Big data in the post-genome era:
What can the human genome sequence do for you?

David Kelley, Harvard University
Joseph Vitti, Harvard University
Casey Gifford, Harvard University
Tonight’s Talks

What are genomes and how does the Human Genome Project help us understand them?

How can we use genomic datasets to study human evolution?

How do many different cell types arise from one individual’s genome?
I. The Human Genome Project

1. What is a genome?

2. Why/how did we sequence the human genome?

3. What did we learn?
What is a genome?
But what IS a genome?
What is a gene?

DNA
What is a gene?

DNA

RNA

Transcription
What is a gene?

DNA

RNA

Protein

Transcription

Translation
Outline

1. The Human Genome Project

   1. What is a genome?

   2. Why/how did we sequence the genome?

   3. What did we learn?
Pre-Genome Genetics

Chromosomal location of a gene

Example gene: CFTR

Chromosomal location: 7 q 3 1 . 2

band # 1
sub-band # 2

region # 3

U.S. National Library of Medicine

Human Genome Project Goals

1. What genes exist?

2. What is their exact sequence?

3. How are they arranged in DNA?
DNA sequencing
DNA sequencing
DNA sequencing
Sequencing the genome
Sequencing the genome
Sequencing the genome
Sequencing the genome

[Image: Whole_genome_shotgun_sequencing_vs_Hierarchical_shotgun_sequencing.png]
Outline

1. The Human Genome Project
   1. What is a genome?
   2. Why/how did we sequence the genome?
   3. What did we learn?
How many genes?

- E. coli: 4,149 genes
- Fruit fly: 14,889 genes
- Chicken: 16,736 genes
- Grape: 30,434 genes
- Human: 22,333 genes

Fewer genes than expected

# Human genes estimate

Gene number ≠ complexity

- E. coli: 4,149
- Fruit fly: 14,889
- Chicken: 16,736
- Grape: 30,434
- Human: 22,333
But humans are complex!
Questions?
Recall the gene

DNA

RNA

Protein

Transcription

Translation
Recall the gene

DNA

RNA

Protein
Gene regulation
Noncoding DNA influences gene expression

2%!?
How do we study gene regulation?
Advances in sequencing technology
Advances in sequencing technology
Illumina sequencing
Advances in sequencing technology

Cost per Raw Megabase of DNA Sequence

Moore's Law

NIH
National Human Genome Research Institute

genome.gov/sequencingcosts

http://www.genome.gov/images/content/cost_per_megabase.jpg
Sequencing to count
Sequencing to count
Sequencing to count

Sequence → Align → Count
Sequencing to count

Sequence   →   Align   →   Count

3  2  1  1
RNA-Seq profiling

Gene

- Lymph
- Adrenal
- Ovary
- Breast
- Adipose
- Prostate
- Colon
- Lung
- Kidney
- Testes
- Brain
- Thyroid
- Heart
- Skeletal_muscle
- Stem_cells
- White_blood_cell
- Liver
Summary

1. The Human Genome Project has provided fascinating insights into how the genome works.

2. The reference sequence and advances in sequencing technology have enabled new experimental approaches that are being applied widely.
Using genomic datasets to understand recent human evolution

Joseph Vitti, Ph.D. Student
Sabeti Lab at Harvard & the Broad Institute
Outline

• Big data made bigger: sequencing populations

• What is evolution by natural selection?
  • Example: lactose tolerance in humans

• How do we find regions of the genome under selection?

• Case study: the gene EDAR
Outline

• Big data made bigger: sequencing populations

• What is evolution by natural selection?
  • Example: lactose tolerance in humans

• How do we find regions of the genome under selection?

• Case study: the gene EDAR
So we sequenced a human genome – what’s next?
Evolutionary genomics
Evolutionary genomics
Outline

• Big data made bigger: sequencing populations

• What is evolution by natural selection?
  • Example: lactose tolerance in humans

• How do we find regions of the genome under selection?

