Pushing the boundaries of whole genome sequencing: From genotype to phenotype with a few extras in between.

Symposium on Advances in Genomics, Epidemiology, and Statistics.

*Rasika Mathias*
Johns Hopkins University
June 7\(^{th}\), 2019
Leveraging whole genome sequencing to identify novel determinants of platelet function.

Novel genetic loci identified for telomere length.
Leveraging whole genome sequencing to identify novel determinants of platelet function
Genome-wide single variant tests for association were performed on ~76 million single nucleotide variants (SNV) in 3,125 European Americans (EA) and 730 African Americans (AA) from the Framingham Heart Study (FHS), Older Order Amish Study (OAA), and the Genetic Study of Atherosclerosis Risk (GeneSTAR) Study.

104 variants associated with platelet aggregation in response to ADP, epinephrine, or collagen ($P$-value $< 5 \times 10^{-8}$)
Iterative conditional analyses refines 16 independent loci

<table>
<thead>
<tr>
<th>MAF</th>
<th>hg38</th>
<th>rsID</th>
<th>ref/alt</th>
<th>Nearest Gene</th>
<th>ADP</th>
<th>Epinephrine</th>
<th>Collagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Loci (N = 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>known</td>
<td>Chr10:111139289</td>
<td>rs7097060</td>
<td>T/A</td>
<td>ADRA2A, GPAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>known</td>
<td>Chr1:156689922</td>
<td>rs12041331</td>
<td>G/A</td>
<td>PEAR1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novel Loci (N = 14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr17:16451482</td>
<td>rs575524466</td>
<td>G/A</td>
<td>LRRC75A-AS1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr20:50142397</td>
<td>rs54270794</td>
<td>CTG/C</td>
<td>TMEM189, TMEM189-UBE2V1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr9:28873884</td>
<td>rs185159562</td>
<td>T/A</td>
<td>LINGO2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr10:75490891</td>
<td>rs138028567</td>
<td>A/G</td>
<td>LRMDA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr12:132589485</td>
<td>rs140148392</td>
<td>G/A</td>
<td>FBRSL1, LRCOL1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr11:92185065</td>
<td>rs183146849</td>
<td>A/T</td>
<td>DISC1FP1, FAT3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr1:67128641</td>
<td>rs142001088</td>
<td>C/T</td>
<td>CIorf141</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr6:19109993</td>
<td>rs112157462</td>
<td>T/C</td>
<td>LINC02223, CDH18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr3:96912429</td>
<td>rs61974290</td>
<td>A/G</td>
<td>HS6ST3, LINC00359</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr1:20567949</td>
<td>rs12137738</td>
<td>A/T</td>
<td>FAM43B, CDA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr18:29059923</td>
<td>rs138845466</td>
<td>TAAATA/T</td>
<td>CDH2, MIR302F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr6:121921871</td>
<td>rs58250884</td>
<td>A/G</td>
<td>GJA1, HSF2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>novel</td>
<td>Chr17:21960955</td>
<td>rs1175170</td>
<td>G/C</td>
<td>KCNJ18, UBBP4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MAF**

- **C**: > 0.05
- **L**: 0.01 – 0.05
- **R**: < 0.01

*P < 5x10^-8*
RGS18 controls platelet generation and function

- **RGS18** -/- mice:
  - lower number of bone marrow Megakaryocytes (MK).
  - peripheral platelets are more prone to be activated at baseline compared to wild type.
  - In presence of platelet agonists, platelets aggregate more compared to RGS18 wild types.

- Differential phosphorylation of RGS18 (Serine49 vs Serine216) modifies Calcium gradient in platelets. This change in gradient of calcium dictates the level of platelet activation.
RGS18 and platelet aggregation to Epinephrine

- 7 variants with p < 5E-08
GeneSTAR Data available for eQTL analysis
RNASeq + WGS

PLATELETS
- AA
  - N=84
  - 5,004,400 SNPs
  - 9,500 genes
- EA:
  - N=101
  - 4,433,801 SNPs
  - 9,662

iPSC-derived MEGAKARYOCYTES
- AA
  - N=110
  - 5,500,942 SNPs
  - 4,998 genes
- EA:
  - N=180
  - 5,064,974 SNPs
  - 4,555 genes
Overlap in platelet aggregation loci and eQTL for the top 22 loci.

eQTL analysis of top variants in PLT RNAseq data:

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>pvalue</th>
<th>beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12041331</td>
<td>PEAR1</td>
<td>3.01E-06</td>
<td>0.16729668</td>
</tr>
<tr>
<td>rs1175170</td>
<td>RGS18</td>
<td>2.29E-03</td>
<td>0.08045384</td>
</tr>
</tbody>
</table>
Co-localization Approaches integrating GWAS and eQTLs

Co-localization also supports the role for multiple causal variants mapping to the GWAS peaks.
Left: EA GWAS p-values (blue) and eQTL q-values (red, only q<0.05 shown) on the -log10 scale near 34Mb on chr20, indicating obvious SNP/eQTL colocalization. Right: Percent eQTL among SNPs (y-axis) as a function of p-value cut-off shows an enrichment of eQTLs among stronger associated SNPs for both EAs and AAs. The much higher percentage among EAs indicates the inadequacy of SNP arrays to capture LD among AAs.

