Clonal Hematopoiesis: genetic and phenotypic associations

Chao Tian, Pierre Fontanillas, Joyce Y Tung, David A Hinds, the 23andMe Research Team

23andMe, Inc., Mountain View, CA.

Introduction
Age-related clonal hematopoiesis is a common condition that is associated with an increase in hematologic cancer and all-cause mortality. Studies have shown that 10% of individuals over the age of 65 years have recurrent somatic mutations and large-scale somatic events.1 We analyzed SNP array data with custom probes for previously characterized single-nucleotide somatic mutations and were able to detect somatic point mutations at DNMT3A R882C/R882H, IDH2 R140Q, JAK2 V617F and SF3B1 K700E. SNP array data are used routinely to scan for large chromosomal anomalies genome-wide, and it is also useful for detecting large-scale chromosomal mosaicism where a relatively high frequency of mutated cells are present (≥5-10%) 2,3. We were able to use our custom SNP array to detect such large-scale mosaic structure variations in the autosomes of 0.5% of study subjects who were unselected for hematologic phenotypes.

Methods
Study Cohort
The study cohort is composed of ~1.6 million research participants of 23andMe, Inc., who were all consented under an IRB-approved protocol4 and were genotyped on a fully custom array from Illumina.

Single-nucleotide Somatic Mutations
The fluorescence intensity data imported from Illumina GenomeStudio software are transformed to intensity R = signal.a + signal.b and minor allelic burden signal.aR, where signal.a and signal.b represent normalized signals from alleles A and B for a particular locus. The detectable single-nucleotide somatic locus has an unusual intensity distribution (Fig 1), with a smear of individuals showing evidence of clonal single-nucleotide burden in the autosomes, but no clear distribution. We required that the carrier status is significantly associated (OR>3 and P value <1e-10) with old age (age as those who have mutation burdens at least 10 standard deviations away from the mean) and were able to detect somatic point mutations at JAK2 v617f and SF3B1 k700e by age.

Somatic Structural Alterations
We detected 8350 (0.5%) large mosaic structural alterations in ~1.6 million individuals (~2093 deletions, ~1921 duplications, ~4386 uniparental disomy (UPD)). The frequency increased from 0.2% in young individual (age<45) to over 2% in old individuals (age≥45) (Fig 3). The distribution across genome is showed in Fig 5.

Genetic Associations
Multiple germline variants (Table 1) were associated with JAK2 V617F clonal hematopoiesis, some of which have been described previously. Novel associations were identified with IDH2 (rs201009932, P=1.0e-16; OR=3.387), splice site variants in HMGA1 (rs77136196, P=4.9e-9; OR=1.73), eQTL in STK25 (rs4675393, P=2.4e-8; OR=0.782) and RASEF (rs145543052, P=4.9e-8, OR=1.727). HMGA1 chromatin binding proteins are potent oncogenes and overexpressed in acute leukemia and have been shown to be drivers of clonal expansion in myeloid disease in humans. RASEF is involved in cell-growth mechanisms and may play role as an oncogene. TET2 is also significantly associated with JAK2 V617F mutation. DNMT3A and SF3B1 have been shown to be associated with high cholesterol, diabetes, and report higher instances of leukemia (P=1.6e-14, OR=4.15) and blood drops (P=3.8e-16, OR=3).

Discussion
Our work suggested important roles of specific germline mutations underlying the clonal hematopoiesis at several loci. We are extending the analysis to many other commonly mutated genes including TET2, ASXL1 and TP53. Our algorithm for detecting somatic structural alterations is not perfect and still requires improvement. The numbers of somatic structural alterations are expected to be higher in the general populations. Also with the growth of our database, we will have higher sensitivity to detect any phenotypic indications of having certain somatic mutations, such as to determine the relationship between detectable mosaic abnormalities and cancers, which has been actively studied recently mostly via next-generation sequencing technology.

Acknowledgments
We thank 23andMe customers who consented to participate in research for enabling this study. We also thank employees of 23andMe who contributed to the development of the infrastructure that made this research possible.

References

Results (cont.)
We observed that the presence of somatic point mutations was often accompanied by large-scale chromosomal alterations (Fig 6), which is consistent with a previous report5 that people who carry the JAK2 v617f mutation are more likely to also have the UPD across chromosome 9p. People who have high burden of DNMT3A R882C/R882H mutations tend to have deletions across the DNMT3A gene.

Figure 1. Intensity distribution of a typical somatic point mutation.

Figure 2. Frequency of the detectable somatic mutations at DNMT3A R882C/R882H, IDH2 R140Q, JAK2 V617F and SF3B1 K700E by age.

Figure 3. Frequency of the detectable somatic structural variations by age.

Figure 4. Manhattan plot of GWAS results.

Figure 5. Genome distribution of somatic structural alterations (UPD, deletion and duplication).

Figure 6. Co-localization of single-nucleotide somatic mutation and somatic structural alterations. The heat map intensity is showing for the mutated allelic burden at JAK2 v617f and DNMT3A R882C/R882H, respectively.

Phenotype associations
The presence of somatic mutations was also analyzed for association with self-reported phenotypes, adjusting for age, gender and ancestry. The sample sizes for each individual mutation are still too small to find any significant phenotypic indications, except for JAK2 v617f which has relatively larger number of carriers in our database. The JAK2 v617f carriers are less likely to have high cholesterol (P = 4.3e-25, OR = 0.66) and seasonal allergies (P=3.68e-12, OR=0.64), and report higher instances of leukemia (P=1.6e-14, OR=4.15) and blood drops (P=3.8e-16, OR=3).

Download Table 1: Top associations from GWAS

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Gene</th>
<th>Effect</th>
<th>Beta</th>
<th>Univariate P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>109625879</td>
<td>IDH2</td>
<td>−</td>
<td>0.345</td>
<td>1.20E-109</td>
</tr>
<tr>
<td>21</td>
<td>10566530</td>
<td>JAK2</td>
<td>−</td>
<td>0.345</td>
<td>4.90E-08</td>
</tr>
<tr>
<td>21</td>
<td>105864530</td>
<td>JAK2</td>
<td>−</td>
<td>0.345</td>
<td>4.90E-09</td>
</tr>
<tr>
<td>21</td>
<td>1287194</td>
<td>IDH2</td>
<td>−</td>
<td>0.345</td>
<td>1.20E-109</td>
</tr>
<tr>
<td>21</td>
<td>134247047</td>
<td>SF3B1</td>
<td>−</td>
<td>0.345</td>
<td>4.90E-08</td>
</tr>
<tr>
<td>21</td>
<td>1387194</td>
<td>JAK2</td>
<td>−</td>
<td>0.345</td>
<td>4.90E-09</td>
</tr>
<tr>
<td>21</td>
<td>150077846</td>
<td>SF3B1</td>
<td>−</td>
<td>0.345</td>
<td>1.20E-109</td>
</tr>
</tbody>
</table>

Table 1: Top associations from GWAS