Approaches to Gene Discovery

Bruce R. Korf, MD, PhD
• The Human Genome
• Genetic Variation
• Gene Identification
98.8% promoter
1.2% exon
1.2% intron
1.2% exon
“Junk DNA”
• Structural Repeats
• Transposable Elements
• Non-coding RNAs
Gene Regulation

- TATA promoter
- Exon
- Intron
- Repressor or activator
- Corepressor/activator or ligand
- Regulatory sequence
Transcription

RNA polymerase

5' 3' 3' 5' 3' 5'
direction of transcription

mRNA

mRNA cap

5' - 5' triphosphate bond

7-methylguanine

mRNA

poly A addition downstream of AAUAAA

poly A tail

Friday, March 29, 13
Splicing

splice donor

branch site

splice acceptor

GU
AG
A

5'

exon

intron

3'

U1
U2
U5
U4
U6
GU
A
U2
U5
U4
U6

RNA lariat

splicosome

Friday, March 29, 13
Repeated Sequences

- simple sequence repeat
  ...GCGACACACACACACACACACACAGT...
- variable number tandem repeat
  14-100 base pair repeat unit
- highly repeated sequences at centromeric and subtelomeric regions
- segmental duplications

- retroviral-like element
- transposon-derived repeats
- LINE
  gene
  promoter
- SINE
  gene
  promoter
- LTR
  gag pol
  repeat
  repeat
- 6 - 11 kb
- 100 - 400 bp
- 6 - 8 kb
- 2 - 3 kb

Friday, March 29, 13
# Transposable Genetic Elements

<table>
<thead>
<tr>
<th>Type</th>
<th>Structure</th>
<th>Copy Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINE</td>
<td>promoter gene 1 gene 2</td>
<td>850,000</td>
<td>21</td>
</tr>
<tr>
<td>SINE</td>
<td>promoter gene</td>
<td>1,500,000</td>
<td>13</td>
</tr>
<tr>
<td>Retroviral-like</td>
<td>LTR gag pol LTR</td>
<td>450,000</td>
<td>8</td>
</tr>
<tr>
<td>Transposon</td>
<td>repeat transposase repeat</td>
<td>300,000</td>
<td>3</td>
</tr>
</tbody>
</table>

**Promoter:** 100-400 bp  
**Gene:** 6-8 kb  
**Transposase:** 2-3 kb  
**LTR:** 6-11 kb
LINE “Life Cycle”

Transcription

RNA

Reverse Transcriptase
LINE “Life Cycle”

Transcription

Reverse Transcriptase

RNA

Recombination

RNA → DNA
Alu Sequences

ENCODE Project

Welcome to the ENCODE explorer

Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.

Select a thread to start
ENCODE Findings

- annotated 20,687 protein-encoding genes
- average 6.3 alternatively spliced isoforms per gene
- 8,801 small RNAs; 9,640 long non-coding transcripts
- >80% genome transcribed in some cell type
- >400,000 enhancers and 70,000 promoters
Non-Coding RNAs

<table>
<thead>
<tr>
<th>tRNA</th>
<th>transfer RNA</th>
<th>protein synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
<td>protein synthesis</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
<td>splicing</td>
</tr>
<tr>
<td>snoRNA</td>
<td>small nucleolar RNA</td>
<td>RNA modification</td>
</tr>
<tr>
<td>miRNA</td>
<td>micro RNA</td>
<td>gene regulation</td>
</tr>
<tr>
<td>siRNA</td>
<td>small interfering RNA</td>
<td>viral defense</td>
</tr>
<tr>
<td>IncRNA</td>
<td>long non-coding RNA</td>
<td>gene regulation/unknown</td>
</tr>
</tbody>
</table>
Long Non-Coding RNAs

- antisense
- intergenic
- sense overlapping
- sense intronic
- processed transcript

Pseudogenes

- **original gene**
  - inactivation by mutation
  - gene duplication
    - one copy inactivated by mutation
  - retrotransposition of mRNA to cDNA back into genome

