Illumina 450k analysis

HSPH Bioinformatics Core
Oliver Hofmann, John Hutchinson & Winston Hide
Three focus areas

Research computing
Array and Next-gen sequencing
Functional significance
Bioinformatics workflows

A typical 450k workflow
Bioinformatics workflows

From samples to function

Biological samples

Annotation & Enrichment
Bioinformatics workflows

- Biological samples
- Array data
- Quality control
Bioinformatics workflows

Check general intensity distributions
Bioinformatics workflows

Check general intensity distributions
Identify outliers
Bioinformatics workflows

Check general intensity distributions
Identify outliers
Test for sample label problems
Bioinformatics workflows

No background normalization necessary
Filter & Quantile normalization
F-test on CpG probe sites

Biological samples
Array data
Quality control
Normalization
Identify sites
Bisulfite conversion

- Unmethylated Cytosine to Thymine
- Methylated Cytosine protected
Bioinformatics workflows

- Identify biological meaning
- Biological samples
- Array data
- Quality control
- Normalization
- Identify sites
- Annotation & Enrichment
A few roadblocks
Not just one implementation

Different packages and implementations
Data formats

- Proprietary formats
- Different requirements for each package
Biases and normalization

Raw

Normalized

Intensity
Probe differences: Infinium I
Probe differences: Infinium II
Probe differences

Smaller batch effect
...but systematic probe differences
Quantile normalization not ideal
Peak-based correction

Re-scale type II based on type I
Works well for peaks
Works poorly for intermediate methylation
SWAN

Subset-quantile within array normalization

1 CpG

2 CpG

3 CpG
Does it matter?

Infinium 450k Methylation Workshop, April 2012, University College London
Does it matter?

Infinium 450k Methylation Workshop, April 2012, University College London

Although differences are noticable, the top hits are very similar in (nearly) all methods.
Does it matter?

- Workflow differences often below technical variance
Does it matter?

- Highly methylated sites with higher variance
- 10% beta value difference with 99% confidence
Identified differences often <10%

EHP120(10):1425–1431; Joubert et al.
Low signal/noise

Study design, replicates, sample size
Probe annotation

- ~20,000 probes matching multiple locations
- ~45,000 probes overlapping SNP
- ~10,000 probes with wrong annotation
Functional analysis
Clustering signals

Multiple CGI/gene interrogated
Signals usually cluster
Signal-to-gene
Window-based approaches
Enrichment analysis

Functional analysis using Gene Ontology, Pathways or Gene Sets
GREAT: Genomic Regions Enrichment of Annotations Tool

GREAT predicts functions of cis-regulatory regions.

1. **Input:** A set of Genomic Regions (such as transcription factor binding events identified by ChIP-Seq).

2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.

3. GREAT uses gene Annotations from numerous ontologies to associate genomic regions with annotations.

Example: SRF ChIP-Seq called peaks

Gene transcription start site

Genome

Association

Ontology annotation (e.g. "actin binding")

http://bejerano.stanford.edu/great/
Prioritizing candidate genes
Guilt-by-association
Target molecules (seeds)
Target molecule interaction

Marker A binds/inhibits Marker B
Target molecule interaction

Marker A binds/inhibits Marker B
Connecting target molecules with ‘linkers’
Transcriptional regulation

Known interactions
Protein-protein binding
Consistent co-expression
Transcriptional regulation
Integrate with expression data

- Impact of methylation on gene expression
- Identify downregulated genes (direct, indirect)
Extending to next-generation sequencing
RRBS
Mspl site

5' - ........CGGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN5'  
3' - ..........GGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN5'  

Mspl site

5' - ..................CGGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN-3'  
3' - ..................GGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN5'  

Mspl digest

5' - CGGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNC-3'  
3' - CNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGC-5'  

end repair + A-tailing

5' - CGGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAGA-3'  
3' - AGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGC-5'  

adapter ligation (X = adapter sequence)

5' - XXXXXXXXTCGGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAGA XXXXXXXXXX-3'  
3' - XXXXXXXXAGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGC-5'  

RRBS
Illumina sequencing
Sequencing by synthesis
Sequence tag

Mapping to a reference genome
Mapping to a reference genome

---

CGCCGTCCTCAGAATGGAAACCTCGCTTCTCTCTGCCCCACAATGCGCAAGTCAG

Reference genome

CGTCCCTCAGAATGGAAACCTCGCTT

Sequence tag

Mapping to a reference genome
Mapping to a reference genome
Mapping to a reference genome
Test for methylation
RRBS alignment

1. Genomic fragment sequence after bisulfite treatment:

   - ...ccggcatgtttaaacgct...
   - TTGGCATGTTTAAACGTT

2. Read conversion:
   - C-to-T
   - G-to-A

   - TTGGTAIGTTTAAATGT
   - TTAACATATTTAAACATT

3. Align to bisulfite converted genomes:
   - Forward strand C-to-T converted genome:
     - ...ttggtatgtttaaattgtt...
     - ...aacataacaaatattttaca...