• Case study: the gene EDAR
Evolutionary genomics: lactose tolerance
Evolutionary genomics: lactose tolerance
Evolutionary genomics: lactose tolerance

$LCT$ (gene)

Covers for…

Lactase (enzyme)

Breaks down…

Lactose (sugar)
Evolutionary genomics: lactose tolerance
Evolutionary genomics: lactose tolerance
Evolutionary genomics: lactose tolerance
From candidate studies to genome scans

DNA is packaged into chromosomes
From candidate studies to genome scans

Grossman et al., *Cell* 2013
Outline

- Big data made bigger: sequencing populations

- What is evolution by natural selection?
  - Example: lactose tolerance in humans

- How do we find regions of the genome under selection?

- Case study: the gene EDAR
Questions?
Outline

• Big data made bigger: sequencing populations

• What is evolution by natural selection?
  • Example: lactose tolerance in humans

• How do we find regions of the genome under selection?

• Case study: the gene EDAR
Signatures of Selection: Population Differences

Habitat 1

Habitat 2

somersault1824.com, commons.wikimedia.org
Signatures of Selection: Long Haplotypes

But first: *recombination*
Remember chromosomes?

Grossman et al., Cell 2013
Background: chromosomes recombine

mother

father

child
Background: chromosomes *recombine*
Signatures of Selection: Long Haplotypes

Before Selection

After Selection

Haplotypes
Signatures of Selection: Derived Variants at High Frequencies

But first: **ancestral** and **derived**

```
...AGG\textcolor{red}{A}TCT...  
...AGG\textcolor{green}{C}TCT...
```
Signatures of Selection: Derived Variants at High Frequencies

Before Selection

After Selection

Haplotype
Signatures of Selection

- Population differences
- Long haplotypes
- Derived variants at high frequencies
Outline

• Big data made bigger: sequencing populations

• What is evolution by natural selection?
  • Example: lactose tolerance in humans

• How do we find regions of the genome under selection?

• Case study: the gene EDAR
Start with a genome scan...
Case study: EDAR

Kamberov et al., Cell 2013
Big Data in the Post-Genome Era: What can the human genome sequence do for you?

III. Epigenetic Regulation and Cellular Identity

Casey Gifford, PhD
Meissner Lab
Harvard University
Broad Institute
I. Human Genome.

II. Epigenetics.

III. How can we use epigenetic information?
DNA

- 46 chromosomes
- 6 billions base pairs
- 20,000 protein coding genes

Each human cell contains approximately 6ft of DNA.
The Human Body Maintains Massive Cellular Complexity

There are over 200 cell types in the human body.

http://www.roadmapepigenomics.org/data/image/adult
Every cell in your body contains your complete genome.
Every cell in your body contains your complete genome.

Two major problems to solve:
Every cell in your body contains your complete genome.

Two major problems to solve:

1. How do you fit all that DNA into the cell?
Every cell in your body contains your complete genome.

Two major problems to solve:

1. How do you fit all that DNA into the cell?

2. How do you ensure that only the required genes are transcribed in the correct cells?
Every cell in your body contains your complete genome.

Two major problems to solve:

1. How do you fit all that DNA into the cell?

2. How do you ensure that only the required genes are transcribed in the correct cells?

Answer: Epigenetics!
Every cell in your body contains your complete genome.

Two major problems to solve:

1. How do you fit all that DNA into the cell?

2. How do you ensure that only the required genes are transcribed in the correct cells?

Answer: Epigenetics!
1. Packing DNA into the cell

DNA is wrapped around nucleosomes, which together are referred to as chromatin.
1. Packing DNA into the cell

= approximately 146 base pairs of DNA wrapped around 1 nucleosome
Every cell in your body contains your complete genome.

Two major problems to solve:

1. How do you fit all that DNA into the cell?

2. How do you ensure that only the required genes are transcribed in the correct cells?

Answer: Epigenetics!
2. How do you ensure that only the required genes are transcribed in the correct cells?

