Genome-wide analysis of DNA methylation in samples from the Genotype-Tissue Expression (GTEx) project

Peter Hickey
@PeteHaitch
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
Single Cell Open Research Endeavour (SCORE),
Walter and Eliza Hall Institute of Medical Research

Slides: www.bit.ly/AGTA2018
GTEx to eGTEx via a ‘pilot’ study
The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study [human] tissue-specific gene expression and regulation.

- GTEx Consortium, 2015, Science 348, 648–660
[eGTEx] extends the GTEx project to combine gene expression with additional intermediate molecular measurements on the same tissues.

Hmm, this eGTEEx study is gonna be huge.
And the human brain is hella cool.
Let’s do a pilot study.
- Artist’s impression of conversation in Hansen and Feinberg labs, c. 2015
BrainEpigenome (the ‘pilot’ study)


UCSC Track Hub: www.bit.ly/BrainEpigenomeHub
Map of human brain methylome was limited (c. 2015)

- Little whole genome bisulfite sequencing (WGBS) data
- Few (if any) biological replicates
- Mostly bulk tissue samples
- Few brain region-specific differentially methylated regions (DMRs)


http://epigenomesportal.ca/ihec/grid.html (Build: 2017-10)
A good map requires biological replicates, multiple brain regions, and multiple cell types.

- **WGBS (bulk)**
  - *n = 27*

- **ATAC-seq (NeuN sorted)**
  - *n = 22*

- **RNA-seq (NeuN sorted)**
  - *n = 20*

**Tissue**
- **BRNCTXB** (frontal cortex)
- **BRNACC** (anterior cingulate cortex)
- **BRNHPP** (hippocampus)
- **BRNNCC** (nucleus accumbens)
Bulk tissue samples are uninformative for brain region-specific mCG due to variation of neuronal proportion in sampled tissue.
Let’s try fluorescence activated nuclei sorting (FANS)

Tissue
BRNCTXB (frontal cortex)
BRNACC (anterior cingulate cortex)
BRNHPP (hippocampus)
BRNNCC (nucleus accumbens)
And let’s do some more assays

- **WGBS (bulk)**
  - NeuN+
  - NeuN-
  - n = 27

- **WGBS (NeuN+, NeuN-)**
  - n = 45

- **ATAC-seq (NeuN+, NeuN-)**
  - n = 22

- **RNA-seq (NeuN+, NeuN-)**
  - n = 20

**Tissue**
- BRNCTXB (frontal cortex)
- BRNACC (anterior cingulate cortex)
- BRNHPP (hippocampus)
- BRNNCC (nucleus accumbens)
FANS & WGBS reveals brain region-specificity of mCG in NeuN+ (but not NeuN-) samples

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeuN+ vs. NeuN-</td>
<td>100,875*</td>
<td>70.0 Mb</td>
</tr>
<tr>
<td>NeuN+</td>
<td>13,074</td>
<td>11.9 Mb</td>
</tr>
<tr>
<td>NeuN-</td>
<td>114</td>
<td>0.1 Mb</td>
</tr>
</tbody>
</table>

*21,802 novel DMRs
NeuN+ samples: mCH shows limited strand specificity, ‘tracks’ mCG, and can be used to identify CH-DMRs

NeuN+ mCH (1kb bins)

<table>
<thead>
<tr>
<th>Context &amp; strand</th>
<th>Region</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: mCA (+)</td>
<td>BRNCTXB</td>
<td></td>
</tr>
<tr>
<td>a: mCA (-)</td>
<td>BRNACC</td>
<td></td>
</tr>
<tr>
<td>T: mCT (+)</td>
<td>BRNHPP</td>
<td></td>
</tr>
<tr>
<td>t: mCT (-)</td>
<td>BRNNCC</td>
<td></td>
</tr>
</tbody>
</table>

CH-DMRs

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeuN+</td>
<td>15,029⁺</td>
<td>39.6 Mb⁺⁺</td>
</tr>
</tbody>
</table>

⁺Before merging across strand and context
⁺⁺After merging across strand and context
Enrichment of DMRs over genomic features

