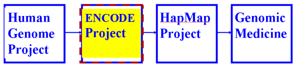
Computational Methods in Genomics
PART TWO

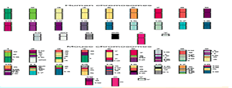
Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

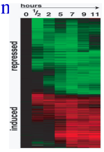
Outline

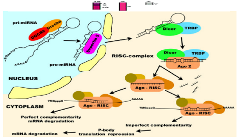
- ENCODE Project
- GENCODE Project
- Micro RNA (MiRNA)
- Genome Rearrangement (2008)
  - Reversals
  - Translocations
  - Fusions
  - Fission
- Microarrays
  - Spotted
  - Affymetrix


Human Genome Project → ENCODE Project → HapMap Project → Genomic Medicine

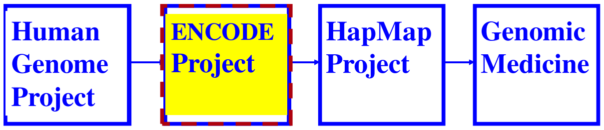
Sequencing of the human DNA | Interpreting the human genome sequence | Implicating genetic variants with human disease | Personalized medicine Cure for diseases







Pathway to Genomic Medicine



Sequencing of the human DNA

Interpreting the human genome sequence

Implicating genetic variants with human disease

Personalized medicine
Cure for diseases

ENCODE

ENCODE: **ENC**yclopedia **O**f **D**NA **E**lements

- **Goal**: compile a comprehensive encyclopedia of all functional elements in the human genome
- Initial Pilot Project: 1% of human genome
- Apply multiple, diverse approaches to study and analyze that 1% in a consortium fashion

The ENCODE Project

- 44 regions of the human genome were selected, spanning 30 megabases (about 1% of the human genome).
- The **ENCODE regions** include:
  - About 50% randomly selected loci
  - About 50% containing well-known genes
    - Example: alpha and beta globins, CFTR
- The **ENCODE Project Consortium** released its findings in a 2007 article (>250 coauthors)

Experiment Redundancy

- The ENCODE pilot project aimed to establish redundancy with respect to the findings represented by different data sets:
  - Multiple experiments based on a similar technique:
    - e.g. study transcriptional activity in different tissues using the same technology.
  - Multiple experiments based on different techniques.
- Such redundancy has allowed methods to be compared and consensus data sets to be generated.

Major Findings of ENCODE

- The majority of all nucleotides are **transcribed** as part of
  - Coding transcripts
  - Noncoding RNAs
  - Random transcripts that may have no biological function.
- Many genes have multiple, previously undetected, **transcription start sites**
  - Regulatory sequences are as likely to be upstream as downstream of the major start sites.

Highlights of ENCODE Project (I)

- The human genome is pervasively transcribed, such that the majority of its bases are associated with at least one primary transcript.
- Many novel non-protein-coding transcripts have been identified:
  - many non-protein-coding genes overlap with protein-coding loci
  - others are located in regions of the genome previously thought to be transcriptionally silent.

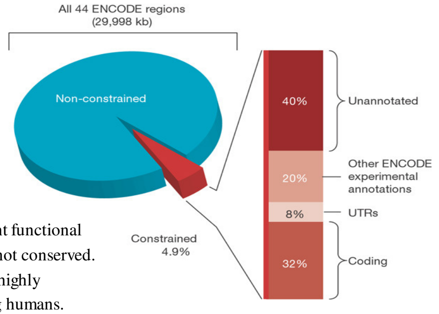
Highlights of ENCODE Project (II)

- Numerous previously unrecognized transcription start sites have been identified, many of which show chromatin structure and sequence-specific protein-binding properties similar to well understood promoters.
- Regulatory sequences that surround transcription start sites are symmetrically distributed, with no bias towards upstream regions.

Comparative Analysis

- 206Mb of genomic sequences orthologous to the human ENCODE DNA sequences were generated from 14 mammalian species
- The orthologous sequences were aligned using three alignment programs: TBA, MAVID and MLAGAN.
- Four independent methods that generated highly concordant results were then used to identify sequences under constraint (PhastCons, GERP, SCONE and BinCons).
- From these analyses, a high-confidence set of 'constrained sequences' was developed that correspond to 4.9% of the nucleotides in the ENCODE regions.
- Constrained sequence is a genomic region associated with evidence of negative selection (that is, rejection of mutations relative to neutral regions).

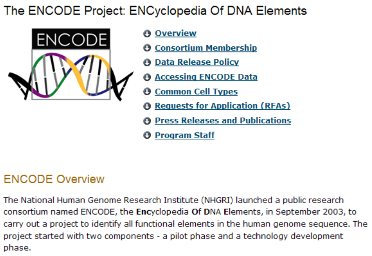
Constrained Sequences



Many important functional sequences are not conserved. They are even highly variable among humans.

ENCODE Portal at UCSC

- The main portal for ENCODE data is provided by the UCSC Genome Browser: genome.ucsc.edu/ENCODE/

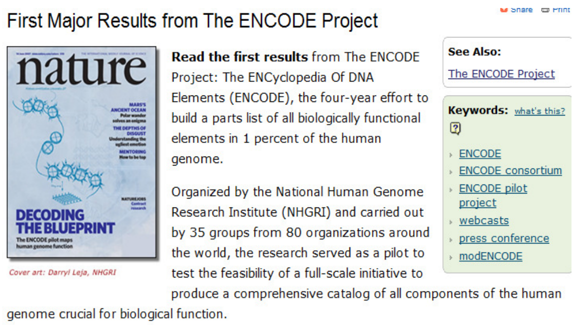


The ENCODE Project: ENCyclopedia Of DNA Elements

ENCODE Overview

The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the **Enc**yclopedia **Of D**NA **E**lements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence. The project started with two components - a pilot phase and a technology development phase.

ENCODE: First Major Results



First Major Results from The ENCODE Project

Cover art: Darryl Leja, NHGRI

Read the first results from The ENCODE Project: The ENCyclopedia Of DNA Elements (ENCODE), the four-year effort to build a parts list of all biologically functional elements in 1 percent of the human genome.

Organized by the National Human Genome Research Institute (NHGRI) and carried out by 35 groups from 80 organizations around the world, the research served as a pilot to test the feasibility of a full-scale initiative to produce a comprehensive catalog of all components of the human genome crucial for biological function.

See Also:
The ENCODE Project

Keywords: what's this?

- ENCODE
- ENCODE consortium
- ENCODE pilot project
- webcasts
- press conference
- modENCODE

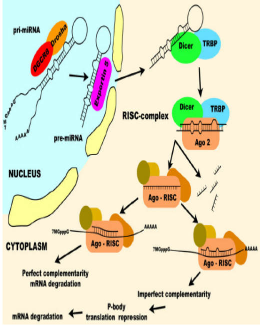
GENCODE Project (I)

- **GENCODE**: [A subproject of ENCODE] aim to produce a reference annotation for ENCODE
- The **GENCODE** consortium was formed to identify and map all protein-coding genes within the ENCODE regions.
  - This was achieved by a combination of initial manual annotation by the HAVANA (**H**uman **A**nd **V**ertebrate **Ana**lysis) team, experimental validation by the GENCODE consortium and a refinement of the annotation based on these experimental results.

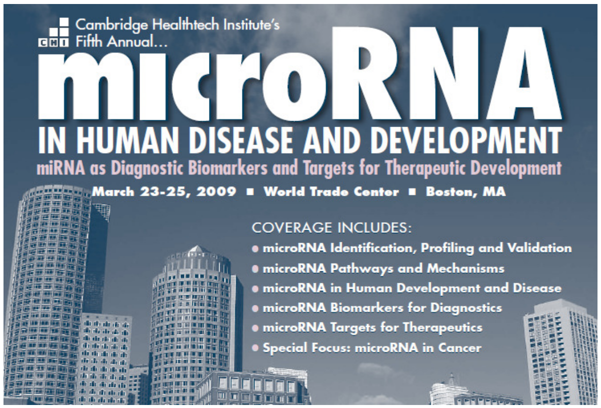
GENCODE Project (II)

- **GENCODE** seeks to identify all protein-coding genes in the ENCODE selected regions.
  - For each protein coding gene this means the delineation of a complete mRNA sequence for at least one splice isoform, and often for a number of additional alternative splice forms.
- Coding sequences for the 44 regions in the study have been ascertained by the Havana group.
  - In total there are 1097 CDS sequences from the 44 selected regions of the human chromosome.

