Functional genomics and transcriptomics

Identify and annotate the complete set of genes encoded within a genome

Determine the entire DNA sequence of an organism

Characterize the gene-expression profiles in different tissues and cell types on a genome-wide scale

Ascertain the function of each and every gene product

Understand the genetic basis of the phenotypic differences between individuals and species
Overview

1. Methods for transcriptome analysis

2. Genome transcription landscape
   - Types of RNAs and functions
   - Analysis of gene structure and alternative splicing

3. Transcriptome profiling (tissues, individuals)

4. Search of regulatory regions

5. Gene expression and evolution

Central dogma of molecular biology
Central dogma of molecular biology

**Genome**
Complete DNA content of an organism with all its genes and regulatory sequences

**Transcription**

**Proteome**
Complete collection of proteins and their relative levels in each cell

**Translation**

**Transcriptome**
Complete set of transcripts and relative levels of expression in a particular cell or tissue under defined conditions at a given time

**Phenotype**

Why study of RNA is so important?

RNA profiling provides clues to:

- Expressed sequences and genes of a genome
- Gene regulation and regulatory sequences
- Function and interaction between genes
- Functional differences between tissues and cell types
- Identification of candidate genes for any given process or disease
Methods for transcriptome analysis

One or few genes
- Northern
- RT-PCR
- 5’ and 3’ RACE
- Quantitative RT-PCR (Real-Time PCR)

Whole transcriptome
- EST sequencing
- SAGE
- Microarrays
- RNA-Seq

Expressed sequenced tags (ESTs)

Arrays for transcriptome analysis

- Gene expression arrays
  - Quantification of transcript abundance
  - Single/multiple 3’ probes

- Genome tiling arrays
  - Identification of transcribed sequences
  - Multiple probes along the genome

Characterization of a novel testis transcript using tiling arrays

RNA-seq

- RNA-tag sequencing
  - Quantification of transcript abundance
  - Single end reads of each RNA species

- Whole RNA sequencing
  - Identification of transcribed sequences
  - Multiple reads along each RNA species
Transcriptome analysis by RNA-seq

<table>
<thead>
<tr>
<th>Technology</th>
<th>Tiling microarray</th>
<th>cDNA or EST sequencing</th>
<th>RNA-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle</td>
<td>Hybridization</td>
<td>Sanger sequencing</td>
<td>High-throughput sequencing</td>
</tr>
<tr>
<td>Resolution</td>
<td>From several to 100 bp</td>
<td>Single-base</td>
<td>Single-base</td>
</tr>
<tr>
<td>Throughput</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Reliance on genomic sequence</td>
<td>Yes</td>
<td>No</td>
<td>In some cases</td>
</tr>
<tr>
<td>Background noise</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Application</td>
<td>Simultaneously map transcribed regions and gene expression</td>
<td>Yes</td>
<td>Limited for gene expression</td>
</tr>
<tr>
<td>Dynamic range to quantify gene expression level</td>
<td>Up to a few hundred fold</td>
<td>Not practical</td>
<td>&gt;10,000 fold</td>
</tr>
<tr>
<td>Ability to distinguish different isoforms</td>
<td>Limited</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ability to distinguish allelic expression</td>
<td>Limited</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Practical issues</td>
<td>Required amount of RNA</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Cost for mapping transcriptomes of large genomes</td>
<td>High</td>
<td>High</td>
<td>Relatively low</td>
</tr>
</tbody>
</table>

Table 1. Wang et al. (2009) Nature Reviews Genetics 10: 57-63

The ENCODE project

The ENCODE project was a major research initiative that aimed to identify all functional elements in the human genome. It was divided into two phases:

- **Pilot phase (2003-2007)**: Identification of all functional elements in 1% of the human genome.
- **Production phase (2007-?)**: Identification of all functional elements in the whole human genome.

You can find more information about ENCODE at:

- [http://www.genome.gov/encode/](http://www.genome.gov/encode/)
- [http://genome.ucsc.edu/ENCODE/](http://genome.ucsc.edu/ENCODE/)
The ENCODE project

International collaboration funded by the NHGRI to build a list of functional elements in the human genome, including those acting at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Main results from ENCODE project

- Pervasive transcription of the human genome
  (14.7% of the analyzed bases are transcribed in at least one tissue sample, but as much as 93% could be present in primary transcripts)

- Identification of many novel non-protein transcripts
  (ncRNAs either overlapping protein-coding loci or in regions thought to be transcriptionally silent)

- Numerous unrecognized transcription start sites

- Great complexity and overlap of transcribed regions
  (multiple overlapping transcripts of different sizes from the same or opposite DNA strands)
The modENCODE project

**Model organism ENCyclopedia Of DNA Elements**
- Identification of functional elements in *D. melanogaster* & *C. elegans*
- Possibility of validation using methods that cannot be applied in humans

[http://www.modencode.org/](http://www.modencode.org/)

---

**Pervasive human transcription**

As much as 93% of DNA could be present in primary transcripts

(Kapranov et al., *Science* 2002)
Most transcribed nucleotides are outside annotated exons!

