applied biosystems

Characterization of Exon-Level Genomic Copy Number Changes in 1,855 Normal Individuals

Alan Roter, Benjamin Bolstad, Stephanie Nguyen, Michael Shapero, and Jeanette Schmidt
Thermo Fisher Scientific, 3450 Central Expressway, Santa Clara, CA, USA, 95051

ABSTRACT: POSTER #737

Introduction: Structural chromosomal variation has been studied in healthy individuals for decades [1], but it was not until the advent of chromosomal microarray (CMA) techniques that the extent of copy number variations (CNVs) in the human genome was discovered [2]. CNVs are thought to account for ~1% of the variation between two individuals, and single-nucleotide polymorphisms (SNPs) are thought to account for 0.1% [3]. In this study, a large cohort of more than 1,800 genomic DNA samples from phenotypically normal individuals was characterized for exon-level genomic CNVs to understand the frequency and genomic distribution in the population(s).

Methods: 1,280 genomic DNA samples were acquired from public repositories in the United States. An additional 69 samples from HapMap [4] populations and 506 whole blood samples from normal individuals increased the total to 1,855 samples. The CMA results were analyzed with Applied Biosystems™ CytoScan™ XON Suite. The HapMap and 1000 Genomes Project [5] samples came from 17 previously characterized populations: AFR, ASW, CDX, CEU, CHB, CHS, EAS, EUR, FIN, GBR, GIH, HIS, IBS, JPT, MXL, PUR, and YRI. Genotypes were extracted for population analysis. Principal component analysis (PCA) of HapMap and 1000 Genomes Project samples defined cluster centers for the 17 populations. Proximity to cluster centers was used to infer population for the 696 samples of unknown population. Data were analyzed for copy number calls, which were extracted and compiled into a large dataset. CNV analysis was performed using 506 samples as a composite reference. The CNV dataset was evaluated for the frequency of finding a loss or gain at every genomic location. Log ratio and CNV call data were extracted and constructed into data sets that were analyzed by PCA and correlation.

Results: Genotypes have been shown to be highly predictive of ethnic origin in numerous studies including the 1000 Genomes Project. We utilized this information to characterize ethnicity of our sample population. PCA of individuals clustered by genotype show a similar pattern to previous studies that allowed us to assign ethnicities to individuals in this study. These assigned ethnicities enabled examination of the distribution of CNVs associated with different ethnic origins. CNVs were characterized using CytoScan XON microarrays, allowing detection of genome-wide CNVs at exon resolution. All non-neutral CNVs were resolved to genomic locations, and a data table was constructed of genomic copy number at each genomic location for every individual. Clustering of this data set did not reveal a strong relationship between ethnicity and CNVs. Our current hypothesis is that CNV variation between populations mostly occurs in noncoding DNA sequence, and is missed by an exon-only analysis. Further evidence for this hypothesis is being evaluated.

INTRODUCTION

Why is an exon-level CNV array an important research tool?

- Detection of deletions and duplications (CNVs) in conditions such as developmental delay
- Exon-level CNVs are important in many congenital pathologies
- Limitations of whole-exome sequencing (WES) in detecting small CNVs (<3 exons) due to poor coverage or highly variable regions
- An exome array is a complementary tool to confirm findings by WES
- An exome array can be helpful when a mutation is found by WES and a deletion/duplication is suspected on the other allele of the gene
- An exome array's SNP content allows detection of loss of heterozygosity/absence of heterozygosity (LOH/AOH)

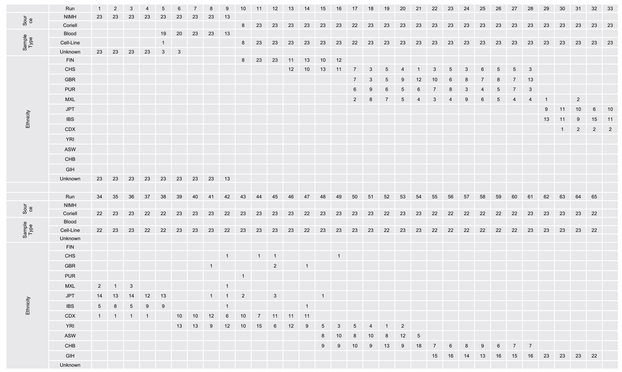
With CytoScan XON Suite, you can:

- Comprehensively detect single-exon deletions and duplications in a cost-effective manner
- Complement NGS mutation analysis with reliable exon-level deletion and duplication detection
- Confirm CNV findings from alternative technologies
- Simplify and streamline sequence variant analysis

RESULTS

1,280 genomic DNA samples were acquired from public repositories in the United States. An additional 69 samples from HapMap populations and 506 whole blood samples from normal individuals increased the total to 1,855 samples. The characteristics of the samples are listed in Table 1.

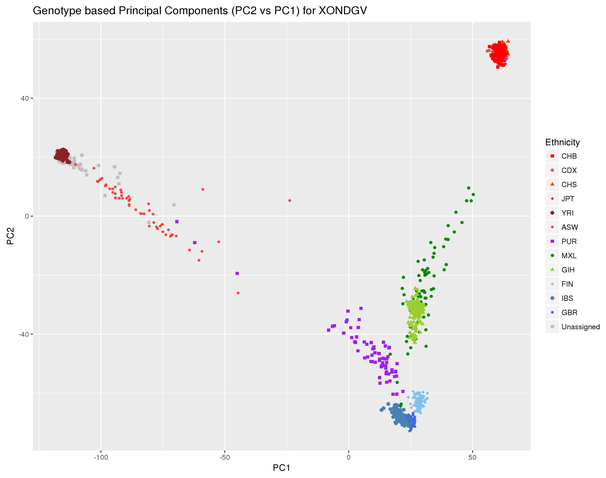
Table 1. Samples used in the study design.



The CMA results were analyzed with CytoScan XON Suite. The HapMap and 1000 Genome Project samples came from 17 previously characterized populations: AFR, ASW, CDX, CEU, CHB, CHS, EAS, EUR, FIN, GBR, GIH, HIS, IBS, JPT, MXL, PUR, and YRI. Genotypes were extracted for population analysis. PCA of HapMap and 1000 Genome Project samples defined cluster centers for the 17 populations.

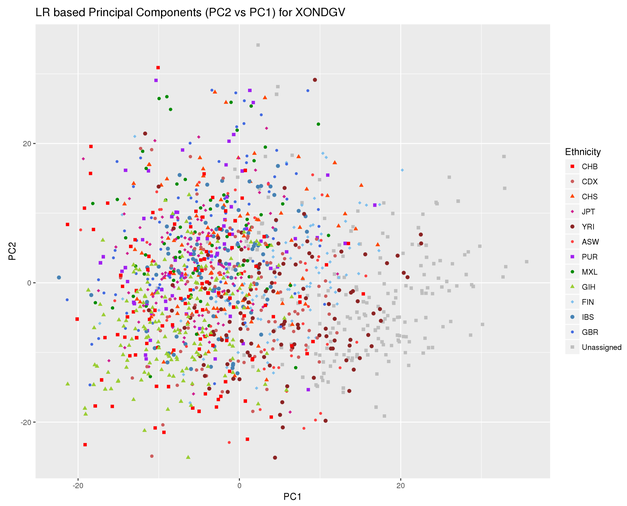
The PCA of genotype data is similar to that seen in other publications. Each sample is plotted as a dot in Figure 1 and is colored by the known population. Gray dots are samples with unknown population. Proximity to cluster centers was used to infer population for the 696 samples of unknown population. The known and inferred populations are used in later analyses.

Figure 1. PCA clustering of ethnicities using genotype profiles.



Having assigned populations for all samples in the study allowed us to evaluate the question of population-specific CNVs. Since log-ratio data is the raw data underlying CNV calls, we first analyzed log-ratio profiles. PCA of the log-ratio profiles of 1,855 samples gave rise to the PC2 vs. PC1 plot in Figure 2. In this plot, samples are colored by population. Each population is generally intermixed with every other population. This indicates that the log-ratio data does not contain information that can separate one population from another. In other words, there are no major population-specific changes in the log-ratio profiles that can discriminate between populations.

Figure 2. PCA clustering using log-ratio profiles.