MicroRNA [MiRNA]



- Pri-miRNA
- Drosha
- Precursor miRNA [Pre-miRNA]
- Dicer
- Translation Repression
- Translation Degradation
- MiRNA Gene Prediction
- MiRNA Target Prediction



Cambridge Healthtech Institute's Fifth Annual...

microRNA

IN HUMAN DISEASE AND DEVELOPMENT

miRNA as Diagnostic Biomarkers and Targets for Therapeutic Development

March 23-25, 2009 ▪ World Trade Center ▪ Boston, MA

COVERAGE INCLUDES:

- microRNA Identification, Profiling and Validation
- microRNA Pathways and Mechanisms
- microRNA in Human Development and Disease
- microRNA Biomarkers for Diagnostics
- microRNA Targets for Therapeutics
- Special Focus: microRNA in Cancer

Micro RNAs

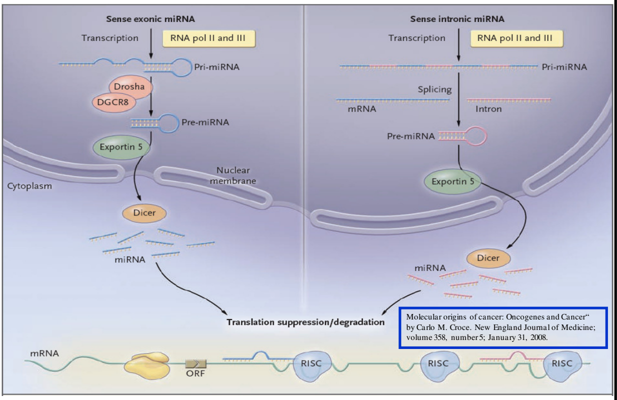
- **Micro RNAs** are a class of non-coding RNA gene whose products are nucleotide sequences (about 22 nucleotides long) that play important roles in regulation of translation and degradation of mRNAs through base pairing of partially complementary sites in the untranslated regions (UTRs) of the message.
- **miRNAs** are a class of small, evolutionarily conserved RNA molecules that regulate gene expression at the post-transcriptional level.

First Micro RNAs

- Micro RNAs (miRNA) were first discovered by Chalfie et al. through genetic studies in the nematode *Caenorhabditis elegans* as essential regulators of development.
  - *lin-4* and *let-7* seemed to be involved in controlling the timing of larval development
- Since then, numerous microRNAs have been found in different species:
  - miRBase (release 13.0) contains 9,499 microRNA entries from 103 species, among which 706 are human microRNAs.
  - many microRNA gene families are conserved among diverse species.

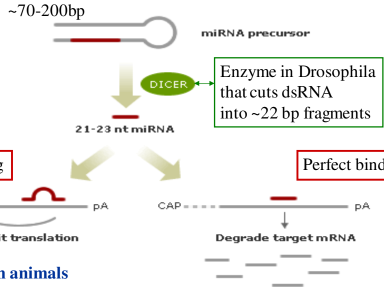
Producing Micro RNAs

- MiRNAs gene encode precursor RNAs that undergo processing to form miRNAs of length approximately 22 nucleotides.



Molecular origins of cancer: Oncogenes and Cancer“ by Carlo M. Croce. New England Journal of Medicine; volume 358, number 5; January 31, 2008.

MicroRNA Binding



Enzyme in Drosophila that cuts dsRNA into ~22 bp fragments

Imperfect binding

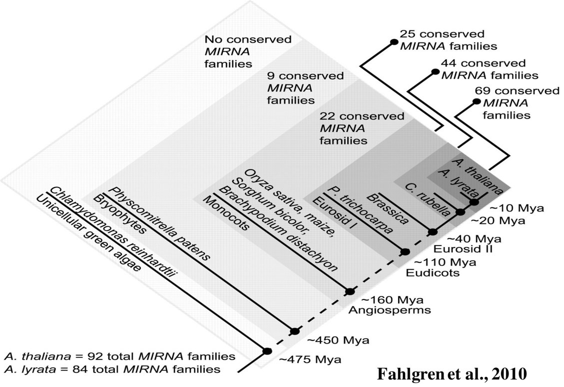
Perfect binding

Usually occurs in animals
Examples: lin-4 and let-7

Usually occurs in plants (scarecrow)

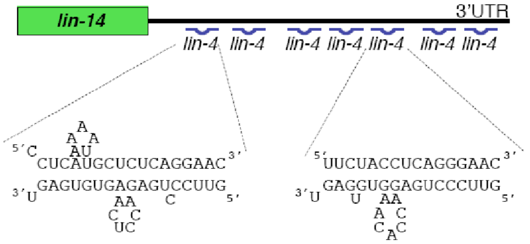
Animal versus Plant Targets

- In plants:
  - microRNAs bind almost perfectly to their target mRNAs
  - targets have been found anywhere on the mRNA
  - relatively few targets because microRNA-mRNA binding requires near-perfect complementarity
- In animals:
  - partial base-pairings with the target mRNAs
  - targets are typically found in the 3'-UTR, where the silencing machinery can easily interact with the initiation complex.
  - multiple targets on the same mRNA and often multiple microRNAs target the same mRNA



Fahlgren et al., 2010

Lin-4 Regulates Lin-14



MiRNA lin-4 acts to developmentally repress the accumulation of the lin-14 protein

miRNA Challenges and Hope

- The challenges are:
  - Predict the **functions** of the miRNAs
  - Identify the potential **target mRNAs** to which miRNAs will bind
  - Characterize the consequences of their **regulatory interactions**.
- The hope is:
  - RNA interference will be used to inactivate tumor genes or viruses.
    - miRNA-based therapies are under investigation

MicroRNAs and Cancer (I)

- More recently, in the past few years, it has been discovered that some of the 250 to 300 human **miRNA** are linked to cancers, such as leukemia, lung, breast, and colon cancers.
- Mapping of numerous **miRNA** genes has shown that many occur in chromosomal regions that undergo rearrangements, deletions, and amplifications in cancer cells.

Molecular origins of cancer: Oncogenes and Cancer" by Carlo M. Croce.
New England Journal of Medicine; volume 358, number 5; January 31, 2008.

MicroRNAs and Cancer (II)

MicroRNAs currently implicated in cancer

MicroRNA	Cancer Role	Cancer Type	Mechanism
miR-15	tumor suppressor	CLL	Bcl-2 inhibition
miR-16	tumor suppressor	CLL	Bcl-2 inhibition
miR-155	oncogene	lymphomas	unknown
let-7	tumor suppressor	lung cancer	ras inhibition
miR-17-92 cluster	oncogene	B cell lymphoma	unknown
miR-372	oncogene	testicular	inhibit p53 pathway
miR-373	oncogene	testicular	inhibit p53 pathway

Chronic Lymphocytic Leukemia (CLL): disease of white blood cells that won't die. It is the most common leukemia.

A **miRNA** can be a tumor suppressor if in a given cell type its target is an oncogene. It can be an oncogene if in a different cell type its target is a tumor-suppressor gene.

Table from "No miR Hype: MicroRNA's Cancer Role Expands" by Ken Garber
Journal of the National Cancer Institute, Vol. 98, No. 13, July 5, 2006

miR-15 and miR-16 in CLL

MicroRNAs currently implicated in cancer

MicroRNA	Cancer Role	Cancer Type	Mechanism
miR-15	tumor suppressor	CLL	Bcl-2 inhibition
miR-16	tumor suppressor	CLL	Bcl-2 inhibition
miR-155	oncogene	lymphomas	unknown
let-7	tumor suppressor	lung cancer	ras inhibition
miR-17-92 cluster	oncogene	B cell lymphoma	unknown
miR-372	oncogene	testicular	inhibit p53 pathway
miR-373	oncogene	testicular	inhibit p53 pathway

miR-15 and **miR-16** induce apoptosis by targeting the key survival protein Bcl-2, which is overexpressed in CLL

Table from "No miR Hype: MicroRNA's Cancer Role Expands" by Ken Garber
Journal of the National Cancer Institute, Vol. 98, No. 13, July 5, 2006

MiRNAs and Treatment

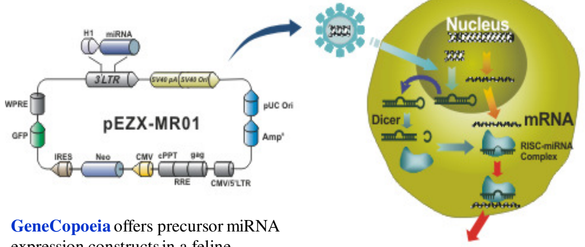
- Examples of the role **miRNA** plays in cancer pathophysiology involve *miR-15a* and *miR-16-1*, which are deleted or down-regulated in most indolent (slow to develop) cases of chronic lymphocytic leukemia.
- The discovery of the involvement of **miRNAs** in the initiation and progression of human cancer may provide additional targets for anticancer treatments.

Antagomirs

- Chemically modified antisense oligonucleotides (i.e., short strings of DNA bases complementary in sequence to their targets) injected into mice potentially silenced a target miRNA in the liver.
- The oligonucleotides were dubbed **antagomirs**.
- It is believed that **antagomirs** should be more effective against cancer-causing miRNAs than classic antisense therapy has been against protein-coding mRNAs:
  - **antagomirs** compete with miRNA targets for binding. An easier task than interfering with the protein translation machinery, which is the classic antisense mechanism.

Silencing of microRNAs in vivo with 'antagomirs by Jan Krützfeldt et al. Nature 438, 685-689 (Dec 2005).

GeneCopoeia



GeneCopoeia offers precursor miRNA expression constructs in a feline immunodeficiency virus (FIV) based lentiviral vector system

- Expression Translation Suppression

And/Or

- mRNA Cleavage and Degradation

www.genecopoeia.com/product/mirna

Computational Problems

- Multiple computational problems exist in microRNA research;

most notably:

- microRNA gene prediction
- microRNA target identification

MicroRNA Gene Prediction

- Traditional gene finding algorithms, which use statistical properties of coding regions, are not appropriate for finding microRNA genes.
- Homology-based searches fail due to the lack of a clear evolutionary model for microRNAs.
- Current techniques for finding microRNA genes take into account the following two properties:
  - the mature microRNA should be approximately 22 nt in length, and
  - it should be processed from a stem-loop precursor of around 65 nt in length.
- Some of the freely available microRNA gene finding tools for mammals are MirScan and ProMir, and for plants, MIRFINDER and FINDmicroRNA
- We concentrate on Target Prediction.

Target Prediction

- MicroRNA target prediction is an active area of research and the search for the best method for target prediction continues.
- Since microRNA targets in plants and animals show significant differences, different computational approaches are used.

Target Prediction

- Algorithms for predicting animal targets can be broadly divided into two categories:
  - The objective of the methods in the first category is to find targets for a given microRNA in the 3'-UTR:
    - use microRNA properties, such as base-pair complementarities, cross-species conservation and minimization of free energy.
  - The methods in the second category use machine learning techniques to classify a given microRNA-mRNA pair as a true or false interaction.

Traditional Target Prediction

- microRNA targets are typically found in the 3'-UTR region of animal mRNA.
- The microRNA-mRNA binding in animals is not perfect, it often contains mismatches, gaps, and wobble pairs (G:U), thus reducing the length of the perfect pairwise alignment between the microRNA and its target.
- However, in most microRNA-mRNA bindings, there is a region that exhibits a nearly perfect complementarity. This region is termed *seed* and it is found in the 5'-end of the microRNA and the 3'-end of the mRNA target.
- The most popular methods use seeds as primary filters: MiRanda, TargetScan, and PicTar.

MiRanda (I)

- Match (align) a microRNA against all 3'-UTR in a genome allowing for wobble pairs and small indels and score the alignments:
  - The algorithm uses higher scores for perfect matches in the 5'-end of the microRNA.
  - The scores are weighted based on the nucleotide position with respect to the 5'-end of the microRNA.
  - Only the alignments with the score above a threshold are kept.
- Each microRNA-mRNA alignment is filtered based on its computed thermostability.
- Retain only those mRNA targets that have been conserved in other (closely related) species.

MiRanda (II)

- Initially developed for *Drosophila melanogaster*
- Extended for target prediction in humans and other animal species.
- The latest version of the algorithm was updated to include microRNA expression profiles derived from sequencing a large set of mammalian tissues and cell lines.
- The MiRanda software currently predicts 1,934,522 target sites in 31,869 human gene isoforms.

TargetScan (I)

- Takes as input microRNAs that are conserved across a group of organisms and scans them against a set of orthologous 3’-UTR from the same organisms.
- All potential seeds, i.e. perfect matches in positions 2-8 of the microRNA, are extended to *target sites,* which may contain wobble pairs, indels and mismatches.
- A folding algorithm is used to determine the secondary structure of the microRNA-mRNA duplex and to compute the folding free energy.

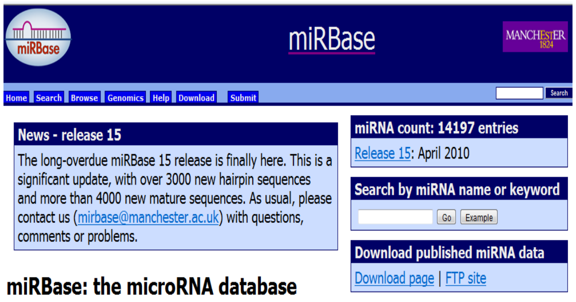
TargetScan (II)

- Each target site receives a score based on the number of matches predicted for the same 3’-UTR and their respective free energies.
- Target sites with a score above a threshold are output.
- Current versions of the software include specialized TargetScanHuman, TargetScanFly and TargetScanMouse.

PicTar

- Search for near-perfect seven nucleotide-long seeds starting at position 1 or 2 in the 5’-end of microRNA.
- Extend seeds into target sites and filter each target site based on the minimum free energy of the resulting microRNA-mRNA duplex.
- A second filter is applied to retain only the target sites that fall into overlapping positions of aligned orthologous sequences.
- The target sites that pass both filters are given a score that takes into account multiple binding sites for a single microRNA.
- Target sites with a score higher than a threshold are output.

miRBase and the Number of miRNAs



News - release 15

The long-overdue miRBase 15 release is finally here. This is a significant update, with over 3000 new hairpin sequences and more than 4000 new mature sequences. As usual, please contact us (mirbase@manchester.ac.uk) with questions, comments or problems.

miRNA count: 14197 entries

Release 15: April 2010

Search by miRNA name or keyword

Download published miRNA data

Download page | FTP site

miRBase: the microRNA database

August 4, 2010. www.mirbase.org

Target Identification

- Main problem: lack of verified targets:
  - thousands of microRNA genes have been experimentally verified, only a few of these genes have been assigned a function:
    - miRBase has 9,499 microRNA entries from 103 species
  - TarBase5.0, the database of experimentally supported targets, contained only 1,300 entries for nine species.

Problems of Traditional Methods

- Lack of high-throughput experimental techniques to confirm the thousands of predicted targets.
- The numbers of predicted targets differ among programs, with only limited overlap in the top-ranking targets:
  - mainly due to differences in selection criteria and the use of numerical cutoffs
  - TargetScan: complementarity in positions 2-7 from 5’ end of microRNA, whereas in PicTar: positions 1-8 or 2-9.

Machine Learning Approaches (I)

- Machine-learning methods try to classify microRNA-mRNA duplexes using a set of experimentally verified positive and negative interactions.
- The aim of these methods is to classify the predicted microRNA-mRNA target interactions as true or false and this is done by considering both, seed and non-seed regions of the target.

Machine Learning Approaches (II)

- Kim et. al implemented miTarget, a support vector machine that considers position, thermodynamic properties and structure of the 5’ and 3’ half of the hybridization site in microRNA-mRNA interactions.
- Saetrom et al. developed a TargetBoost algorithm that combines genetic programming with boosting.
  - The genetic programming component evolves a series of patterns which try to generalize properties of microRNA target sites, i.e. existence of a seed or a bulge of unpaired nucleotides.
  - Each of these patterns is a classifier itself. The boosting technique assigns a weight to each classifier, depending on its performance on the training set.

Weaknesses of Machine Learning Approaches

- The main weakness of the classifier methods is the small size (or even lack of) negative training data.

  [Negative Training Data: Known miRNA binds to mRNA but it is experimentally known that the binding site is NOT a target].
- Some authors attempt to overcome this drawback by artificially generating negative target sites and using them in their classifiers.
- However, until one has the technology for verifying microRNA-mRNA interactions, this weakness will persist.