Poster # 23: Identifying SNP Associations in Under-Powered Whole-Genome Sequencing Association Studies Using eQTLs. Julius Ngwa.
New Results

**Genome Sequencing Unveils a New Regulatory Landscape of Platelet Reactivity**


doi: https://doi.org/10.1101/621565

This article is a preprint and has not been certified by peer review [what does this mean?].
Novel genetic loci identified for telomere length.
Estimating telomere length from WGS data

**Telseq:**
Filters WGS reads with a specified number of occurrences of the telomere hexamer TTAGGG, adjusting for counts of overall reads with similar GC content (Z. Ding et al., Nucl. Acids Res. 2014)

**Computel:**
Realigns all reads to “telomere reference genome” using bowtie alignment software (L. Nersisyan & A. Arakelyan, PLoS One, 2015)
Estimating telomere length: TOPMed data

[A] Pearson correlation between TelSeq and Computel length estimates on 3362 TOPMed samples. [B] Comparison of computational times for TelSeq and Computel [C] Pearson correlation between TelSeq (left) and Computel (right) and Southern Blot TL estimates on 2429 samples from JHS. Colors indicate sequencing plate in Panels A and C.
TL analysis in TOPMed will be the single largest dataset to investigate genetics of TL in the most diverse sample available!

Rasika Mathias, ScD

Race/Ethnicity
- African ancestry
- Asian ancestry
- European ancestry
- Hispanic/Latino
- Samoan

Percentage

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>N~47,000</th>
<th>N~29,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>African ancestry</td>
<td>28%</td>
<td>29%</td>
</tr>
<tr>
<td>Asian ancestry</td>
<td>5%</td>
<td>8%</td>
</tr>
<tr>
<td>European ancestry</td>
<td>59%</td>
<td>37%</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>3%</td>
<td>26%</td>
</tr>
<tr>
<td>Samoan</td>
<td>5%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Multiethnic genomewide tests for association using 82M sequence identified variants on N=75,176 samples with sequence generate telomere length from TOPMed. All loci had a peak p<5x10^{-8} in the combined meta-analysis.
Multiethnic genomewide tests for association in TOPMed: 22 identified loci

<table>
<thead>
<tr>
<th>Locus Name</th>
<th>Known vs Novel</th>
<th>Discovery (n=46458)</th>
<th>Replication (n=28718)</th>
<th>Meta-Analysis</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AAF</td>
<td>P-Value</td>
<td>AAF</td>
<td>P-Value</td>
</tr>
<tr>
<td>TERT*</td>
<td>Novel</td>
<td>31%</td>
<td>3.3E-24</td>
<td>28%</td>
<td>1.3E-18</td>
</tr>
<tr>
<td>TERC*</td>
<td>Novel</td>
<td>21%</td>
<td>6.7E-19</td>
<td>22%</td>
<td>1.2E-16</td>
</tr>
<tr>
<td>RTEL1*</td>
<td>Novel</td>
<td>70%</td>
<td>1.7E-13</td>
<td>71%</td>
<td>1.0E-09</td>
</tr>
<tr>
<td>SH3PXD2A,OBFC1(STN1),SLK*</td>
<td>Novel</td>
<td>69%</td>
<td>1.8E-19</td>
<td>66%</td>
<td>2.2E-18</td>
</tr>
<tr>
<td>RFWD3*</td>
<td>Novel</td>
<td>44%</td>
<td>4.1E-15</td>
<td>43%</td>
<td>3.6E-03</td>
</tr>
<tr>
<td>NAF1*</td>
<td>Novel</td>
<td>78%</td>
<td>1.8E-09</td>
<td>78%</td>
<td>1.3E-04</td>
</tr>
<tr>
<td>ACYP2*</td>
<td>Novel</td>
<td>17%</td>
<td>2.4E-08</td>
<td>17%</td>
<td>2.0E-04</td>
</tr>
<tr>
<td>TERTF1*</td>
<td>Novel</td>
<td>58%</td>
<td>1.4E-07</td>
<td>54%</td>
<td>2.1E-03</td>
</tr>
<tr>
<td>LINC01592</td>
<td>Novel</td>
<td>0%</td>
<td>5.8E-09</td>
<td>0%</td>
<td>4.5E-02</td>
</tr>
<tr>
<td>TINF2</td>
<td>Novel</td>
<td>1%</td>
<td>1.1E-07</td>
<td>1%</td>
<td>4.3E-07</td>
</tr>
<tr>
<td>SAMHD1</td>
<td>Novel</td>
<td>23%</td>
<td>7.6E-08</td>
<td>26%</td>
<td>1.1E-03</td>
</tr>
<tr>
<td>TERTF2</td>
<td>Novel</td>
<td>31%</td>
<td>2.6E-06</td>
<td>30%</td>
<td>2.9E-04</td>
</tr>
<tr>
<td>ZNF676,ZNF729</td>
<td>Novel</td>
<td>59%</td>
<td>4.5E-07</td>
<td>57%</td>
<td>7.3E-03</td>
</tr>
<tr>
<td>TCL1A</td>
<td>Novel</td>
<td>34%</td>
<td>3.8E-06</td>
<td>37%</td>
<td>1.4E-03</td>
</tr>
<tr>
<td>YY1P2,LRP1B</td>
<td>Novel</td>
<td>0%</td>
<td>3.7E-07</td>
<td>0%</td>
<td>9.7E-03</td>
</tr>
<tr>
<td>LOC100507516</td>
<td>Novel</td>
<td>0%</td>
<td>3.7E-06</td>
<td>0%</td>
<td>2.2E-03</td>
</tr>
<tr>
<td>LINC01429</td>
<td>Novel</td>
<td>14%</td>
<td>3.1E-06</td>
<td>15%</td>
<td>3.2E-03</td>
</tr>
<tr>
<td>RPN1</td>
<td>Novel</td>
<td>26%</td>
<td>4.5E-06</td>
<td>23%</td>
<td>2.4E-03</td>
</tr>
<tr>
<td>DCAF4</td>
<td>Novel</td>
<td>10%</td>
<td>1.2E-05</td>
<td>10%</td>
<td>7.1E-05</td>
</tr>
<tr>
<td>POT1</td>
<td>Novel</td>
<td>21%</td>
<td>2.6E-04</td>
<td>19%</td>
<td>2.0E-05</td>
</tr>
<tr>
<td>ATM</td>
<td>Novel</td>
<td>50%</td>
<td>2.2E-05</td>
<td>49%</td>
<td>6.0E-04</td>
</tr>
<tr>
<td>CHKB-AS1,MAPK8IP2</td>
<td>Novel</td>
<td>30%</td>
<td>9.5E-05</td>
<td>26%</td>
<td>1.2E-04</td>
</tr>
</tbody>
</table>
Association signal for the 22 loci showing all variants having a $p<1\times10^{-5}$ in the meta-analysis, and the ancestry specific signal at each of these variants.
New Results