To further explore the relationships between CNVs and populations, we constructed profiles of the copy number calls for each sample in the study. PCA analysis of the copy number call profiles of 1,855 samples gave rise to the PC2 vs. PC1 plot in Figure 3. In this plot, samples are colored by population. This plot forms an interesting shape similar to the genotype PCA plot, but each population is generally intermixed with every other population. This indicates that the copy number call data does not contain information that can separate one population from another. In other words, there are no major population-specific changes in the copy number profiles that can discriminate between populations.

Figure 3. PCA clustering using copy number call profiles.

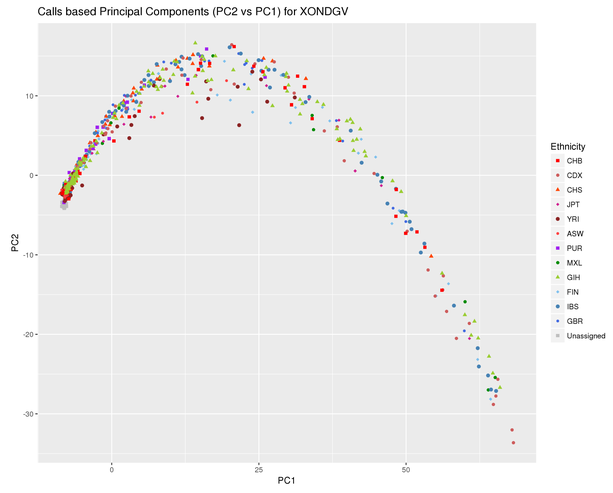


Table 2. Impact of filtering results using the CytoScan XON Database of Genomic Variants (DGV).

Another aspect of this work was to characterize common CNVs in normal samples across multiple populations. Conceptually, CNVs commonly found in normal samples are unlikely to cause disease. This database of common CNVs can be used to filter calls from a product like the CytoScan XON Suite to help one focus on the important changes. Table 2 shows the reduction of the number of relevant calls when using a resource like the data set generated in these studies. This resource has a huge impact on simplifying work when analyzing exon-level copy number data.

For a description of the different gene levels, please see Poster 775, Figure 1.

	Unfiltered	DGV: all	DGV: blood	DGV: cell line
All	179	79	81	85
Level 1 only	43	20	24	22
Level 2 only	30	10	12	11
Level 3 only	40	15	18	17
Level 4 only	61	28	30	33

CONCLUSIONS

In this study and several previous studies, sequence polymorphisms are highly predictive of population origin. In contrast, in this study we do not see a strong relationship between copy number variations and population. Clustering of log ratio data, the underlying signal used for copy number calling, does not result in co-clustering of members of each specific population. Clustering of profiles of copy number calls also does not result in populations clustering together. Some previous studies have demonstrated a relationship between ethnic populates and the prevalence of CNVs. The main difference between those studies and this study is that the former evaluated the whole genome, which includes predominantly intergenic DNA and many segmental duplications. In contrast, this study used CytoScan XON Suite, focused on exon-level copy number changes. Segmental duplications and deletions of intergenic DNA are less likely to cause disease, while deletions of exons and regions of genes are often associated with disease. While there are several examples of important genes being deleted in founder populations (e.g., *GSTT1*, *GSTM1*, *HEXA*, *DMD*, *HBA1/2*, and *HBB*), these represent a very small fraction of the genome and of the genes and would not have yielded a strong signal in this analysis.

This study suggests that population-specific copy number references are likely not needed for copy number studies that focus on functional genes. In addition, having a database of common CNVs in a large population of phenotypically normal individuals is a very useful tool in understanding the relevance of detected CNVs.

REFERENCES

1. Harrison CJ, Jack EM, Allen TD et al. (1985) Investigation of human chromosome polymorphisms by scanning electron microscopy. *J Med Genet* 22:16–23.

2. Pang AW, MacDonald JR, Pinto D et al. (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* 11:R52.

3. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. *Nat Rev Genet* 7:85–97.

4. The International HapMap Consortium (2003) The International HapMap Project. *Nature* 426:789–796.

5. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.

ACKNOWLEDGEMENTS

We would like to thank Chuan Chen, Vicky Huynh, Renee Zamora, and Gargi Mamtora for their contributions in generating laboratory data.

TRADEMARKS/LICENSING

ThermoFisher SCIENTIFIC