Alternatives

- "Assessing potential microRNA targets based on a Markov model" by Fu et al., 2009:
  - developed a Markov model to learn from known microRNA-mRNA duplexes and applied it to filter out predictions of traditional algorithms:
    - only 30% of MiRanda predictions were picked up by the model.
    - but when the model was applied to the intersection of MiRanda and PicTar predictions, the model picked 70% of the targets.

Binding Representation

The 0-1 sequences generated by the model represent microRNA-target complementarity base-pairing binding patterns:

```
target  5' A         C A G A G G U       C               U 3'
             A A C U               C C U   C U A C C U C
             | | | |               | | |   | | | | | | |
             U U G A               G G A   G A U G G A G
miRNA   3'           U A U G U U         U               U 5'
-----------------------------------------------------------
target  5' A A A C U C A G A G G U C C U C C U A C C U C U 3'
             | | | |               | | |   | | | | | | |
miRNA   3'   1 1 1 1 0 0 0 0 0 0   1 1 1 0 1 1 1 1 1 1 1 0
-----------------------------------------------------------
miRNA   3'   1 1 1 1 0 0 0 0 0 0   1 1 1 0 1 1 1 1 1 1 1 0
```

Lack of Uniqueness

The 0-1 sequence can represent different binding scenarios:

```
----------------------------------------------------------------
a.  target  5' A         C A G A G G U       C               U 3'
                 A A C U               C C U   C U A C C U C
                 | | | |               | | |   | | | | | | |
                 U U G A               G G A   G A U G G A G
    miRNA   3'           U A U G U U         U               U 5'
----------------------------------------------------------------
b.  target  5'             G A G                             U 3'
                 A A C U               C C U   C U A C C U C
                 | | | |               | | |   | | | | | | |
                 U U G A               G G A   G A U G G A G
    miRNA   3'           U A U G U U         U               U 5'
----------------------------------------------------------------
c.  target  5'                                               U 3'
                 U U G A               G G A   G A U G G A G
                 | | | |               | | |   | | | | | | |
                 A A C U               C C U   C U A C C U C
    miRNA   3'           U A U G U U         U               U 5'
----------------------------------------------------------------
d.  target  5' A A A C U C A G A G G U C C U C C U A C C U C U 3'
                 | | | |               | | |   | | | | | | |
    miRNA   3'   1 1 1 1 0 0 0 0 0 0   1 1 1 0 1 1 1 1 1 1 1 0
----------------------------------------------------------------
e.  target  5' A A A C U               C C U C C U A C C U C U 3'
                 | | | |               | | |   | | | | | | |
    miRNA   3'   1 1 1 1 0 0 0 0 0 0   1 1 1 0 1 1 1 1 1 1 1 0
----------------------------------------------------------------
```

Selecting the Model

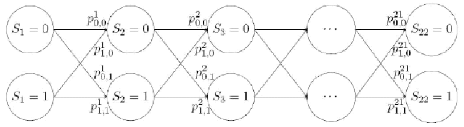
- On the basis of the 0-1 sequence representation, a generative chain model was chosen:
  - Each 0-1 sequence can be viewed as being generated by the model with a certain probability.
  - Since the 0-1 sequence stemming from biological molecular sequence (i.e., the microRNA sequence), we assume that the 0-1 sequence has a first-order Markov property.
- Since different sites have significantly different properties with respect to base-pairing binding statistical characteristics, a non-homogeneous rather than homogeneous Markov model is adopted.

Model States

- Two states were assigned for the *ith* base site of 22-nucleotide-length microRNA from 5' to 3', state 1 and state 0.
- Either state can be viewed as a value of a variable s_i.
  - State 1 stands for forming Watson-Crick pairing, while the state 0 stands for unforming pairing.
- From the state s_i, either the next state 1 or 0 can be transferred with probability $p^{i}_{si,0}$ or $p^{i}_{si,1}$, respectively.

Probability of a Sequence

pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
miRNA_1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	0	1	0	0	1
miRNA_2	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0
miRNA_3	0	1	0	1	0	0	0	1	0	0	0	1	0	0	1	1	1	1	1	1	1	1
:																						
:																						
miRNA_N	0	1	1	1	0	0	0	1	1	0	1	0	0	0	0	0	1	1	1	1	1	1



For Example: **0111111101110000001111**

$$P_0^{1} * P_{0,1}^{1} * P_{1,1}^{2} * \ldots * P_{1,1}^{7} * P_{1,0}^{8} * P_{0,1}^{9} * \ldots * P_{1,1}^{21}$$

Estimating Model Parameters

- All of the model parameters were estimated by maximum likelihood estimation method.
- Used 128 0-1 sequences, corresponding to 128 known human microRNA-mRNA bindings from TarBase.
- Discovered that not all transition probabilities were needed, probability parameters corresponding to sites 2-11 achieve the maximal recognition rates.
- Model identified 110 targets from the set of 128 (86%).

Links for MiRNA

- Ambion, Inc,

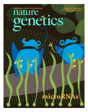
  www.ambion.com/main/explorations/mirna.html
- Nature Genetics

  www.nature.com/ng/supplements/micrornas/rosetta_video.mpg
- Rosetta Genomics

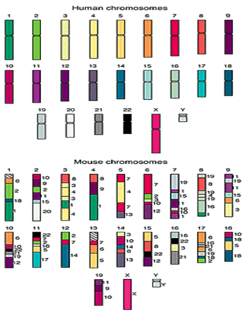
  Developer of microRNA-based diagnostic tests and therapeutic tools

  www.rosettagenomics.com/inner_video.asp?first_tier=97
- Nature Genetics – Several articles on miRNAs

  www.nature.com/ng/journal/v38/n6s/index.html



Genome Rearrangement



- Reversals
- Translocations
- Fusions
- Fissions
- Mammalian Evolution
- Mathematical Models

Chromosomal Exchanges

- During biological evolution, inter- and intra-chromosomal exchanges of chromosomal fragments disrupt the order of genes on a chromosome.
- **Genome rearrangement** approach:

  The use of combinatorial optimization techniques to infer a sequence of rearrangement events to account for the differences among the genomes

Genome Rearrangement

- Genes are arranged along the genome.
- Distinct species often have surprisingly many genes in common, but in different order and with different orientations.
- Periodic, large-scale **genome rearrangement** events, that alter the order and/or orientation of gene sequences occurred.

Genome Rearrangement Events

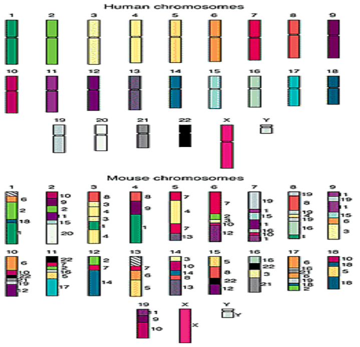
- The most common rearrangement events:
  - **Reversal** (**inversion**): reverses the order of genes in a chromosome.
  - **Transposition**: removes a sequence of genes from the chromosome and inserts it into another place on the same chromosome.
  - **Translocation**: same as transposition but the sequence of genes is inserted in a different chromosome.
  - **Fusion**: concatenates 2 chromosomal regions into one.
  - **Fission**: does the opposite work of fusion.

The Mighty Mouse

Mouse versus Human

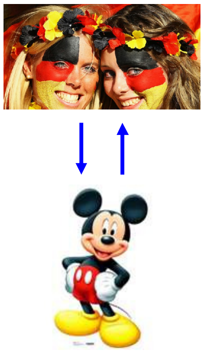
Approximately the same number of chromosomes and local gene order in mammals.

Insights into mouse genetics are likely to **illuminate** human genetics as well.