- mRNA
- reverse transcriptase
- cDNA
MicroRNA

- Drosha
- Dicer
- RISC
- mRNA
- degraded mRNA

nucleus
Genetic Variation

- Nucleotide change
- Short sequence repeat
- Variable number tandem repeat
- Intrageneric deletion/duplication
- Genomic deletion/duplication
- Chromosome abnormality

Number of nucleotides:
- 1
- 10
- $10^2$
- $10^3$
- $10^4$
- $10^5$
- $10^6$
- $10^7$
- $10^8$
Point Mutations

TCC CAA ATC GTC CCT CGA GTT
ser  gln  ile  val  pro  arg  val
wild type sequence

TCC CAG ATC GTC CCT CGA GTT
ser  gln  ile  val  pro  arg  val
silent mutation

TCC CAA ATC CTC CCT CGA GTT
ser  gln  ile  leu  pro  arg  val
conservative mutation

TCC CAA ATC GTC GCT CGA GTT
ser  gln  ile  val  ala  arg  val
non-conservative mutation

TCC CAA ATC GTC CCT TGA GTT
ser  gln  ile  val  pro  stop
stop mutation

TCC CAG AAT CGT CCC TCG AGT T
ser  gln  asn  arg  pro  ser  ser
frameshift mutation
**Indel**

**normal sequence**
ATA-AAA-ACG-AAA-CTG-TGT-CAA-TTA-GTT-
Ile Lys Thr Lys Leu Cys Gln Leu Val

**mutant sequence**
ATA-AAA-ACG-TTC-AAT-TAG-TT-
Ile Lys Thr Phe Asn Stop

AAA-CTG-TG

delete

insert

T
Splicing Mutations

- Normal splicing
  - Splice acceptor mutation
  - Splice donor mutation
  - Exon splice enhancer mutation
  - Exon skip mutations

- Cryptic splice acceptor mutations
  - Cryptic splice acceptor
  - Truncated exon
  - Mutation creates new splice acceptor

- Inclusion of intron in processed mRNA

Friday, March 29, 13
Triplet Repeat Expansions

- CGG
- GAA
- CAG
- CTG

Diseases:
- Fragile X syndrome
- Friedreich ataxia
- Spinocerebellar ataxias
  - Huntington disease
  - DRP atrophy
  - Spinal & bulbar atrophy
  - Machado-Joseph disease
- Myotonic dystrophy
**Multiexon Deletion**

- **Becker** dystrophy: small deletion juxtaposing out-of-frame exons.
  - Numbers indicate 1st, 2nd, or 3rd position in a triplet codon.
  - Normal dystrophin transcript.
  - Normal dystrophin protein.
- **Duchenne** dystrophy: large deletion juxtaposing in-frame exons.
  - Normal dystrophin transcript.
  - Truncated dystrophin protein.
  - Shortened dystrophin protein.

Dystrophin gene – letters indicate different exons.
Chromosome Microdeletion

deletion

Velo-cardio-facial, Smith-Magenis, Charcot-Marie-Tooth, Disease.
LCR Mispairing

mispairing of LCR1 and LCR2 & recombination

duplication

deletion
DNA Repair

Nucleotide Excision Repair

Base Excision Repair

Mismatch Repair
Frequency of Mutation

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of de novo germline base substitution mutations to be approximately $10^{-8}$ per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

If there are $10^8$ sperm per ejaculate, in principle every base could be mutated in at least one sperm cell and each germ cell has around 10 mutations.
Human Mendelian Phenotypes

OMIM Entry Statistics:

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Autosomal</th>
<th>X Linked</th>
<th>Y Linked</th>
<th>Mitochondrial</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Gene description</td>
<td>13,041</td>
<td>640</td>
<td>48</td>
<td>35</td>
<td>13,764</td>
</tr>
<tr>
<td>* Gene and phenotype, combined</td>
<td>161</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>169</td>
</tr>
<tr>
<td># Phenotype description, molecular basis known</td>
<td>3,064</td>
<td>258</td>
<td>4</td>
<td>28</td>
<td>3,354</td>
</tr>
<tr>
<td>% Phenotype description or locus, molecular basis unknown</td>
<td>1,654</td>
<td>136</td>
<td>5</td>
<td>0</td>
<td>1,795</td>
</tr>
<tr>
<td>Other, mainly phenotypes with suspected mendelian basis</td>
<td>1,799</td>
<td>129</td>
<td>2</td>
<td>0</td>
<td>1,930</td>
</tr>
<tr>
<td>Totals</td>
<td>19,719</td>
<td>1,169</td>
<td>59</td>
<td>65</td>
<td>21,012</td>
</tr>
</tbody>
</table>

http://www.genome.gov/Pages/News/PaceofDiseaseGeneDiscovery.pdf
Approach to Genetic Disorders
Approach to Genetic Disorders

- Phenotype
- Genes
- Environment
- Physiology

Diagnostics
Therapeutics
Genetic Linkage

haplotype
Polymorphism: occurrence of at least two alleles at a locus having a frequency of at least 1%

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNTR</td>
<td>14-100 bp repeat unit with variable number of repeats</td>
</tr>
<tr>
<td>STR</td>
<td>di, tri, tetranucleotide repeats</td>
</tr>
<tr>
<td>SNP</td>
<td>Single base change</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy number variation</td>
</tr>
</tbody>
</table>
Haplotype Analysis

marker 1  4  3
marker 2  7  6
marker 3  3  3
marker 4  5  8
marker 5  9  7
marker 6  2  1
marker 7  8  3
marker 8  5  2
marker 9  3  4

marker 1  4  3
marker 2  5  6
marker 3  1  3
marker 4  7  8
marker 5  9  7
marker 6  4  1
marker 7  5  1
marker 8  4  3
marker 9  7  4
Linkage Disequilibrium

http://estrip.org/articles/read/tinypliny/44920
Positional Cloning
Genome Browser
Candidate Genes

Massively Parallel Sequencing

Diagram showing the process of adding bases sequentially.

Legend:
- A
- T
- G
- C

Steps:
1. One base added
2. Next base added
3. Next base added

Genetic codes:
- GTG
- CCC
- AUG
- TCA
- ACC
- AGC
Exome vs. Genome Sequencing

Genome

Exome

Genome

Exome
Gene Discovery

Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

<table>
<thead>
<tr>
<th>Filter</th>
<th>Kindred 1-A</th>
<th>Kindred 1-B</th>
<th>Kindred 1 (A+B)</th>
<th>Kindreds 1+2</th>
<th>Kindreds 1+2+3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dominant</td>
<td>Recessive</td>
<td>Dominant</td>
<td>Recessive</td>
<td>Dominant</td>
</tr>
<tr>
<td>NS/SS/I</td>
<td>4,670</td>
<td>2,863</td>
<td>4,687</td>
<td>2,859</td>
<td>3,940</td>
</tr>
<tr>
<td></td>
<td>2,362</td>
<td>1,810</td>
<td>3,099</td>
<td>1,810</td>
<td>2,654</td>
</tr>
<tr>
<td>Not in dbSNP129</td>
<td>641</td>
<td>102</td>
<td>647</td>
<td>114</td>
<td>369</td>
</tr>
<tr>
<td></td>
<td>369</td>
<td>53</td>
<td>105</td>
<td>25</td>
<td>63</td>
</tr>
<tr>
<td>Not in HapMap B</td>
<td>898</td>
<td>123</td>
<td>923</td>
<td>128</td>
<td>506</td>
</tr>
<tr>
<td></td>
<td>117</td>
<td>7</td>
<td>38</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Not in either</td>
<td>456</td>
<td>31</td>
<td>464</td>
<td>33</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>1*</td>
<td>8</td>
<td>1*</td>
<td></td>
</tr>
<tr>
<td>Predicted damaging</td>
<td>204</td>
<td>6</td>
<td>204</td>
<td>12</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Ng, S., et al. Nature Genetics 2010;42:30
Cytogenomics

Reference DNA
Test DNA

Mix
Block repeated sequences
Hybridize

Microarray with oligonucleotides
Mendelian Disorders Sequencing Centers

Program Rationale
Discovering the genes and genetic variants underlying human Mendelian disorders is of significant biomedical relevance. The knowledge of these variants, which are rare and highly penetrant, will facilitate rapid and accurate diagnosis of Mendelian disorders and might lead to new therapeutic approaches. This knowledge can also lead to insight about the common or more complex phenotypes that involve similar genes, pathways, and phenotypes. In the long run, a comprehensive collection of rare and highly penetrant variants would be a highly valuable resource for understanding basic human genetics and would identify entry points into fundamental developmental and physiological pathways.