- OCR (union)
- H3K27ac
- FANTOM5
- CH–DMRs (NeuN+)
- DEGs
- CG–DMRs (NeuN+)
- DEG promoters
- intronic
- Shelves
- exonic
- three_utr
- Shores
- promoter
- CGI
- OpenSea
- SINE
- DNA
- Simple_repeat
- Low_complexity
- five_utr
- intergenic
- LTR
- LINE
- Satellite

log2(OR)

Color Key

CG − DMR (NeuN+)

CH − DMR (NeuN+)
Enrichment of DMRs over genomic features

CG-DMRs and CH-DMRs co-occur

log2(OR)

Value

Color Key

OCR (union)
H3K27ac
FANTOM5
CH-DMRs (NeuN+)
DEGs
CG-DMRs (NeuN+)
DEG promoters
intrinsic
Shelves
exonic
three_utr
Shores
promoter
CGI
OpenSea
SINE
DNA
Simple_repeat
Low_complexity
five_utr
intergenic
LTR
LINE
Satellite

CG-DMRs are enhancer-centric
CH-DMRs are enriched over
Enrichment of DMRs over genomic features

CG-DMRs and CH-DMRs co-occur
CG-DMRs are enhancer-centric

log2(OR)
Enrichment of DMRs over genomic features

CG-DMRs and CH-DMRs co-occur
CG-DMRs are enhancer-centric
CH-DMRs are DEG-centric

log2(OR)

Value

Color Key

OCR (union)
H3K27ac
FANTOM5
CH-DMRs (NeuN+)
DEGs
CG-DMRs (NeuN+)
DEG promoters
intronic
Shelves
exonic
three_utr
Shores
promoter
CGI
OpenSea
SINE
DNA
Simple_repeat
Low_complexity
five_utr
intergenic
LTR
LINE
Satellite
CG-DMRs in NeuN+ samples are enriched for GWAS heritability of neuropsychiatric traits

Stratified linkage disequilibrium score regression*

27 ‘brain-linked’ traits (e.g., Schizophrenia, ADHD)

3 ‘non-brain-linked’ traits (e.g., height)

eGTEx (work in-progress)

*eGTEx Project* Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease.

*Nature Genetics* (2017), doi: 10.1038/ng.3969
# eGTEX Study Design

<table>
<thead>
<tr>
<th>Molecular phenotype</th>
<th>Primary assay(s)</th>
<th>Targeted tissues (phase II)</th>
<th>Targeted sample number</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA accessibility</td>
<td>DNase I hypersensitivity</td>
<td>Brain regions, heart, lung, muscle, esophagus, breast, prostate, skin</td>
<td>~1,135</td>
</tr>
<tr>
<td>Histone modifications</td>
<td>ChIP-seq</td>
<td>Brain regions, heart, lung, muscle</td>
<td>~600</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>WGBS and capture bisulfite sequencing</td>
<td>Brain regions, heart, lung, muscle, thyroid</td>
<td>~2,000</td>
</tr>
<tr>
<td>Allele-specific expression</td>
<td>mmPCR-seq</td>
<td>All tissues</td>
<td>~2,000</td>
</tr>
<tr>
<td>Post-transcriptional RNA modifications</td>
<td>m6A methylation capture sequencing</td>
<td>Brain regions, heart, lung, muscle</td>
<td>~300</td>
</tr>
<tr>
<td>Proteomic variation</td>
<td>MS, targeted arrays for transcription factors and cell signaling proteins</td>
<td>Brain, heart, lung, muscle, thyroid, colon, liver, prostate, pancreas, ovary, testis, breast</td>
<td>~1,000 (MS) ~2,500 (arrays)</td>
</tr>
<tr>
<td>Somatic variation</td>
<td>Deep exome sequencing, RNA-seq, SNP arrays</td>
<td>~20–25 tissues</td>
<td>~800</td>
</tr>
<tr>
<td>Telomere length</td>
<td>Luminex-based assay for telomere-repeat abundance</td>
<td>~20 tissues</td>
<td>~5,000</td>
</tr>
</tbody>
</table>

Molecular assays, targeted tissues, and sample number for eGTeX.

---

*eGTEX Project* Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease.  