Mighty Mouse vs Human Genome (I)

- Humans and mice have similar genomes, but their genes are ordered differently
- ~245 rearrangements
  - **Reversals**
  - Fusions
  - Fissions
  - Translocation



Translocation – Fusion - Fission

Translocation

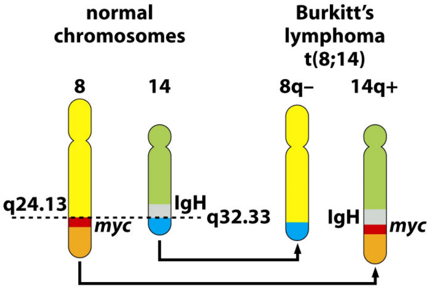
1 2 3 9 8 7 → 1 2 6 2 4 1
4 5 6 2 4 1 → 4 5 3 9 8 7

Fusion

1 2 3 4
5 6 ⇄ 1 2 3 4 5 6

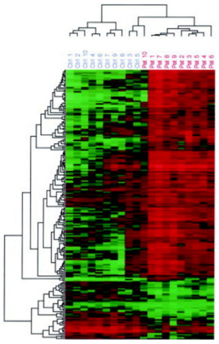
Fission

Burkitt's Lymphoma t(8;14)



The Biology of Cancer by R. Weinberg. Garland Science 2007

Introduction to Microarrays

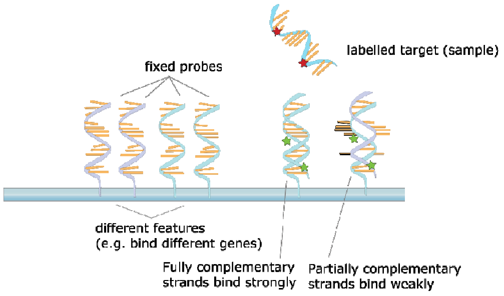


- ❖ Spotted arrays
- ❖ Affymetrix
- ❖ Learning basic biology
- ❖ Yeast
- ❖ Gene Expression
- ❖ Guilt by association
- ❖ "What if" questions

Fundamentals of Microarrays

- Microarrays are composed of short DNA sequences attached to a slide at high density
- Microarrays work by exploiting the ability of an mRNA molecule to bind specifically to the DNA sequence from which it originates
- This RNA (or its DNA derivative) is fluorescently labeled so that the amount of hybridization can be quantitatively measured

Hybridization of Target to Probe



History of DNA Microarrays

- Microarrays descend from Southern and Northern blotting. Unknown DNA is transferred to a membrane and then probed with a known DNA sequence with a label.
- In microarrays, the known DNA sequence (or probe) is on the membrane while the unknown labeled DNA (or target) is hybridized and then washed off so only specific hybrids remain.
- Dot blots of different genes in an array were used to assay gene expression as early as 1987.
- Complete genome of all Saccharomyces cerevisiae ORFs on a microarray were published in 1997 by Lashkari et al.

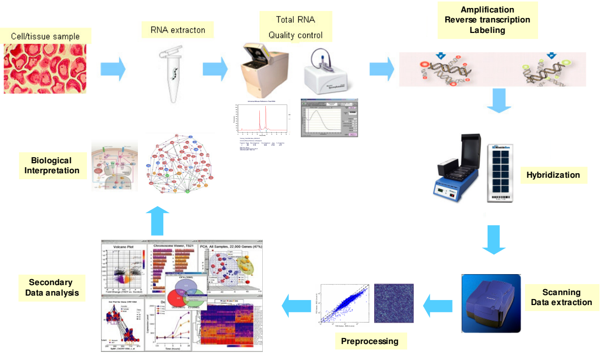
Microarray Applications

Array	Probes *on the array*	Targets *to be hybridized*	Large-scale Analysis of…
Gene Expression	DNA (cDNA, oligos: gene representatives)	mRNA/cDNA	transcriptional alterations
CGH	DNA (clones, oligos)	DNA	Genomic changes in cancers
SNP	DNA (oligos)	DNA	Genotyping; Genomic changes
Methylation	DNA (CpG island)	DNA (IP or bisulfite-treated)	Methylation-status in genes
Promoter	DNA (promoter ~1kb)	DNA (ChIP-enriched)	Transcription factor binding sites; histone modifications
Tiling	DNA	All of the above	All of the above; sequencing; gene annotation
Protein	antibody	protein	Protein expression (ELISA)
Tissue	tissues	proteins	Histology; protein expression (immunohistochemistry)

Definitions

- Gene expression:
  - A process by which inheritable information from a gene is made into a functional gene product
- Gene expression profiling:
  - A measurement of the activity of thousands of genes at once, creating a global profile of cellular function.
  - Profiles can for example distinguish between cells that are actively dividing, or show how the cells react to a particular treatment.
  - The sequence tells us only what the cell could possibly do, the expression profile tells us what the cell is actually doing at that moment.

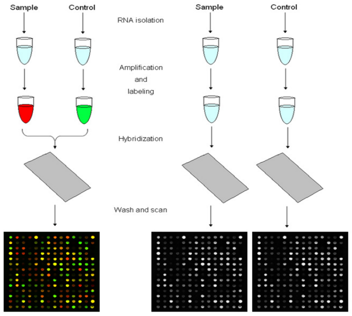
Microarray Assay



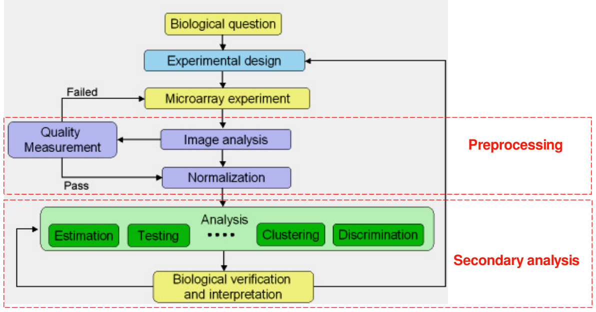
Microarray Technology

- Basic idea: mRNA hybridizes best to exactly complementary sequences.
- Method**:**
  - Probes are attached to a substrate in a known location
  - mRNA in one or more samples are fluorescently labelled
  - samples are hybridized to probe array, excess is washed off, and fluorescence reading are taken for each position
- Two major classes**:**
  - “custom” cDNA arrays (probes are full length cDNAs)
  - “Affymetrix” oligonucleotide arrays (probes are unique ~25bp segments from genes & ESTs)

Two Color versus One Color Microarray

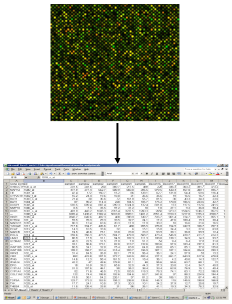


Gene Expression Experiment Pipeline



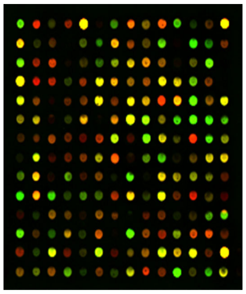
Data Preprocessing

- Objective
  - Convert image of thousands of signals to a signal value for each gene or probe set.
- Multiple step
  - Image analysis
  - Background and noise subtraction
  - Normalization
  - Expression value for a gene or probe set
- Usually done by proprietary software



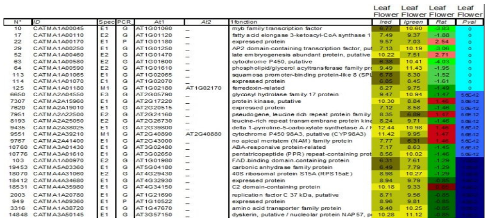
Microarray Image Analysis

- **Red** dots indicate induced (more mRNA) gene in the experiment
- **Green** dots indicate repressed (less mRNA) gene in the experiment
- **Yellow** dots indicate no change