**Novel genetic determinants of telomere length from a multi-ethnic analysis of 75,000 whole genome sequences in TOPMed**


**doi:** https://doi.org/10.1101/749010

This article is a preprint and has not been certified by peer review [what does this mean?].
Lessons from the TOPMed illustration

High success of the opportunity created to call TL and to understand the genetics of TL.

Multiple novel loci with strong biological plausibility.

New opportunity created to examine TL-phenotype associations for HLB disorders.
Epigenetic Clocks

Clonal Hematopoiesis

Telomere length / GRS

Metabolomic signatures of aging

**Metabolomic signatures of aging**

\[ e = \text{RNAseq expression} / p = \text{proteomics} / m = \text{methylation} / \text{mt} = \text{metabolomics} \]

Bayesian models to test for SNP association including QTL priors

**APPROACH 2**

Traditional GWAS / ignoring omics

**APPROACH 1**

Imputed transcript and protein for gene-based association tests

**APPROACH 3**

WGS

eQTL / pQTL

EXPOSOME: traditional and omics-based approaches

Age          Race          Gender

Environment

Co-morbidities

Clonal Hematopoiesis

Telomere length / GRS
Acknowledgement

Framingham Heart Study
Andrew Johnson
Ming-huei Chen
Benjamin Rodriguez
Jennifer Huffman

Older Order Amish Study
Joshua Lewis
Braxton Mitchell
Brady Gaynor
Kathleen Ryan

GeneSTAR
Ali R. Keramati
Lew Becker
Nauder Faraday
Lisa Yanek
Kruthika Iyer
Margaret Taub
Kai Kammers

Grants and Funding:
The Framingham Heart Study is conducted and supported by the NHLBI in collaboration with Boston University (Contract No. N01-HC-25195). M.H.C., J.D.E. and A.D.J. were supported by National Heart, Lung and Blood Institute Division of Intramural Research funds.
GeneSTAR was supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, and HL112064) and by a grant from the National Institutes of Health/ National Center for Research Resources (M01-RR00052) to the Johns Hopkins General Clinical Research Center. Genotyping services were provided through the RS&G Service by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under U.S. Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute.
This investigation was also supported by National Institutes of Health grants U01 GM074518, U01 HL105198, R01 HL137922, R01 HL121007 and the University of Maryland Mid-Atlantic Nutrition and Obesity Research Center (P30 DK072488).
This study makes use of data generated by the BLUEPRINT Consortium. A full list of the investigators who contributed to the generation of the data is available from www.blueprint-epigenome.eu. Funding for the project was provided by the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 BLUEPRINT.
NHLBI Trans-Omics for Precision Medicine
Whole Genome Sequencing Program