   - Forward strand G-to-A converted genome:
     - ...ccacataatatatcact...
     - ...ggattgtatatattttgtga...

4. Determine unique best alignment
genomic fragment
sequence after bisulfite treatment

...ccggcatgtttaaaacgct...

TTGGCATGTTAAACGTT

C-to-T

TTGGTATGTTAAATGTT

(1)

(2)

G-to-A

TTAACATATTAAACATT

(3)

(4)

read conversion
align to bisulfite converted genomes

...ttggtatgtttaaatgtt...
...aacatacaatatttacaa...
forward strand C-to-T converted genome

...ccaacatatttaaacact...
...ggtttgtataatttgtga...
forward strand G-to-A converted genome

determine unique best alignment
genomic fragment sequence after bisulfite treatment

TTGGCATGTGTTAAACGTT

C-to-T

TTGGTATGTGTTAAATGTT

G-to-A

TTAACATATTTAAACACATT

read conversion

align to bisulfite converted genomes

forward strand C-to-T converted genome

forward strand G-to-A converted genome

determine unique best alignment
genomic fragment sequence after bisulfite treatment

...ccggcatgtttaaacgct...

TTGGCATGTAAAACGTT

C-to-T

TTGGTAATTTAAAATGTTT

(1)

G-to-A

TTAACATATTTAAACATT

(2)

read conversion

align to bisulfite converted genomes

...ttggtatgtttaaatgttt...

...aaccatatcaaaatttacaal...

forward strand C-to-T converted genome

...ccaacatatattaaacact...

...ggttgtataatttttgtga...

forward strand G-to-A converted genome

(1) (2) (3) (4)

determine unique best alignment
More roadblocks

- No probe differences
- No normalization issues
Development of aligners

http://www.ebi.ac.uk/hts_mappers/
Whole genome, deep coverage v1

Best practices workflow

GATK Best Practices
Whole genome, deep coverage v1

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v2

1000 Genome Project Best Practices
Whole genome, deep coverage v1

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v2

Warning: the material on this page is considered out of date by the GSA team.

Best Practice Variant Detection with the GATK v3

1000 Genome Project Best Practices
1000 Genome Project Best Practices
Start/end bias
Detecting sequencing errors

~1% error rate
Degrading read quality
Good agreement between 450k/RRBS

80% of calls highly concordant (<20% difference)
Good agreement between 450k/RRBS

Correlation ~0.95 for sites interrogated by both
Size-selection step

MspI fragment length

MspI fragment count

40-220bp
Building research computing infrastructure
Servers

Array - high memory
RRBS - large number of CPUs
Storage & backup

Redundant storage
MISSION

The research computing group enables advanced research by providing leading edge computing services across the FAS.

Our staff maintains expertise in constantly changing computing technologies, while “speaking the language” of the FAS researchers, to help them use computing more effectively. Research Computing technologies sit atop of support and infrastructure provided by the FAS IT organization.
Harvard Helps Build $168M Supercomputing Facility

By AKUA F. ABU, CRIMSON STAFF WRITER

Published: Monday, October 31, 2011

Harvard is taking another stride in revolutionary computing by participating in the development of the Massachusetts Green High Performance Computing Center, a state-of-the-art research computing facility in Holyoke, Massachusetts.

The center is the product of an unprecedented partnership among five of the leading research universities in Massachusetts—Harvard, Boston University, MIT, Northeastern, and the University of Massachusetts system—in conjunction with the state government and private industry.

Scaling beyond Odyssey
Sequencing as a commodity

- IonTorrent, MiSeqs: easy fit for an R01
Yikes: Peter Tonellato (BIDMC): The sequenced genome of every patient in the world = 1 Yottabyte (10 to 24 power). Big data indeed. #Tricon
Take-home messages

Current best practices “good enough”
Most studies require *large* cohorts
RRBS starting to become cost-competitive
Set aside funding for IT
Working with us

Scientific Director: Winston Hide (whide@hsph.harvard.edu)
Associate Director: Oliver Hofmann (ohofmann@hsph.harvard.edu)

http://catalyst.harvard.edu/services/biostatsconsult/
Exome-seq in the news

With more than 200 publications using exome sequencing approaches to study disease-causing variants for everything from lung cancer to autism a major challenge is the data management, analysis and presentation of ever-larger sequencing data sets.

A series of articles from Brad Chapman, one of CHB’s senior bioinformaticians, outlines a semi-automated workflow that combines all required steps in a flexible, parallelized manner and currently runs at the FAS Research Computing environment as part of our Galaxy analysis system.

Keep Reading...

http://compbio.sph.harvard.edu/chb/