Molecular Biology of the Cell (Fifth addition) 2008 Fig 8-73

Biological Interpretation

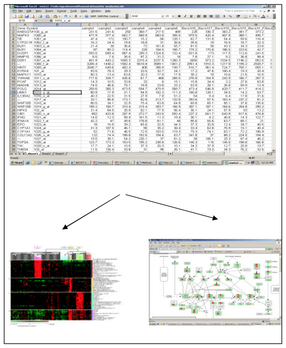


							Leaf Flower	Leaf Flower	Leaf Flower	Leaf Flower
N°	ID	Spec	PCR_	At1	At2	1fonction	Ired	Igreen	Rat	Pval
10	CATMA1A00045	E1	G	AT1G01060	_	myb family transcription factor	6.77	10.60	-3.83	0
17	CATMA1A00110	E2	G	AT1G01120	_	fatty acid elongase 3-ketoacyl-CoA synthase 1	7.49	9.37	-1.88	0
22	CATMA1A00170	E1	P	AT1G01180	_	expressed protein	9.57	7.03	2.54	0
29	CATMA1A00250	E1	G	AT1G01250	_	AP2 domain-containing transcription factor, put	7.13	10.19	-3.06	0
52	CATMA1A00460	E2	G	AT1G01470	_	late embryogenesis abundant protein, putative	10.22	7.51	2.71	0
63	CATMA1A00580	E2	G	AT1G01600	_	cytochrome P450, putative	6.38	10.41	-4.03	0
64	CATMA1A00590	E2	G	AT1G01610	_	phospholipid/glycerol acyltransferase family pro	9.49	11.43	-1.95	0
113	CATMA1A01065	E1	G	AT1G02065	_	squamosa promoter-binding protein-like 8 (SPL	6.78	8.30	-1.52	0
114	CATMA1A01070	E1	G	AT1G02070	_	expressed protein	6.85	8.45	-1.61	0
125	CATMA1A01180	M1	G	AT1G02180	AT1G02170	ferredoxin-related	8.27	9.75	-1.49	0
6650	CATMA2A04550	E3	G	AT2G05790	_	glycosyl hydrolase family 17 protein	9.47	10.94	-1.47	5.6E-12
7307	CATMA2A15960	E1	G	AT2G17220	_	protein kinase, putative	10.30	8.84	1.46	5.6E-12
7620	CATMA2A19010	E1	G	AT2G20515	_	expressed protein	7.12	8.58	-1.46	5.6E-12
7951	CATMA2A22500	E2	G	AT2G24160	_	pseudogene, leucine rich repeat protein family	8.35	6.89	1.47	5.6E-12
8193	CATMA2A25050	E2	G	AT2G26730	_	leucine-rich repeat transmembrane protein kina	8.24	9.71	-1.46	5.6E-12
9435	CATMA2A38025	E1	G	AT2G39800	_	delta 1-pyrroline-5-carboxylate synthetase A / P	12.44	10.98	1.46	5.6E-12
9551	CATMA2A39210	M1	G	AT2G40890	AT2G40880	cytochrome P450 98A3, putative (CYP98A3)	11.42	9.95	1.47	5.6E-12
9767	CATMA2A41400	E1	G	AT2G43000	_	no apical meristem (NAM) family protein	7.77	6.31	1.46	5.6E-12
10760	CATMA3A01430	E1	G	AT3G02480	_	ABA-responsive protein-related	7.17	8.63	-1.45	5.6E-12
10776	CATMA3A01610	E1	G	AT3G02650	_	pentatricopeptide (PPR) repeat-containing prot	8.56	10.02	-1.46	5.6E-12
103	CATMA1A00970	E2	G	AT1G01980	_	FAD-binding domain-containing protein	6.31	7.61	-1.29	[illegible]
19453	CATMA5A03360	E1	G	AT5G04180	_	carbonic anhydrase family protein	6.49	7.79	-1.29	[illegible]
18070	CATMA4A31060	E2	G	AT4G29430	_	40S ribosomal protein S15A (RPS15aE)	8.98	10.27	-1.29	[illegible]
18412	CATMA4A34680	E1	G	AT4G32930	_	expressed protein	8.94	9.79	-0.85	[illegible]
18531	CATMA4A35980	E1	G	AT4G34150	_	C2 domain-containing protein	10.18	9.33	0.85	[illegible]
2003	CATMA1A20780	E1	G	AT1G21690	_	replication factor C 37 kDa, putative	8.71	9.56	-0.85	[illegible]
949	CATMA1A09360	E1	P	AT1G10522	_	expressed protein	8.96	9.81	-0.85	[illegible]
3316	CATMA1A38720	E1	G	AT1G47670	_	amino acid transporter family protein	9.40	10.25	-0.85	[illegible]
14848	CATMA3A50145	E1	G	AT3G57150	_	dyskerin, putative / nucleolar protein NAP57, pu	10.28	11.12	-0.85	[illegible]

- Annotation
- Clustering
- Pathway analysis
- Gene Set Enrichment Analysis (GSEA)

Data Analysis

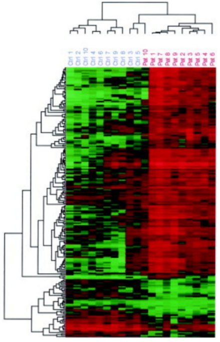
- Class comparison:
  - identify genes that are expressed differently, e.g. tumour vs. normal tissue or treated vs. untreated samples
- Class discovery:
  - divides samples into reproducible classes that have similar behaviour or properties
  - Classification unsupervised
- Class prediction:
  - Classification supervised
  - Biological annotation
  - Pathway analysis



Class Comparison

- Differential expression analysis:
  - What genes are up regulated between control and test or multiple test conditions?
    - Normal vs. tumor
    - Treated vs. untreated
- Fold change
- Statistics
  - t-test, non-parametric tests, ANOVA, SAM

Class Discovery



- Which classes are similar?
- Are there subgroups?
- Methods:
  - Unsupervised methods:
    - Cluster analysis
    - K-means clustering,
    - Principal Component Analysis (PCA)
    - Self-organizing maps (SOM)
  - Supervised methods:
    - Partial Least Square (PLS)
    - Hierarchical clustering

Functional Genomics

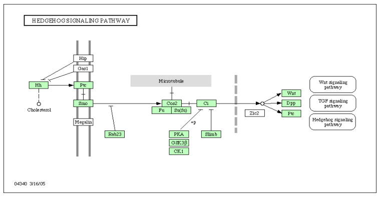
- Take a list of "interesting" genes and find their biological relationships.
- Gene lists may come from significance/classification analysis of microarrays, proteomics, or other high-throughput methods.
- Requires a reference set of “biological knowledge”.

Gene Annotation

- Gene ontology is a gene annotation system, a controlled vocabulary used to describe gene products:
  - What does a gene product do?
  - Why does it perform these activities?
  - Where does it act?
  - Are several genes involved in the same process?
- Different platforms provide gene ontology mining tool which returns GO terms for probe sets:
  - Panther (Applied Bios stems)
  - NetAffx Analysis Center (Affymetrix)
  - DAVID database
  - www.geneontology.org/GO.tools.microarray.shtmlm
- Annotation challenges
  - Databases change regularly
  - Various databases refer to the same protein by different names
  - A changing understanding of protein function
  - A gene product may be part of several different ontologies

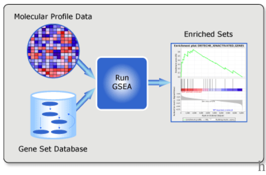
Pathway Analysis

- Discover relationships between the annotated genes:
  - KEGG http://www.genome.jp/kegg/
  - Pubgene http://www.pubgene.com/



GSEA

- Gene Set Enrichment Analysis (GSEA) uses patterns to find regulated genes.
- A computational method that determines whether an a priori defined set of genes shows statistically significant differences between two biological states:
  - Do any of the previously defined gene sets exhibit unusual behavior in the current expression profile?



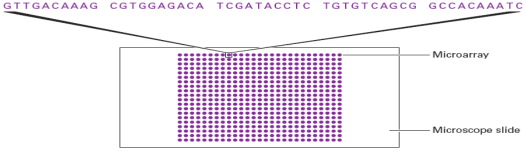
http://www.broad.mit.edu/gsea/

The Importance of Yeast

The Yeast Genome

- The yeast genome has about 6,200 genes.
- All 6,200 genes are amplified by polymerase chain reaction (PCR).
- The PCR products are verified, purified, and spotted onto an ordinary glass microscope slide by a robot.
- The spotted DNA is denatured and covalently linked to the glass slide.
- Each spot contains many amplified copies of a single gene.

