Comparative analysis of natural selection effects across human populations

Student: Julia Kornienko

Supervisor: Yury Barbitoff, Bioinformatics Institute
Analysis of protein-coding genetic variation in 60,706 humans - Lek et al., Nature, 2016 (assembled by the Exome Aggregation Consortium - ExAC)

Example of ExAC data:

1 69516 rs776332430 G A 488.39 PASS
AC=1;AF=6.34180e-06;AN=15768

Exon

hsa.11.Indel.1:rs776332430|1

Example of ExAC data:
• 60706 individuals
• Selection of only those PTV alleles (stop gained, splice donor/acceptor)
• Estimation of the selective effects of PTV alleles with very low allelic frequency (AF<<1), therefore contribution of the homozygous PTVs was neglected (as a very low AF in the square).

\[ E(n) = \frac{N U}{S_{\text{het}}} \]

- \( E(n) \) - amount of the loss of function alleles among \( N \) chromosomes, \( U \) - estimated frequency of mutations in the neutral selection model. It is considered that \( n \) has a Poisson distribution with an expected value \( E(n) \) => it is possible to estimate \( S_{\text{het}} \)

Estimating the selective effects of heterozygous protein-truncating variants from human exome data - Cassa et al., Nature Genetics, 2017
**The aim of the project:** estimate the selective effects of heterozygous PTV alleles from gnomAD data (123136 exomes) and perform the comparative analysis of these selective effects both for individual genes and for gene sets among the different human populations.

**Project objectives:**

- Create filtered data set as it had been done in the paper Cassa *et al.*, but for the updated gnomAD data for the 123,136 individuals (instead of 60,706).
- Estimate the selective coefficients per individual genes in the global population for the gnomAD data and to compare them to the published in the Cassa *et al.* paper selective coefficients for the ExAC data.
- Estimate the selective coefficients for the different populations and to perform their comparative analysis.
- Search for the genes and gene sets with population-dependent selective effects.
We calculated the sum of PTV allele counts per each gene (gencode19) with an average coverage > 30x and therefore created the data set containing AC for 17412 genes both for global and local populations.

- For further analysis we left genes with sum of AC < 0.001AN and with Shets (calculated in the naive way as Shet = AN*U/AC) no more than 10 times differ from published Shets for the ExAC data.

- Finally we created data set for further estimation of the selective effects of 12367 genes.
We performed calculation of expected value of the $\text{Pois}(\text{AC, lambda} = \text{AN*U/Shet})$ distribution for each gene, where Shets were varied from 0 to 2 with the step = 0.0001. Thus the selective coefficients were estimated both for global and local populations.
Dependence of distribution of the coefficients for individual genes on the population

Distributions of estimated coefficients for AFR, AMR, EAS and SAS populations were considered to be comparable. As the FIN population is much smaller than other populations and its coefficients distribution differ from other populations, this population was excluded from further analysis. NFE coefficients distribution is more comparable to the global population due to its big size.
Search for the genes with population-dependent selective effects

<table>
<thead>
<tr>
<th>Population</th>
<th>GLOBAL</th>
<th>AFR</th>
<th>AMR</th>
<th>EAS</th>
<th>SAS</th>
<th>NFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>AN</td>
<td>178435</td>
<td>9479</td>
<td>27069</td>
<td>12467</td>
<td>25309</td>
<td>72395</td>
</tr>
</tbody>
</table>

\( \text{p.value} = 5.23e-26 \)

<table>
<thead>
<tr>
<th>Population</th>
<th>GLOBAL</th>
<th>AFR</th>
<th>AMR</th>
<th>EAS</th>
<th>SAS</th>
<th>NFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>32</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>AN</td>
<td>243601</td>
<td>9479</td>
<td>27069</td>
<td>12467</td>
<td>25309</td>
<td>72395</td>
</tr>
</tbody>
</table>

\( \text{Chisq.test} \)
\( \text{p.value} = 0.99 \)

- **2040** of **12367** genes had \( \text{p.value} \) with Bonferroni correction < 0.05

- **30** of **2040** genes had more than 90% PTV alleles (with AC > 10) in one of four populations (AFR = 6, AMR = 6, EAS = 6, SAS = 12)
Some interesting genes with different distribution of AC among the populations

**GDNF - GLIAL CELL LINE-DERIVED NEUROTROPHIC FACTOR**
(26 of 26 variant alleles are among South Asians)

- highly conserved neurotrophic factor
- supports survival and differentiation of dopaminergic and motoneurons

**PAX3 - PAIRED BOX GENE 3** (48 of 53 variant alleles are among South Asians)

- together with Sox10 activates transcription of MITF and RET genes
- controls a cascade of transcriptional events that are necessary and sufficient for skeletal myogenesis.