Genome transcription relevance?

1. Transcriptional noise?

2. Functional?
Types of RNAs

Total RNA

Coding RNA 4% of total

Functional RNA 96% of total (Non-coding)

Pre-mRNA (hnRNA)

Pre-rRNA Pre-tRNA snRNA snoRNA miRNA siRNA piRNA

Structural RNAs

Regulatory RNAs

KEY

All organisms
Eukaryotes only

Figure 1.12: Genomes 3 (© Garland Science 2007)
### Types of RNAs and characteristics

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Size</th>
<th>Number</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
<td>Several kb</td>
<td>~30,000</td>
<td>Coding for proteins</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNAs</td>
<td>114-5000 nt</td>
<td>~300-400</td>
<td>Component of ribosome</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNAs</td>
<td>74-95 nt</td>
<td>~1000</td>
<td>Translation</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNAs</td>
<td>100-300 nt</td>
<td>?</td>
<td>Splicing</td>
</tr>
<tr>
<td>snoRNA</td>
<td>small nucleolar RNAs</td>
<td>60-300 nt</td>
<td>~375</td>
<td>RNA modification</td>
</tr>
<tr>
<td>miRNA</td>
<td>micro RNAs</td>
<td>21-23 nt</td>
<td>~1000</td>
<td>Gene expression regulation</td>
</tr>
<tr>
<td>piRNA (rasi RNA)</td>
<td>piwi RNAs</td>
<td>26-30 nt</td>
<td>~33000</td>
<td>Gametogenesis, transposon defense</td>
</tr>
<tr>
<td>IncRNA</td>
<td>Long non-coding RNAs</td>
<td>&gt;200 bp</td>
<td>~35000</td>
<td>Regulation, imprinting</td>
</tr>
</tbody>
</table>

### Messenger RNAs (mRNAs)

![Diagram of messenger RNAs (mRNAs)]
Micro RNAs (miRNAs)

- Small non-coding RNAs (21-23 nt) involved in post-transcriptional regulation of gene expression by binding to the 3' UTR of target mRNAs
- Identified in the early 1990s, but recognized as a distinct class of regulators in the early 2000s
- Abundant in many cell types and may be involved in different developmental processes (complex organisms)
- Target around 60% of mammalian genes and each miRNA can repress hundreds of genes

Other new non-coding RNAs

- From the same family of non-coding RNAs, some have been identified in humans (miR-200 family) by different groups
- miR-200 family has been shown to play a role in the regulation of cell proliferation and differentiation
- miR-200 family has been shown to be downregulated in various cancers, such as breast cancer and sarcoma
- miR-200 family has been shown to be upregulated in other diseases, such as diabetes and obesity
- miR-200 family has been shown to be expressed in different tissues, such as brain, heart, and liver
- miR-200 family has been shown to be regulated by various factors, such as hormones and environmental cues
- miR-200 family has been shown to be involved in the regulation of various pathways, such as Wnt, TGF-β, and Notch pathways

Figure 2. He and Hannon (2004) Nature Reviews Genetics 5: 522-531.

Other new non-coding RNAs

- From the same family of non-coding RNAs, some have been identified in humans (miR-200 family) by different groups
- miR-200 family has been shown to play a role in the regulation of cell proliferation and differentiation
- miR-200 family has been shown to be downregulated in various cancers, such as breast cancer and sarcoma
- miR-200 family has been shown to be upregulated in other diseases, such as diabetes and obesity
- miR-200 family has been shown to be expressed in different tissues, such as brain, heart, and liver
- miR-200 family has been shown to be regulated by various factors, such as hormones and environmental cues
- miR-200 family has been shown to be involved in the regulation of various pathways, such as Wnt, TGF-β, and Notch pathways

Figure 2. He and Hannon (2004) Nature Reviews Genetics 5: 522-531.
Transcriptional complexity

(61%-72% genes)

Long-range transcription could be more frequent than previously thought and involve both exons located very far away (up to Mb) and exons in different chromosomes

Figure 2] Fusion transcripts combining exons of different genes and unannotated regions. Two different transcripts combine novel 5' exons with selected exons of cancer1 (CAV1) and cancer2 (CAV2). The exons of the two fusion transcripts (GenBank accession numbers EF179190 and EF179190) and CAV1 and CAV2 mRNAs are shown as vertical bars. Introns are represented as horizontal lines; slanted lines indicate a gap of -295 kb, used to simplify the depiction of this genomic region. The coordinates are taken from the hg18 NCBI35 version of the genome.

(Gingeras et al., Nat. Rev. Genet. 2007)
Transcription start site analysis

Global mapping of CAGE (Cap Analysis of Gene Expression) tags from mRNA ends identifies multiple transcription start sites (TSS):

CAGE tags can be mapped to different positions in a ~100 kb upstream region of the transcribed gene. This analysis helps identify different transcriptional start sites. (From Carninci et al., 2006)

Identification of new TSS

Mapping of CAGE-tags to the human genome

(Caminci et al., Nat. Genet. 2006)
Alternative splicing

Alternative splicing affords extensive proteomic and regulatory diversity from a limited repertoire of genes

Figure 1. Nielsen and Graveley (2010) Nature 463: 457-463

Figure 1. Li et al. (2007) Nature Reviews Neuroscience 8: 819-831

Large-scale analysis of alt. splicing

Alternative isoform regulation in human tissue transcriptomes

Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing

Nature Genetics 40 | NUMBER 12 | DECEMBER 2008
Alternative splicing in human tissues

- 92-98% of multi exon genes are subjected to alternative splicing (almost 86% of human genes)
- ~100,000 AS events of different frequency detected in human tissues and many show variation between them
- Most alternative transcripts are variable between different tissues as a result of specific regulation
- Alternative splicing could be a key mechanism in generating proteomic diversity

Figure 2. Wang et al. (2008) Nature 456: 470-476.

Alternative splicing in α-tropomyosin

Figure 8.22. Evolution. Barton et al. (2007) Cold Spring Harbor Laboratory Press.
Unanswered questions

- How many of the observed isoforms are functionally relevant?
- Can alternative slicing account for higher complexity in organisms?

### Table 2 | Examples of functionally relevant alternative splicing events

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drsper1</td>
<td>Alternative splicing has the potential to generate 152000 alternative isoforms that show tissue-specific binding, regulating protein expression and immune system interactions with pathogens</td>
<td>76</td>
</tr>
<tr>
<td>Pi</td>
<td>Splicing-specific transcript splicing behavior</td>
<td>77</td>
</tr>
<tr>
<td>Mut571</td>
<td>Alternative splicing generates three protein isoforms, each of which assembles into a functionally distinct form of the complex involved in remodeling complex</td>
<td>78</td>
</tr>
<tr>
<td>Ganciclovir oligos</td>
<td>Mutually exclusive splicing generates two isoforms of a growth factor with distinct ligand specificities</td>
<td>79</td>
</tr>
<tr>
<td>Pk1</td>
<td>Alternative splicing generates two isoforms of a protein that has distinct but complementary roles in tumor suppression and organization</td>
<td>80</td>
</tr>
<tr>
<td>unq-6</td>
<td>Alternative splicing generates two isoforms of a protein that have distinct functional activities</td>
<td>81</td>
</tr>
<tr>
<td>Perseus</td>
<td>Alternative splicing modulates the interactions between the neuronal synaptic cell adhesion molecule 1 and receptors</td>
<td>82</td>
</tr>
<tr>
<td>SLC25</td>
<td>Alternative splicing alters the amino-terminal domain of SLC25</td>
<td>82</td>
</tr>
<tr>
<td>ucs-25</td>
<td>Alternative splicing generates two isoforms that differ in their ability to stabilize synaptic vesicles in the primed state</td>
<td>84</td>
</tr>
</tbody>
</table>


---

### A new gene concept

**Classic definition**

A unit of hereditary material that codifies a protein (or polypeptide chain)

**Novel definition**

(wikipedia)

A hereditary unit consisting of a sequence of DNA or RNA that occupies a specific location in a genome and is associated with regulatory regions, transcribed regions, and/or other functional sequence regions
Transcript profiling across tissues

- Relationship between tissues
- Functional differences between tissues
- Identification of tissue specific promoters
- Co-regulation and functional relationship of gene products (regulation networks and functional pathways)

Zapala et al. 2005

Transcript profiling in cancer

Molecular classification of cancer types

ALL – acute lymphoblastic leukemia
AML – acute myeloid leukemia

Transcriptome reflects Biology

- Predictive
- Treatment
Transcript profiling across individuals

- Characterization of expression levels in lymphoblastoid cell lines from individuals of hapMap populations
- Many genes with variation across individuals
- Gene expression phenotypes are highly heritable
- Higher proportion of expression differences related to cis-variants (although they are also easier to detect)
Mapping gene expression variants

(a) 3' (exon)

(b) 3' (intron)

(bigger effects, few genes)

(moderate effects, more genes)

Table 2: Gene associations in single populations and multipopulation subsets

<table>
<thead>
<tr>
<th>Population</th>
<th>Chr</th>
<th>Gene</th>
<th>p-value</th>
<th>rsID</th>
<th>Effect</th>
<th>Haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTT</td>
<td>1</td>
<td>A</td>
<td>5.6e-4</td>
<td>123</td>
<td>Effect</td>
<td>Haplotype</td>
</tr>
<tr>
<td>GTT</td>
<td>2</td>
<td>B</td>
<td>1.2e-5</td>
<td>456</td>
<td>Effect</td>
<td>Haplotype</td>
</tr>
<tr>
<td>GTT</td>
<td>3</td>
<td>C</td>
<td>3.4e-6</td>
<td>789</td>
<td>Effect</td>
<td>Haplotype</td>
</tr>
</tbody>
</table>

Table 3: Rare allele associations in single-population and multipopulation analysis

<table>
<thead>
<tr>
<th>Population</th>
<th>Chr</th>
<th>Gene</th>
<th>p-value</th>
<th>rsID</th>
<th>Effect</th>
<th>Haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTT</td>
<td>1</td>
<td>A</td>
<td>2.1e-3</td>
<td>123</td>
<td>Effect</td>
<td>Haplotype</td>
</tr>
<tr>
<td>GTT</td>
<td>2</td>
<td>B</td>
<td>3.2e-4</td>
<td>456</td>
<td>Effect</td>
<td>Haplotype</td>
</tr>
<tr>
<td>GTT</td>
<td>3</td>
<td>C</td>
<td>4.3e-5</td>
<td>789</td>
<td>Effect</td>
<td>Haplotype</td>
</tr>
</tbody>
</table>

Gene-level QTL (TSP50)

rs7639979: GG

(rs = 18)

GA (35)

AA (16)

Ensembl gene model

Position (Mb):

46,730 - 46,734
Persistence of lactase expression

In most mammals, the ability to digest milk disappears with age and is related to the production of the lactase enzyme:

Lactase production in adults shows large variability in human populations and seems related with pastoralism.
Persistence of lactase expression

<table>
<thead>
<tr>
<th>SNP</th>
<th>Position</th>
<th>Allele changes</th>
<th>Evidence of association with lactase persistence</th>
<th>Evidence of function</th>
<th>Regulator (Kedr et al. 2003 nonconsensus)</th>
<th>Geographic location of highest observed frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>14810</td>
<td>C &gt; G</td>
<td>Not included in dbSNP</td>
<td>-</td>
<td>Toshkoff et al. (2007)</td>
<td>B</td>
<td>Kenya/Tanzania</td>
</tr>
<tr>
<td>13585</td>
<td>T &gt; G</td>
<td>041980547</td>
<td>-</td>
<td>Toshkoff et al. (2007), Toshkoff et al. (2007), Toshkoff et al. (2007)</td>
<td>C</td>
<td>South Africa</td>
</tr>
<tr>
<td>13910</td>
<td>C &gt; T</td>
<td>e4892235</td>
<td>Toshkoff et al. (2007), O'Sullivan and Sibley (2008)</td>
<td></td>
<td>A</td>
<td>Europe</td>
</tr>
<tr>
<td>13907</td>
<td>C &gt; G</td>
<td>e0277574</td>
<td>Toshkoff et al. (2007), Toshkoff et al. (2007)</td>
<td></td>
<td>A</td>
<td>Ethiopia/Rwanda</td>
</tr>
</tbody>
</table>

Note: we and others have identified a total of six other alleles (including -13910C>T) within the 159 bp region -15,096 to -14,938 for which studies of their association and function are ongoing.

Identification of regulatory elements

Gene regulatory regions very difficult to predict:
- Small (<50 bp)
- Variable sequence motifs
- Few really important positions
- Poorly conserved & not defined location

Regulatory elements:
- Core promoter
- Proximal elements
- Distal enhancers (upstream/downstream)
ChIP-chip or ChIP-seq

Chromatin Immunoprecipitation (ChIP)
- Transcription factor binding sites
- DNA methylation
- Histone modifications

Regulatory data from ENCODE

Extraordinary resource to predict enhancers!
Regulatory data from ENCODE

Figure 1. Ernst (2011) Nature 473: 43–49

Human enhancers maps
Where does this info could be found?

http://genome.ucsc.edu

http://www.ensembl.org/index.html