Gene Expression Microarrays



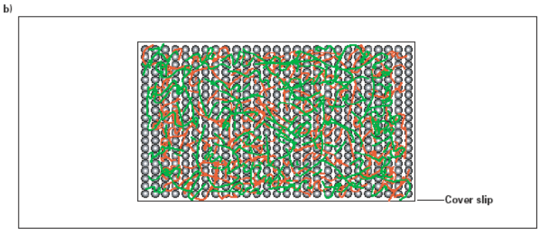
```
TCCTTTCCGG  AACGGTTGGC  GTCTGCGCAC  GGCGGTGTGG  GGCATGACAT
GCCGCCCCAG  GAACAACCCC  GACACGGCTT  TAAGCCTCTC  AAATCGCTGT
AGACATCATC  TTTACGTGCT  TGCCACCATT  TGCCACCATT  AGGGCTGTTC
CCGCGACGAC  TCGCCATTCA  ACCTCAGTCC  TTCGGGTTGA  GCGAGTGGGT
CGCGCGCAAG  GTGCGAATGG  GTCGCGCGCA  AAGTGTTGCG  CTGGCTGTAT
TATATGCTGC  CTATAGCGAG  ACTAACGACC  CACACTTTCA  CACAAGGATT
TCCCGCTAAT  GGGTACCTCG  CGTCAGGACC  TTGACGCAAG  CGCGCCTTCG
GTTGGCCCCA  AGCTTGCTAG  GACTACTTAT  CTTGAGCTCA  TTTAACATCC
CGGCGCCTCT  CCGGGAGCGG  TCGTCGCGAA  GAAGTCAAAC  CCGGAACGGC
GTTGACAAAG  CGTGGAGACA  TCGATACCTC  TGTGTCAGCG  GCCACAAATC
```

Each purple spot indicates the location of a PCR product on the glass slide. One particular spot has been chosen to illustrate the presence of one gene’s sequence.
On a real microarray, each spot is about 100 µm in diameter.

Yeast DNA Chip Experiment

- Cells are grown in two different conditions:
  - In the presence of oxygen
  - In the absence of oxygen.
- The 2 populations of mRNAs are harvested from each population of cells and separately converted into cDNAs.
- The two populations of cDNAs are colored either green or red, each representing the transcriptome from one population of cells.

Hybridization



The red and green cDNAs are mixed, placed on the chip, covered by a glass cover slip, and incubated overnight with the DNA microarray.

Results From a Single DNA Chip

The microarray is put under a scanner that uses light to excite the dyes and sensors to detect the dyes to record the location and two-color intensities for each spot.

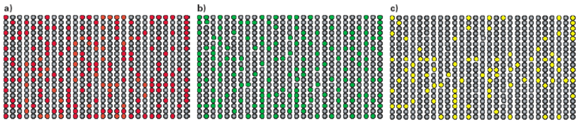


Diagram of a single DNA chip showing:
a) the red transcriptome, b) the green transcriptome,
c) which genes are expressed in both transcriptomes.
Some genes are not expressed in either growth condition (gray spots).

Image Analysis

- Scanning one microarray chip takes about 20 minutes.
- When completed, the green color image and the red color image are stored in a computer for image analysis.
- The computer also generates a new merged image, with yellow spots indicating the open reading frames (ORFs) that are transcribed in both transcriptomes

Conversion of Color Spots to Table

- Yellow spots are a visual way of depicting a red-to-green ratio 1:1. More typically, the merged image will be a bit more green or a bit more red.
- The color spots are then converted to numbers that represent the light intensity of
  - red dye,
  - green dye, and
  - the ratio of red to green.

Conversion Table for 14 Genes (I)

Block	Column	Row	Gene Name	Red	Green	Red : Green Ratio
1	1	1	*Tub1*	2,345	2,467	0.95
1	1	2	*Tub2*	3,589	2,158	1.66
1	1	3	*Sec1*	4,109	1,469	2.80
1	1	4	*Sec2*	1,500	3,589	0.42
1	1	5	*Sec3*	1,246	1,258	0.99
1	1	6	*Act1*	1,937	2,104	0.92
1	1	7	*Act2*	2,561	1,562	1.64
1	1	8	*Fus1*	2,962	3,012	0.98
1	1	9	*Idp2*	3,585	1,209	2.97
1	1	10	*Idp1*	2,796	1,005	2.78
1	1	11	*Idh1*	2,170	4,245	0.51
1	1	12	*Idh2*	1,896	2,996	0.63
1	1	13	*Erd1*	1,023	3,354	0.31
1	1	14	*Erd2*	1,698	2,896	0.59

Conversion Table for 14 Genes (II)

Block	Column	Row	Gene Name	Red	Green	Red : Green Ratio
1	1	1	*Tub1*	2,345	2,467	0.95
1	1	2	*Tub2*	3,589	2,158	1.66
1	1	3	*Sec1*	4,109	1,469	2.80
1	1	4	*Sec2*	1,500	3,589	0.42
1	1	5	*Sec3*	1,246	1,258	0.99
1	1	6	*Act1*	1,937	2,104	0.92
1	1	7	*Act2*	2,561	1,562	1.64
1	1	8	*Fus1*	2,962	3,012	0.98
1	1	9	*Idp2*	3,585	1,209	2.97
1	1	10	*Idp1*	2,796	1,005	2.78
1	1	11	*Idh1*	2,170	4,245	0.51
1	1	12	*Idh2*	1,896	2,996	0.63
1	1	13	*Erd1*	1,023	3,354	0.31
1	1	14	*Erd2*	1,698	2,896	0.59

The fluorescence intensity of each color was determined and the ratio of red signal to green signal was calculated.

The table shows data for 14 of the 6,200 genes on the full microarray.

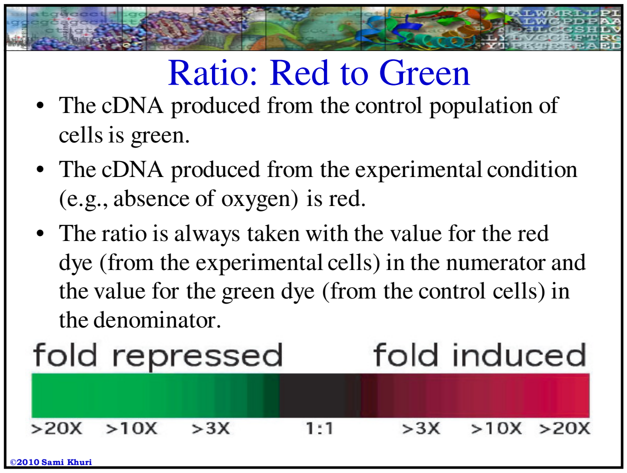
Note that the location of each spot is shown in the table.

Red–Green Color Scale

- Humans process visual data much faster than numbers.
- So the numerical data ratios of red to green is converted into a visually comprehensive system.

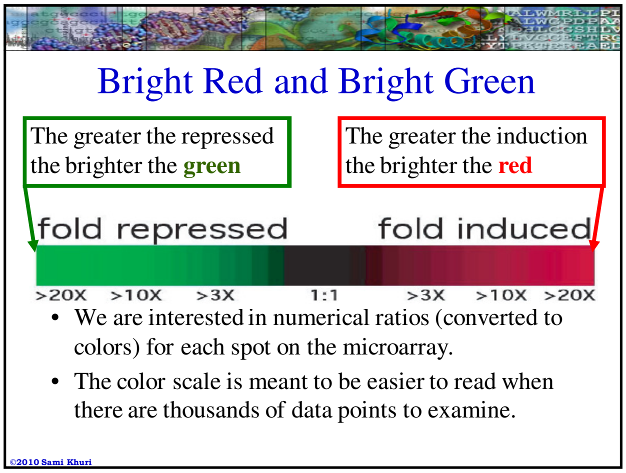
Induced and Repressed Genes

- Genes that are equally transcribed in both conditions are coded black.
- Experimentally induced genes (stimulated to produce more mRNA) are colored red.
- Experimentally repressed genes (stimulated to produce less mRNA) are colored green.



Ratio: Red to Green

- The cDNA produced from the control population of cells is green.
- The cDNA produced from the experimental condition (e.g., absence of oxygen) is red.
- The ratio is always taken with the value for the red dye (from the experimental cells) in the numerator and the value for the green dye (from the control cells) in the denominator.

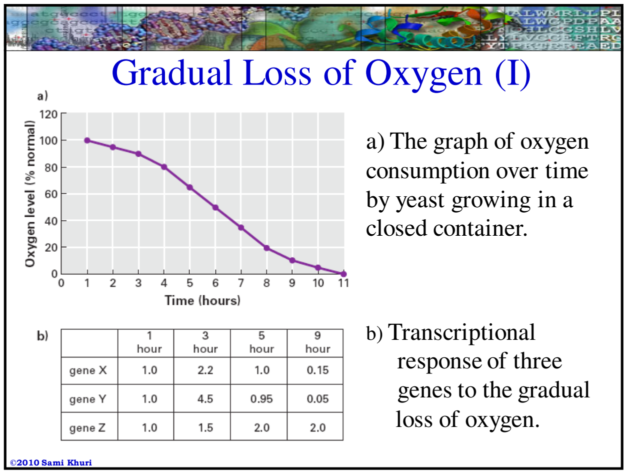


Bright Red and Bright Green

The greater the repressed the brighter the **green**

The greater the induction the brighter the **red**

- We are interested in numerical ratios (converted to colors) for each spot on the microarray.
- The color scale is meant to be easier to read when there are thousands of data points to examine.



