A deep learning based approach for genetic risk prediction

Raquel Dias, PhD.
Senior Staff Scientist
Scripps Research Translational Institute
raqueld@scripps.edu, @RaquelDiasSRTI

Ali Torkamani, PhD.
atorkama@scripps.edu, @ATorkamani
# Whole Genome Sequencing vs. Genotype array

<table>
<thead>
<tr>
<th>Full Data (whole genome sequencing)</th>
<th>Sparse Data (genotype array)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1</td>
<td>0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?</td>
</tr>
<tr>
<td>0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1</td>
<td>0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?</td>
</tr>
<tr>
<td>0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1</td>
<td>0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?</td>
</tr>
<tr>
<td>0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0</td>
<td>0 0 ? ? ? 0 0 1 1 ? ? 1 ? ? ? ?</td>
</tr>
<tr>
<td>0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1</td>
<td>0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?</td>
</tr>
<tr>
<td>0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1</td>
<td>0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?</td>
</tr>
</tbody>
</table>

![Scripps Research Translational Institute Logo](image)
### Whole Genome Sequencing vs. Genotype array

<table>
<thead>
<tr>
<th>Full Data</th>
<th>Sparse Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>~80M genetic variants</td>
<td>~4 million genetic</td>
</tr>
</tbody>
</table>

#### Full Data

- 0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
- 0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
- 0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
- 0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0
- 0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
- 0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1

#### Sparse Data

- 0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?
- 0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?
Genetic imputation problem

<table>
<thead>
<tr>
<th>HapMap or 1,000 Genomes (whole genome)</th>
<th>Reference haplotypes</th>
</tr>
</thead>
</table>
| ...                                   | ...
| ...                                   | ...
| ...                                   | ...
| ...                                   | ...

Cases and Controls typed Genotype array

<table>
<thead>
<tr>
<th>Study genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 ? 0 0 1 1 ? 0 ? ? ? 0 ? 1 ? 1</td>
</tr>
<tr>
<td>0 0 ? 0 0 1 1 ? 0 ? ? ? 0 ? 1 ? 1</td>
</tr>
<tr>
<td>0 0 ? 0 0 1 1 ? 0 ? ? ? 0 ? 1 ? 1</td>
</tr>
<tr>
<td>0 0 ? 0 0 1 1 ? 0 ? ? ? 0 ? 1 ? 1</td>
</tr>
<tr>
<td>0 0 ? 0 0 1 1 ? 0 ? ? ? 0 ? 1 ? 1</td>
</tr>
<tr>
<td>0 0 ? 0 0 1 1 ? 0 ? ? ? 0 ? 1 ? 1</td>
</tr>
<tr>
<td>0 0 ? 0 0 1 1 ? 0 ? ? ? 0 ? 1 ? 1</td>
</tr>
</tbody>
</table>
A typical imputation approach

Multiethnic Haplotype Reference Consortium (HRC) Study genotypes

Reference panel

Mapping

Linkage disequilibrium (LD r²) structure

Prediction

Scripps Research Translational Institute
A typical imputation approach

<table>
<thead>
<tr>
<th>Mulit-ethnic Haplotype Reference Consortium (HRC)</th>
<th>Reference panel</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Genotype table" /></td>
<td><img src="image2.png" alt="Reference panel" /></td>
</tr>
</tbody>
</table>

Mapping

Linkage disequilibrium (LD r²) structure

Study genotypes

Prediction
Polygenic Risk Score (PRS)
Polygenic Risk Calculation

Design
100,000+ subjects

Results
Millions of known variants

Polygenic Risk Score
Cumulative sum

*Trait can often be heterogeneous
   e.g. coronary artery = heart attack, stroke, bypass surgery, etc.
Objectives

1. More accurate and faster imputation
2. Find important genetic variants
3. Better polygenic risk score calculation
Our proposed approach
Denoising autoencoder for image restoration


Genotype imputation case study example

Ground truth
(whole genome sequencing)

Masked input
(genotype array)

0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1
0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1

Mask
Case study: 9p21.3 region of the genome

- Length: 59846 bp
- 846 genetic variants in reference panel (whole genome data)
  - Approx. 200 common variants
  - Approx. 600 rare variants
- Only 17-47 variants in genotype array!!!
- Strong association to coronary artery disease (CAD)
- Genotyped and sequenced in many studies
Training on the reference panel:
Data augmentation strategy

```
... 0 0 0   1  1 1 0  0  1  1 0 0  0  1  1 1   1 ...
... 0 0 0   0  0 1 1  0  0  1  0 1  0  0  1  0  1...
... 0 1 0   1  0 1 0  1  0  1  0 1  1  0  0 1   1 ...

... 0 0 0   0  0 1 1  0  0  1  0 1  0  0  1  0  1...
... 0 0 0   0  0 1 1  0  0  1  0 1  0  0  1  0  1...
... 0 1 0   1  0 1 0  1  0  1  0 1  1  0  0 1   1 ...
```

