Impact of regulatory variation from RNA to protein

Alexis Battle,¹,²‡ Zia Khan,³‡ Sidney H. Wang,³‡ Amy Mitrano,³ Michael J. Ford,⁴ Jonathan K. Pritchard,¹,²,⁵§ Yoav Gilad³§
Genotype to Phenotype

Genome Wide Association Studies (GWAS)
Where are the trait associated variants?

Controls

Cases

Published Genome-Wide Associations through 12/2013
Published GWA at p≤5X10⁻⁸ for 17 trait categories

Majority of the complex trait associated variants lie in the noncoding regions of the genome
Coding vs. Noncoding variation

DNA

Coding variant

Gene A

Synonymous vs. Nonsynonymous

Variation in protein function

Noncoding variant

Gene C

Regulatory vs. Nonregulatory

Variation in gene expression level

Variation in translation

Variation in protein level
Genetic mapping of:

1. mRNA Level (RNA-Seq)
2. Translation Level (Ribosome Profiling)
3. Protein Level (Quantitative Mass Spectrometry)

HapMap Yoruba (Ibadan, Nigeria) lymphoblastoid cell lines (LCLs)
Genotype Data (N=75)

- ~3.1 million SNPs genotyped (Phase II HapMap data)
- Imputation → ~15.8 million variants (14.9 million SNPs, 0.9 million indels)

RNA-Seq Data (N=75)

- Median of 8.6 million reads per individual uniquely mapped to Ensembl genes
- 16,614 genes

Pickrell et al. (2010). Nature. 464, 768-772
Ribosome Profiling (N=72)

30-50 million live LCLs

Cycloheximide treatment (blocks translational elongation)

Cell lysis and nuclease digestion

Monosome isolation (gradient centrifugation)

Purify rRNA depleted RNA (rRNA removal probes)

Library preparation and sequencing
Ribosome Profiling (QC)
Protein Quantification

SILAC = Stable Isotope Labeling by Amino Acids in Cell culture
Protein Quantification (QC)

Correlation of groups of distinct peptide quantifications from the same protein from cell line GM18916.

Spearman's correlation = 0.838 N = 3540

Histogram of correlations from all LCLs.
Number of Genes Quantified

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Genes tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein abundance</td>
<td>4,381</td>
</tr>
<tr>
<td>Ribosome occupancy</td>
<td>15,059</td>
</tr>
<tr>
<td>mRNA expression</td>
<td>16,614</td>
</tr>
</tbody>
</table>

4,322 common
**cis-Quantitative Trait Loci (QTLs) Mapping**

1. For every SNP within a ±20kbp window of the gene with minor allele frequency (MAF) of 0.1 or greater, record the Pearson correlation with the expression phenotype along with the corresponding p-value.

2. Take the minimum p-value $p_{\text{best}}$ among all SNPs tested for the gene.

3. Repeat (1-2) for 10,000 permutations of the genotype sample labels, obtaining $p_{\text{perm}(1)}\ldots p_{\text{perm}(10000)}$.

4. Estimate an empirical gene-level p-value for the most significant QTL based on the permutations as the fraction of permutation p-values at least as significant as the original p-value: $\sum_i (p_{\text{perm}(i)} \leq p_{\text{best}})/10,000$.

### Table 1. Number of cis-QTLs identified at FDR of 10%.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Genes tested</th>
<th>No. of cell lines</th>
<th>cis-QTLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein abundance</td>
<td>4,381</td>
<td>62</td>
<td>278</td>
</tr>
<tr>
<td>Ribosome occupancy</td>
<td>15,059</td>
<td>72</td>
<td>939</td>
</tr>
<tr>
<td>mRNA expression</td>
<td>16,614</td>
<td>75</td>
<td>2,355</td>
</tr>
</tbody>
</table>
Overlap of cis-QTLs – Genome-wide Testing

4,322 genes common to all three datasets

~15% of the genes with expression QTL (eQTL) have protein QTL (pQTL)

~25% of the genes with ribosome occupancy QTL (rQTL) have protein QTL (pQTL)
Overlap of *cis*-QTLs – Replication Testing

SNP-gene pairs identified as QTLs in one phenotype tested in the second phenotype

35% of eQTLs are pQTLs
51% of rQTLs are pQTLs

- Majority of the genetic regulation at mRNA level are not reflected at the protein level
→ Majority of genetic variants affecting transcript levels also alter ribosomal occupancy, typically with a similar magnitude of effect

→ Many eQTLs have attenuated (or absent) effects on steady-state protein levels
Reduction in effect size in protein data is robust with respect to number of transcripts per gene and absolute protein quantification.
Protein- & Expression-Specific QTLs

68 protein-specific QTLs (psQTLs)

76 expression-specific QTLs (esQTLs)

esQTLs have similar effect sizes on ribosome occupancy
psQTLs have larger effect sizes on protein levels compared to ribosome occupancy
→ psQTLs are not involved in regulation of transcription/translation
→ psQTLs are likely QTLs of protein degradation
Enrichment of Genomic Annotations

![Graph showing enrichment of genomic annotations](image)

<table>
<thead>
<tr>
<th>Annotation</th>
<th>No. of SNPs</th>
<th>Background</th>
<th>Protein</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exonic</td>
<td>12,568</td>
<td>Intergenic</td>
<td>$2.8 \times 10^{-14}$</td>
<td>$2.3 \times 10^{-21}$</td>
</tr>
<tr>
<td>5’ UTR</td>
<td>6,488</td>
<td>Intergenic</td>
<td>$3.2 \times 10^{-5}$</td>
<td>$5.9 \times 10^{-19}$</td>
</tr>
<tr>
<td>3’ UTR</td>
<td>15,139</td>
<td>Intergenic</td>
<td>$2.0 \times 10^{-6}$</td>
<td>$1.7 \times 10^{-16}$</td>
</tr>
<tr>
<td>Intronic</td>
<td>628,591</td>
<td>Intergenic</td>
<td>$7.1 \times 10^{-3}$</td>
<td>$2.9 \times 10^{-38*}$</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>2,099</td>
<td>Exonic</td>
<td>$5.7 \times 10^{-3}$</td>
<td>$9.7 \times 10^{-2}$</td>
</tr>
<tr>
<td>Ribo SNitch</td>
<td>414</td>
<td>Exonic</td>
<td>$5.2 \times 10^{-2}$</td>
<td>$2.5 \times 10^{-2}$</td>
</tr>
<tr>
<td>Acetylation site</td>
<td>22</td>
<td>Nonsynonymous</td>
<td>$3.2 \times 10^{-2}$</td>
<td>0.62</td>
</tr>
</tbody>
</table>

*Depletion relative to background.
Comments

• Within each data set
  – Data generation + processing + QC checks suggest good quality data

• Three sets of data generated at different time points
  – Age of HapMap LCLs bias gene expression profiles. Yuan Y. et al. (2015). Scientific Reports. 5:7960
  – Inter-individual variation in gene expression levels decrease with freeze-thaw cycles. Caliskan et al. (2014). PLoS One. 9(9):e107166
a  effect size of eQTLs ascertained in the GEUVADIS study stratified by number of transcripts per gene
Comments

• Why eQTLs have significantly reduced effect sizes on protein levels?

• Potential mechanisms to explain how eQTL effects on downstream phenotypes are attenuated or buffered?
Questions?