**Gene-Phenotype Relationships (OMIM)**

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Inheritance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central hypoventilation syndrome</td>
<td>AD</td>
</tr>
<tr>
<td>{Hirschsprung disease, susceptibility to, 3}</td>
<td>AD</td>
</tr>
<tr>
<td>{Pheochromocytoma, modifier of}</td>
<td>AD</td>
</tr>
</tbody>
</table>

**Gene-Phenotype Relationships (OMIM)**

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Inheritance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Craniofacial-deafness-hand syndrome</td>
<td>AD</td>
</tr>
<tr>
<td>Rhabdomyosarcoma 2, alveolar</td>
<td>AR</td>
</tr>
<tr>
<td>Waardenburg syndrome, type 1</td>
<td>AD</td>
</tr>
<tr>
<td>Waardenburg syndrome, type 3</td>
<td>AR, AD</td>
</tr>
</tbody>
</table>

+ **EIF4G3** (translation initiation factor 92% AFR), **NUMBL** (Numb-related gene, 94% EAS), **TERF1** (Telomeric-repeat binding factor, 93% SAS)
+ different olfactory receptor genes in each population
+ other genes
We used curated gene set and hallmark gene set derived from MSigDB Collections GSEA to estimate Shets per gene sets (1377 gene sets in total). To do this we used sums of all ACs and ANs for all genes included in the current pathway for each population. Shets were estimated in the naive way as Shet = AN*U/AC.
Search for the gene sets with population-dependent selective effects

<table>
<thead>
<tr>
<th>Pathway</th>
<th>AC_GLOBAL</th>
<th>AFR</th>
<th>AMR</th>
<th>EAS</th>
<th>SAS</th>
<th>NFE</th>
<th>p_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROTEASOME PATHWAY</td>
<td>185</td>
<td>0.53</td>
<td>0.08</td>
<td>0.02</td>
<td>0.09</td>
<td>0.21</td>
<td>2.20e-133</td>
</tr>
<tr>
<td>IONOTROPIC ACTIVITY OF KAINEATE RECEPTORS</td>
<td>125</td>
<td>0.03</td>
<td>0.55</td>
<td>0.04</td>
<td>0.04</td>
<td>0.27</td>
<td>3.50e-35</td>
</tr>
<tr>
<td>IL 10 PATHWAY</td>
<td>326</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
<td>0.51</td>
<td>0.20</td>
<td>5.91e-94</td>
</tr>
<tr>
<td>APOPTOSIS INDUCED DNA FRAGMENTATION</td>
<td>132</td>
<td>0.05</td>
<td>0.14</td>
<td>0.02</td>
<td>0.52</td>
<td>0.24</td>
<td>1.07e-33</td>
</tr>
<tr>
<td>REGULATION OF IFNG SIGNALING</td>
<td>186</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.60</td>
<td>0.24</td>
<td>9.67e-66</td>
</tr>
<tr>
<td>IL 13 PATHWAY</td>
<td>215</td>
<td>0.03</td>
<td>0.08</td>
<td>0.06</td>
<td>0.56</td>
<td>0.25</td>
<td>8.24e-64</td>
</tr>
</tbody>
</table>

- 746 of 1379 gene sets had p.value with Bonferroni correction < 0.05
- 6 of 746 gene sets had more than 50% PTV alleles in one of four populations (AFR = 1, AMR = 1, EAS = 0, SAS = 4)
- 14 of 746 gene sets had more than 40% PTV alleles in one of four populations (AFR = 2, AMR = 1, EAS = 3, SAS = 8)
Conclusions:

- Data set with calculated sum of PTV ACs for global and local populations was created based on the gnomAD data for 12367 genes.
- Selective effects of heterozygous PTVs were estimated both for individual genes and gene sets for global and local populations.
- With an increase of population size the estimated Shet values have a tendency to decrease.
- Selective effects of 2040 of 12367 genes are population-dependent and 30 of these genes have more than 90% PTV alleles in one of four populations (AFR, AMR, EAS, SAS).
- Selective effects of 746 of 1379 gene sets are population-dependent and 6 of these gene sets have more than 50% PTV alleles in one of four populations (AFR, AMR, EAS, SAS).

To sum up, selective effects for some genes do vary among the populations.

Thank you!
Data Set:
• 60706 individuals
• Mean coverage depth > 30
• Stop gained, splice donor, splice acceptor
• Selection only variants which lead to complete loss of function of the gene

\[
X = \sum x_j
\]
частота loss of function аллелей в гене как сумма частот аллелей по всем PTV сайтам внутри этого гена (на основе отобранного data set)

При \( X \ll 1 \) изменение частоты \( X \) обусловливается притоком de novo мутаций и их оттоку благодаря действию отбора (без учета дрейфа генов):

\[
\frac{d}{dt}X = - S_{het} X(1-X) - S_{hom} X^2(1-X)^2 + U
\]
\( S_{het} \) и \( S_{hom} \) искомые коэффициенты отбора, действующие на PTV варианты (в силу \( X \ll 1 \) \( S_{hom} \) можно пренебречь) \( U \) - оценочная частота мутаций при нейтральном отборе

Для \( N \) хромосом, число loss of function аллелей \( n \) зависит от частоты \( X \):

\[
n = NX = N \sum x_j
\]
И описывается распределением Пуассона с мат. ожиданием:

\[
E(n) = NU/S_{het}
\]