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 #373

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 array of CNVs 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 Hamap Project 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 Hamap and 1000 Genomes Project [5] samples came from 17 previously characterized populations: AFR, ASW, CDX, CEU, CHB, CHS, EAS, EUR, FIN, GBR, GIH, HJS, IBS, JPT, MXL, PUR, and YRI. Genotypes were extracted for population analysis. Principal component analysis (PCA) of CytoScan 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. The known population is generally intermixed with every other population. This indicates 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.

RESULTS

1,280 genomic DNA samples were acquired from public repositories in the United States. An additional 69 samples from Hamap Project 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.

<table>
<thead>
<tr>
<th>Sample Type</th>
<th>Number</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamap</td>
<td>1,280</td>
<td>Public repositories in the United States.</td>
</tr>
<tr>
<td>Hamap</td>
<td>69</td>
<td>Additional samples from Hamap Project populations.</td>
</tr>
<tr>
<td>1000 Genomes</td>
<td>506</td>
<td>Whole blood samples from normal individuals.</td>
</tr>
</tbody>
</table>

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

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 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.

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

<table>
<thead>
<tr>
<th>DGV Type</th>
<th># DGVs</th>
<th># Filters</th>
<th>Filter Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1,793</td>
<td>79</td>
<td>0.043%</td>
</tr>
<tr>
<td>GVG: all</td>
<td>1,441</td>
<td>24</td>
<td>0.016%</td>
</tr>
<tr>
<td>GVG: blood</td>
<td>667</td>
<td>11</td>
<td>0.016%</td>
</tr>
<tr>
<td>GVG: cell line</td>
<td>482</td>
<td>17</td>
<td>0.035%</td>
</tr>
</tbody>
</table>

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 variation 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 does not also result in populations clustering together. Some previous studies have demonstrated a relationship between ethnicity and the prevalence of CNVs. The main difference between those studies and this study is the overlap between populations in the 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 expansions and deletions of intergenic DNA are less likely to be associated with disease. While there are several examples of important genes being deleted in founder populations (e.g., DMD, DMD, HEXA, HEXC, and DMB), 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


ACKNOWLEDGEMENTS

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

TRADEMARKS/LICENSING

For Research Use Only. Not for use in diagnostic procedures. © 2019 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified.