dels/Dups ratio (expectation: 1:1)
– CoNVex score (≥10) for at CNV level

Results and Discussion

CoNVex can generate plots summarising the CNV calls in the analysis batch (Fig. 2). Comparison with ExomeDepth (a HMM based algorithm for detecting CNVs from exomes) and aCGH are shown in Fig. 3. CoNVex and ExomeDepth offer comparable results (size distribution and number of detections) in 100 UK10K sandbox samples (Fig. 3a and 3c). The detections in common have a high % of known CNVs, while the detections by only one algorithm have a lower % of known CNVs. Distribution of CNV size in aCGH differs due to presence of probes in the flanking non-exonic regions in which exome coverage is scarce.

Similarly, comparison with aCGH shows 98% of detections in common are at known CNVs (Fig. 3d) which decreases for single platform detections. [A] and [C] indicate percentages of detectable calls with enough coverage of probes on both platforms. [B] and [D] are the percentages of detectable calls that fall in the regions of known common CNVs. Relatively higher percentage in [D] indicates the maturity of the aCGH platform, while a moderate percentage in [B] indicates that CoNVex can also detect CNVs that are missed by high-resolution aCGH platforms (e.g., single exon deletions) in large scale variation analyses.

Fig. 3e shows the relationship between the CoNVex score and the proportion of calls that are at known CNVs (%). For low scoring calls, the overlap with known CNVs is low and increases markedly for high-scoring detections (≥5). This enables specificity of calling to be adjusted. Similarly, CoNVex's sensitivity for calling detectable CNVs (common CNVs [MAF>5%[2,3]] containing exome baits shows that 82% of known CNVs are detected by CoNVex and most of the undetected CNVs are due to having only a single probed region per CNV. Fig. 4 shows examples of large duplications and deletions. CNV calls' log2 ratio from aCGH and CoNVex are visualised using cBrowse, a web-based browser for the manual review of CNV calls.

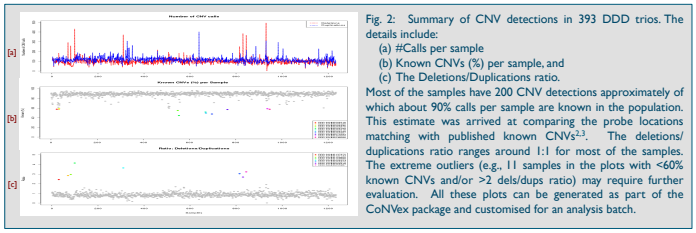


Fig. 2: Summary of CNV detections in 393 DDD trios. The details include:
(a) #Calls per sample
(b) Known CNVs (%) per sample, and
(c) The Deletions/Duplications ratio.
Most of the samples have 200 CNV detections approximately of which about 90% calls per sample are known in the population. This estimate was arrived at comparing the probe locations matching with published known CNVs[2,3]. The deletions/duplications ratio ranges around 1:1 for most of the samples. The extreme outliers (e.g., 11 samples in the plots with <60% known CNVs and/or >2 dels/dups ratio) may require further evaluation. All these plots can be generated as part of the CoNVex package and customised for an analysis batch.

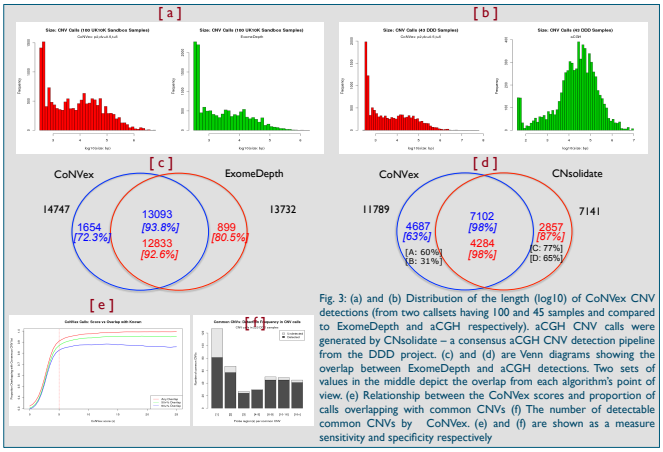
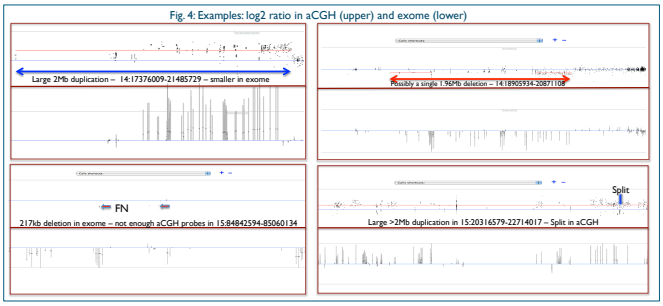


Fig. 3: (a) and (b) Distribution of the length (log10) of CoNVex CNV detections (from two callsets having 100 and 45 samples and compared to ExomeDepth and aCGH respectively). aCGH CNV calls were generated by CNsolidate – a consensus aCGH CNV detection pipeline from the DDD project. (c) and (d) are Venn diagrams showing the overlap between ExomeDepth and aCGH detections. Two sets of values in the middle depict the overlap from each algorithm's point of view. (e) Relationship between the CoNVex scores and proportion of calls overlapping with common CNVs (f) The number of detectable common CNVs by CoNVex. (e) and (f) are shown as a measure sensitivity and specificity respectively



Summary

– We have developed CoNVex – an algorithm to detect copy number variation from exome sequence data that uses an error-weighted score and the Smith-Waterman algorithm for detecting change points (CNV segments). The sensitivity and specificity of the CoNVex scores are high enough to be used for the selection of CNV calls. It has been tested and optimised using large datasets (DDD and UK10K) with thousands of exomes and performs robustly, with a low sample failure rate.

– We have compared our algorithm against ExomeDepth. Initial results indicate that both algorithms perform similarly well despite the underlying differences in their mathematical models.

– We have compared our algorithm against aCGH CNV detections on the same samples. Although the results vary due to the differences in probe design, the overlapping call set's specificity is higher than that of individual algorithms' detections. This indicates that both platforms have false positives and false negatives, and one can be complementary to the other.

– Comparison with other exome CNV algorithms and experimental validation of CNVs are ongoing to further improve the algorithm and QC pipeline.

– CoNVex will be made available as an R package in the near future.

References

1. Tewhey, R., M. Nakano, et al. (2009). "Enrichment of sequencing targets from the human genome by solution hybridization." Genome Biol 10(10): R116.
2. Conrad, D. F., D. Pinto, et al. (2009). "Origins and functional impact of copy number variation in the human genome." Nature.
3. Durbin, R. M., G. R. Abecasis, et al. "A map of human genome variation from population-scale sequencing." Nature 467(7319): 1061-73.
4. Plagnol, V. et al. (2012). "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling" 28(21): 2747-2754.