Epigenetics: the study of changes in gene expression caused by mechanisms other than changes to the underlying DNA sequence.
Epigenetic Information Provides Another Layer of Regulatory Information
Epigenetic Information Provides Another Layer of Regulatory Information

aaaaaaaawomanwithouthermanisnothingaaaaa
Epigenetic Information Provides Another Layer of Regulatory Information

1. A woman without her man is nothing.
Epigenetic Information Provides Another Layer of Regulatory Information

1. A woman without her man is nothing.

2. A woman. Without her, man is nothing.
2. How do you ensure that only the required genes are transcribed in the correct cells?

Multiple types of modifications can affect gene expression or the production of RNA, depending on the area of the histone protein that the modification is added and the type of modification.
2. How do you ensure that only the required genes are transcribed in the correct cells?
2. How do you ensure that only the required genes are transcribed in the correct cells?

DNA methylation prevents proteins from binding to DNA, which prevents RNA from being transcribed.
2. How do you ensure that only the required genes are transcribed in the correct cells?
2. How do you ensure that only the required genes are transcribed in the correct cells?
2. How do you ensure that only the required genes are transcribed in the correct cells?

Histone modification and DNA methylation patterns contribute to cellular identity by controlling the production of RNA!
2. Regulation of Gene Expression

DNA methylation

Histone Modifications
2. Regulation of Gene Expression

Compaction/Accessibility
2. Regulation of Gene Expression
2. Regulation of Gene Expression
2. Regulation of Gene Expression

Certain modifications can create space, allowing proteins necessary for transcription to access DNA.
2. Regulation of Gene Expression

x RNA

Nucleosome

Closed
2. Regulation of Gene Expression

Histone modifications and DNA can prevent proteins from accessing DNA, resulting in gene silencing.
Epigenetics helps to organize DNA, and controls gene expression, therefore controlling cellular identity.
Epigenetics helps to organize DNA, and controls gene expression, therefore controlling cellular identity.

Hepatocytes = liver
- break down toxic substances

Neurons = brain
- process information

http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg
Epigenetics helps to organize DNA, and controls gene expression, therefore controlling cellular identity.

Hepatocytes = liver
- break down toxic substances

Neurons = brain
- process information
Each Cell Type Exhibits a Unique Epigenetic Landscape

While each cell type contains the same genomic sequence, the epigenetic patterns create a unique identity that allows for cellular function.

http://www.roadmapepigenomics.org/data/image/adult
Questions?
How do we use the human genome sequence to study these mechanisms?
Sequencing-based Approaches for Studying Epigenetics

Histone Modifications

DNA
Sequencing-based Approaches for Studying Epigenetics

Histone Modifications

DNA
Sequencing-based Approaches for Studying Epigenetics

Histone Modifications

Protein-Protein Interactions
Sequencing-based Approaches for Studying Epigenetics

Histone Modifications

DNA

Protein-Protein Interactions

Isolate DNA bound by histones
Sequencing-based Approaches for Studying Epigenetics

DNA methylation

![DNA methylation diagram]

ACGTTCCGAGAC
TGCAAGGCTCTG

Identify mismatches after alignment, which delineates methylated versus unmethylated cytosines.
Sequencing-based Approaches for Studying Epigenetics

DNA methylation

Chemical treatment
DNA methylation

Identify mismatches after alignment, which delineates methylated versus unmethylated cytosines.
Primary Data

DNA
Primary Data

- Count reads that have mismatches to calculate DNA methylation.
Primary Data

- Count reads that have mismatches to calculate DNA methylation.

- Identify regions with many reads to identify regions bound by histones with specific modifications.
How can we use this information for therapeutic application?
How can we use this information for therapeutic application?

.... Cancer
.... Inheritance
.... Disease states
Epigenetic Regulation Is Essential for Cellular Stability and Function

Human embryo

Hepatocytes (liver)
Epigenetic Regulation Is Essential for Cellular Stability and Function

….and Development

Human embryo

Hepatocytes (liver)

development

http://www.hsci.harvard.edu/newsroom/examining-ethics-embryonic-stem-cell-research-conversation-hscis-michael-j-sandel-dphil
http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg
Epigenetic Regulation Is Essential for Embryonic Development

Human embryo

Hepatocytes (liver)

Closed

Open

development

http://www.hsci.harvard.edu/newsroom/examining-ethics-embryonic-stem-cell-research-conversation-hscis-michael-j-sandel-dphil
http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg
Epigenetic Regulation Is Also Essential for Differentiation

Human Embryonic Stem Cells
Epigenetic Regulation Is Also Essential for Differentiation

Human Embryonic Stem Cells → differentiation → Hepatocytes → Blood → Neurons
Stem Cells Can Differentiate to Any Cell Type

Human Embryonic Stem Cells

hESCs also created the opportunity to study development of adult human tissues, and potentially represent an alternative to organ transplant.