Reference Whole Genome

Mask

Masked input

Autoencoder

Reconstructed Output
Customized Sparsity Loss Function

- Sparsity loss with Kullback-Leibler (KL) / cross entropy element:
  \[ D_{KL}(\rho || \hat{\rho}) = \rho \log \left( \frac{\rho}{\hat{\rho}} \right) + (1 - \rho) \log \left( \frac{1 - \rho}{1 - \hat{\rho}} \right) \]

- Customized loss adjusted for hidden activation sparsity:
  \[ \text{loss} = MSE + \beta \sum_{i=1}^{n} D_{KL(i)} \]

- Mean Squared Error:
  \[ MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) \]
Hyper parameters to be optimized

• $\beta$
• $\rho$
• Activation functions
• L1/L2 regularizers
• Learning rate
• Batch size
Parallel Grid Search
Hyperparameter optimization approach

Hyperparameter combinations

100 X grid search samples

9 GPUs available:
- 7 GTX 1080,
- 1 Titan V,
- 1 Titan Xp

Trained model performance

860 hours
(sequential run, 100 epochs)

Accuracy, loss
Sparsity, MSE
Grid Search Results: training accuracy

Accuracy

Density

Density

0.0 0.2 0.4 0.6 0.8 1.0

0 1 2 3 4 5

0 5 10 15

0.80 0.85 0.90 0.95

Scripps Research
Translational Institute
Grid Search results: assessing best hyperparameter values
Effect of hyper parameter values in training accuracy

- $\beta$
- $\rho$
- Learning rate
Effect of hyper parameter values in training accuracy

- $\beta$
- $\rho$
- Learning rate

Pearson correlation ($r^2$)
Effect of hyper parameter values in training accuracy

- $\beta$
- $\rho$
- Learning rate
Effect of hyper parameter values in training accuracy

- $\beta$
- $\rho$
- Learning rate
Optimizing batch size: training accuracy

- Accuracy
- Loss

- 10 batches
- 50 batches
- 100 batches
- 1000 batches
Optimizing batch size: training run time

- 10 batches
- 50 batches
- 100 batches
- 1000 batches
Testing on multiple case studies

- **Atherosclerosis Risk in Communities (ARIC)**
  - More than 3000 samples
  - Whole genome sequencing (846 variants, 0% mask, ground truth)
  - Affymetrix 6.0 genotype array (17 variants, 98% mask, input data)

- **Framingham Heart Study (FHS)**
  - More than 500 samples
  - Whole genome sequencing (846 variants, 0% mask, ground truth)
  - Illumina 500K genotype array (47 variants, 95% mask, input data)
  - Illumina 5M (93 variants, 89% mask, input data)
Accuracy in additional case studies: Proposed approach versus common statistic methodology

Performance: all variants

Performance: rare variants
Accuracy in additional case studies: Proposed approach versus common statistic methodology

Performance: all variants

Performance: common variants
Run time: Proposed approach versus common statistic methodology
Linkage disequilibrium structure: ARIC

Ground truth

Prediction

All variants  Rare variants  Common variants

Linkage disequilibrium ($r^2$)
Linkage disequilibrium structure: FHS

Ground truth

Prediction

All variants  Rare variants  Common variants

Linkage disequilibrium (LD) $r^2$
Interpretability: identifying representative genetic variants

Maximal information criteria
Conclusions

- **Grid search** was able to find high accuracy models (>0.90)
- **Hyperparameters** played an important role in training performance
- Reconstruction of genetic variants from very sparse data with **high accuracy** (>0.80)
- Superior computational performance, **faster predictions**
- **Fine parameter tuning** may be necessary
Future steps

- Expand to other genomic regions, fine parameter tuning
- Use imputation autoencoder results as input for polygenic risk score calculation
Future steps

• Focal loss to compensate for rare variants

\[
\text{CE}(p_t) = -\log(p_t)
\]

\[
\text{FL}(p_t) = -(1 - p_t)^\gamma \log(p_t)
\]

Limitations: expanding the methodology to other genomic regions
Acknowledgements

Ali Torkamani PhD
Shang-Fu Chen
Elias Salfati PhD
Doug Evans
Shuchen Liu
Alex Lippman
Nathan W. PhD
Emily Spencer PhD
Eric Topol MD

NVIDIA collaborators
Johnny Israeli
Carla Leibowitz
Fernanda Foertter
Brian Welker
David Nola

Funding
NIH/NCRR flagship CTSA Grant

Contact:
raqueld@scripps.edu, @RaquelDiasSRTI
atorkama@scripps.edu, @Atorkamani
Thanks for your attention!!