*Nature Genetics (2017), doi: 10.1038/ng.3969*
Re-wrote *bsseq* to process and analyse eGTEx-sized (and bigger) datasets

- Processed data is too large to store and operate on in-memory (10s – 100s of GB)
  - Data stored on-disk in HDF5 file
- Improved parallelization of key steps
  - Importing files
  - Smoothing
  - DMR calling
  - Permutation testing
- Available through Bioconductor
  - [https://bioconductor.org/packages/bsseq/](https://bioconductor.org/packages/bsseq/)
mCG distinguishes eGTEx samples by tissue

Tissue
- BRAIN
- BREAST
- ESPMSL
- HRTLV
- LUNG
- MSCLSK
- PTTARY
- SKINS
- THYROID

NeuN
- bulk
- pos
eGTEX NeuN+ samples are (mostly) consistent with BrainEpigenome NeuN+ samples.
eGTEX NeuN+ samples are (mostly) consistent with BrainEpigenome NeuN+ samples
eGTEX NeuN+ samples are (mostly) consistent with BrainEpigenome NeuN+ samples
5-group: 16x as many CG-DMRs in eGTEX NeuN+ samples as in BrainEpigenome NeuN+ samples

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-group</td>
<td>181,146</td>
<td>196.9 Mb</td>
</tr>
</tbody>
</table>
Basal ganglia: Discover 2x as many CG-DMRs in eGTEX NeuN+ samples as in BrainEpigenome NeuN+ samples.
Hippocampus: What the hell is going on?

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-group</td>
<td>181,146</td>
<td>196.9 Mb</td>
</tr>
<tr>
<td>Basal ganglia</td>
<td>16,866</td>
<td>24.0 Mb</td>
</tr>
<tr>
<td>Hippocampus</td>
<td>11,702</td>
<td>24.4 Mb</td>
</tr>
</tbody>
</table>
Ongoing eGTEx analyses

- Complete analyses of CG-DMRs
- Identify CH-DMRs and analyse
- Stratified linkage disequilibrium score regression
  - Do BrainEpigenome results replicate?
  - What can brain region-specific DMRs tell us?
- Variably methylated regions (VMRs)
- Allele-specific methylation using phased GTEx genomes
- Use sorted data to deconvolute bulk brain samples
- Integration with other GTEx and eGTEx data
Summary

• BrainEpigenome
  • FANS + WGBS reveals many brain region-specific CG-DMRs and CH-DMRs for NeuN+ (but not NeuN-) samples.
  • Neuronal CG-DMRs are enriched for heritability of several neurological, psychiatric, behavioral-cognitive phenotypes.

• eGTEx
  • More tissues + more replicates = huge increase in DMRs.

• The scale of these projects necessitated extensive improvements to computational methods and software engineering.

• There will still be heaps of analyses on the table after publication of initial eGTEx publication(s).
  • Get involved!
Acknowledgements

Sequencing gurus: Rakel Tryggvadóttir, Adrian Idrizi, Colin Callahan
ATAC-seq experiments: Varenka Rodriguez DiBlasi
Flow sorting: Hao Zhang and Hopkins Flow Facility
Funding: eGTEx (U01MH104393), CFAR (5P30AI094189-04, 1S10OD016315-01, and 1S10RR13777001), AGTA Travel Award
Donors and families: NIH NeuroBioBank at the University of Maryland & University of Pittsburgh

eGTEx Project Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. Nature Genetics (2017), doi: 10.1038/ng.3969

Genome Browser
www.bit.ly/BrainEpigenomeHub

Slides
www.bit.ly/AGTA2018

Software
http://bioconductor.org/packages/bsseq/

@PeteHaitch
Bonus slides
eGTEx capture bisulfite-sequencing study

• Aim: Study genetic influence on DNA methylation in human brain
• Assay: Targeting 46 Mb (1 million CpGs) with Roche NimbleGen capture
  • 55% of CpGs not captured by microarrays or other targeted panels
  • CG-DMRs
    • Neuronal (BrainEpigenome and eGTEx)
    • NeuN+ vs. NeuN- (BrainEpigenome)
    • GABAergic vs. glutamatergic
  • CG-VMRs (eGTEx)
  • Haplotype-dependent allele-specific DMRs and meQTLs
  • Fetal brain meQTLs
  • ‘Epigenetic age’ CpGs
• Samples: > 100 donors (BRNCTXB, BRNCDT, BRNNCC, BRNHPP, and THYROID)

1Dracheva et al., unpublished