Hepatocytes

Blood

Neurons

http://commons.wikimedia.org/wiki/File:Ground_glass_hapatocyes_intermed_mag.jpg
Induced Pluripotent Stem Cells Created the Potential for Patient-Specific Tissues

You!

Your cells

Takahashi et al., 2007.
http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg
Induced Pluripotent Stem Cells Created the Potential for Patient-Specific Tissues

You! + Genes

Your cells

iPSCs

Patient-specific tissues!

http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg
Induced Pluripotent Stem Cells Created the Potential for Patient-Specific Tissues

You!

Your cells

+ Genes

iPSCs

http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg

Induced Pluripotent Stem Cells Created the Potential for Patient-Specific Tissues

You! + Genes → iPSCs → Patient-specific tissues!

Takahashi et al., 2007.
http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg
Induced Pluripotent Stem Cells Created the Potential for Patient-Specific Tissues

You! + Genes → iPSCs

Cell types currently being derived are not functional or stable.

Patient-specific tissues!

http://commons.wikimedia.org/wiki/File:Ground_glass_hepatocytes_intermed_mag.jpg
Can an improved understanding of epigenetic regulation improve derivation of therapeutically relevant, functional cell types?
Can an improved understanding of epigenetic regulation improve derivation of therapeutically relevant, functional cell types?

Yes.
Can an improved understanding of epigenetic regulation improve derivation of therapeutically relevant, functional cell types?

Human Embryonic Stem Cells

Differentiation

Epigenetics

Distinct cell types that represent intermediate stages of differentiation.
Can an improved understanding of epigenetic regulation improve derivation of therapeutically relevant, functional cell types?

Human Embryonic Stem Cells → Differentiation → Epigenetics

Distinct cell types that represent intermediate stages of differentiation.

How do the epigenetic patterns change as you acquire a new identity?
Epigenetic Changes Can Be Predictive

hESCs → Intermediate → Differentiated cell type

The epigenetic state changes prior to the appearance of RNA.
Epigenetic Changes Can Be Predictive

hESCs

Intermediate

Differentiated cell type

closed

open-Primed

open-Active

The epigenetic state changes prior to the appearance of RNA.
Epigenetic Changes Can Be Predictive

Regions that should be transcribed during the next stage of development!
Improving Differentiation

RNA tells you where you are, but the epigenetic state tells you about where you’re going.

By looking for “epigenetic priming” in ESC-derived cell types, can we figure out how which regions are incorrectly primed, preventing acquisition of a functional cellular identity?
Summary

- The genome is packaged in to the cell in a structure called chromatin.
- Epigenetics includes the study of chromatin and DNA methylation, and how they affect the production of RNA.
- Each cell type contains a unique epigenetic landscape, which is required for stability and functionality.
- Epigenetic patterns change during development, and ESC/iPSCs are a great tool for studying these changes.
- Mapping the epigenetic landscape may help us derived therapeutic tissues in the lab.
Lecture Summary

• Sequencing a human genome has transformed biology and medicine, opening up new fields of research

• By looking for population-wide patterns of variation, we can infer how natural selection has shaped our species

• Epigenetic mechanisms of gene regulation explain how the same genome can produce many cell types
Thank you!

*SITN would like to acknowledge the following organizations for their generous support.*

**Harvard Medical School**
Office of Communications and External Relations
Division of Medical Sciences

*The Harvard Graduate School of Arts and Sciences (GSAS)*

*The Harvard Graduate Student Council (GSC)*

*The Harvard Biomedical Graduate Students Organization (BGSO)*

*The Harvard/MIT COOP*