While the genetic basis of more than 3000 Mendelian disorders has been determined so far, the genetic basis remains to be determined for a larger number of confirmed or suspected Mendelian disorders. Recent advances in genome technology and computational methods have made it possible to identify the genetic basis of Mendelian disorders using genome-wide approaches in a more rapid and cost-effective way than linkage mapping and candidate gene approaches.

The Mendelian Disorders Genome Centers Program aims to contribute to the discovery of the genetic basis of most Mendelian disorders in two main ways. The first is to use genome-wide sequencing and other genomic approaches to discover the genetic basis underlying as many disorders and health-related traits as possible, spanning the various Mendelian inheritance patterns, during the funding period. The second is to build a better foundation for elucidating the genetic basis of Mendelian disorders by 1) establishing and disseminating information about effective approaches to the identification of the causative genetic variants, and gaining insight about the overall tractability of Mendelian disorders to state-of-the-art genomic approaches, and 3) compiling a comprehensive list of available human samples of Mendelian disorders and other health-related Mendelian traits as a public resource to help coordinate genetic variant discovery activities that will be carried out by many groups.

Grantees of the Program
The currently funded centers are:

- University of Washington Center for Mendelian Genomics
- Yale Center for Mendelian Disorders
- Baylor-Johns Hopkins Center for Mendelian Genetics

In addition to these centers, the Genome Sequencing and Analysis Centers also carry out efforts to discover the genetic basis of Mendelian disorders (see above).

Program Contacts
For general inquiries about the program, please contact:
Lu Wang, Ph.D.
Program Director
E-mail: wanglu@nih.gov

If you wish to provide samples with confirmed or suspected Mendelian disorders or traits for the Mendelian Disorders Genome Centers to study, please contact the Coordination Site of the Program at mendelian@uw.edu. The Program will decide on the feasibility and priority of sequencing these samples.
Diagnostic Odyssey

Clinical problem → Try again

Interpretation → Genetic testing

Differential diagnosis → Clinical problem
Genomic Diagnosis

NTSE Mutations and Arterial Calcifications
Cynthia St. Hilaire, Ph.D., Shira G. Ziegler, B.A., Thomas C. Markello, M.D., Ph.D., Alfredo Brusco, Ph.D., Catherine Groden, M.S., Fred Gill, M.D., Hannah Carlson-Donohoe, B.A., Robert J. Lederman, M.D., Marcus Y. Chen, M.D., Dan Yang, M.D., Ph.D., Michael P. Siegenthaler, M.D., Carlo Arduino, M.D., Cecilia Mancini, M.Sc., Bernard Freudenthal, M.D., Horia C. Stanescu, M.D., Anselm A. Zdebik, M.D., Ph.D., R. Krishna Chaganti, M.D., Robert L. Nussbaum, M.D., Robert Kleta, M.D., Ph.D., William A. Gahl, M.D., Ph.D., and Manfred Boehm, M.D.

Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease
Elizabeth A. Worhey, PhD,2,3 Alan N. Mayer, MD, PhD2,3, Grant D. Spiverson, MD2,3, Daniel Hellings, BSc1, Benedetta B. Bonacci, MSc2, Brennan Docker, BSc1, Jaime M. Serpe, BSc2, Trivikram Dance, PhD2, Michael R. Tynhannen, BSc2, Regan L. Veith, MSc1, Monica J. Basehore, PhD2,3, Ulrich Brown, MD, PhD2,3, Amy Tomita-Mitchell, PhD2,3, Marjorie J. Arca, MD2,3, James T. Casper, MD2,3, David A. Margolis, MD2,3, David P. Bick, MD2,3, Martin J. Heenan, PhD2,3, John M. Routes, MD2,3, James W. Verhelst, MD, PhD2,3, Howard J. Jacob, PhD2,3, and David P. Dinneen, MD2,3

Genetics in Medicine • Volume 13, Number 3, March 2011