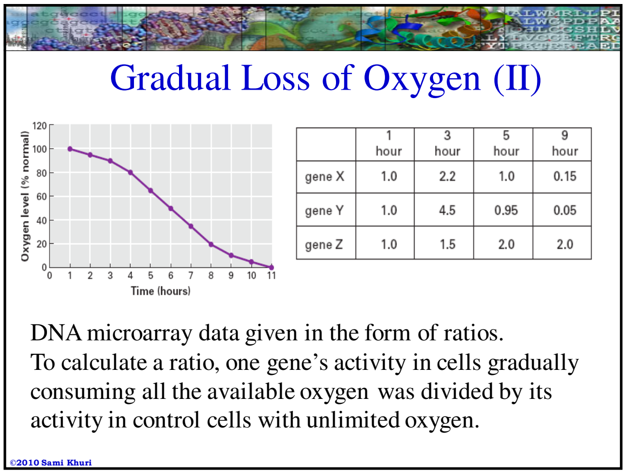
Gradual Loss of Oxygen (I)

a) The graph of oxygen consumption over time by yeast growing in a closed container.

b)

	1 hour	3 hour	5 hour	9 hour
gene X	1.0	2.2	1.0	0.15
gene Y	1.0	4.5	0.95	0.05
gene Z	1.0	1.5	2.0	2.0

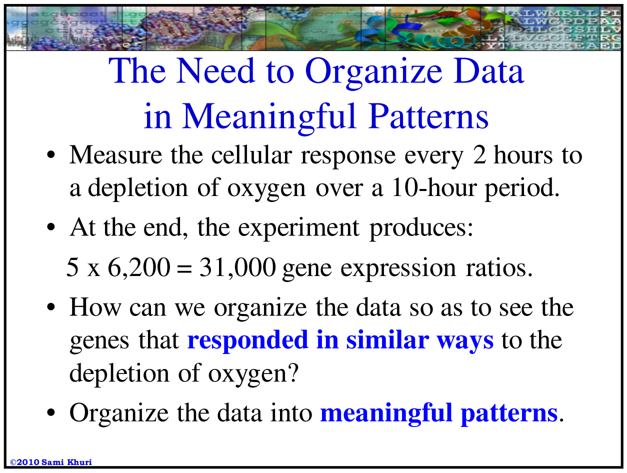
b) Transcriptional response of three genes to the gradual loss of oxygen.



Gradual Loss of Oxygen (II)

	1 hour	3 hour	5 hour	9 hour
gene X	1.0	2.2	1.0	0.15
gene Y	1.0	4.5	0.95	0.05
gene Z	1.0	1.5	2.0	2.0

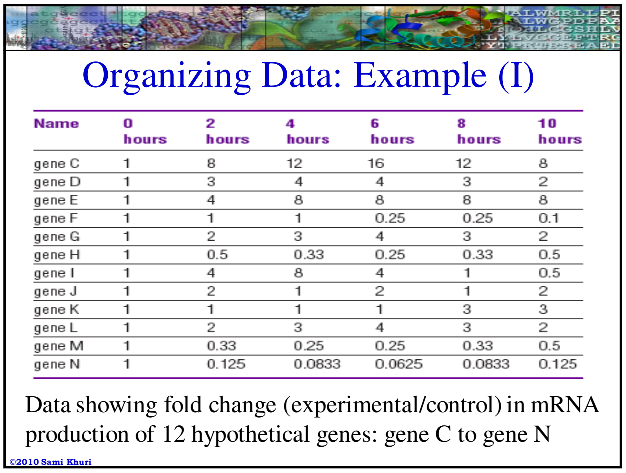
DNA microarray data given in the form of ratios.
To calculate a ratio, one gene's activity in cells gradually consuming all the available oxygen was divided by its activity in control cells with unlimited oxygen.



The Need to Organize Data in Meaningful Patterns

- Measure the cellular response every 2 hours to a depletion of oxygen over a 10-hour period.
- At the end, the experiment produces:

  5 x 6,200 = 31,000 gene expression ratios.
- How can we organize the data so as to see the genes that **responded in similar ways** to the depletion of oxygen?
- Organize the data into **meaningful patterns**.



Organizing Data: Example (I)

Name	**0 hours**	**2 hours**	**4 hours**	**6 hours**	**8 hours**	**10 hours**
gene C	1	8	12	16	12	8
gene D	1	3	4	4	3	2
gene E	1	4	8	8	8	8
gene F	1	1	1	0.25	0.25	0.1
gene G	1	2	3	4	3	2
gene H	1	0.5	0.33	0.25	0.33	0.5
gene I	1	4	8	4	1	0.5
gene J	1	2	1	2	1	2
gene K	1	1	1	1	3	3
gene L	1	2	3	4	3	2
gene M	1	0.33	0.25	0.25	0.33	0.5
gene N	1	0.125	0.0833	0.0625	0.0833	0.125

Data showing fold change (experimental/control) in mRNA production of 12 hypothetical genes: gene C to gene N

Organizing Data: Example (II)

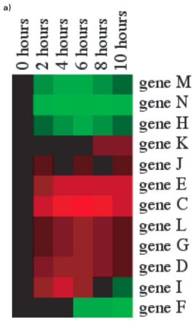
Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene M	1	0.33	0.25	0.25	0.33	0.5
gene N	1	0.125	0.0833	0.0625	0.0833	0.125
gene H	1	0.5	0.33	0.25	0.33	0.5
gene K	1	1	1	1	3	3
gene J	1	2	1	2	1	2
gene E	1	4	8	8	8	8
gene C	1	8	12	16	12	8
gene L	1	2	3	4	3	2
gene G	1	2	3	4	3	2
gene D	1	3	4	4	3	2
gene I	1	4	8	4	1	0.5
gene F	1	1	1	0.25	0.25	0.1

Reorganization (clustering) of gene order from the previous table based on similarity of expression patterns or profiles

Organizing Data: Example (III)

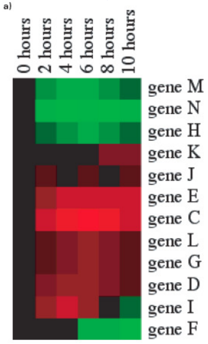
- As can be seen from the example, when the genes are reorganized, or clustered, according to the **similarity of their expression ratios**, it is easier to detect genes with similar activity.
- If these **ratios are then converted to colors**, one can quickly understand patterns of gene activity.

Organizing Data: Example (IV)



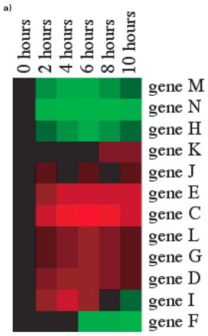
a) **Gene expression ratios** from the previous table were converted to colors. Color conversion of ratios facilitates the interpretation of changes in gene expression profiles.
b) Color scale used to convert induced genes to red and repressed genes to green.
Genes with ratios of 1:1 are colored black.

Organizing Data: Example (V)



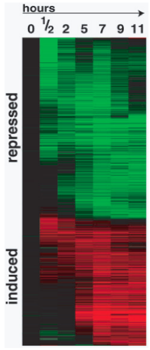
Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene M	1	0.33	0.25	0.25	0.33	0.5
gene N	1	0.125	0.0833	0.0625	0.0833	0.125
gene H	1	0.5	0.33	0.25	0.33	0.5
gene K	1	1	1	1	3	3
gene J	1	2	1	2	1	2
gene E	1	4	8	8	8	8
gene C	1	8	12	16	12	8
gene L	1	2	3	4	3	2
gene G	1	2	3	4	3	2
gene D	1	3	4	4	3	2
gene I	1	4	8	4	1	0.5
gene F	1	1	1	0.25	0.25	0.1

Organizing Data: Example (VI)



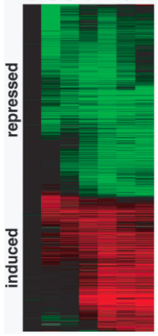
Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene M	1	0.33	0.25	0.25	0.33	0.5
gene N	1	0.125	0.0833	0.0625	0.0833	0.125
gene H	1	0.5	0.33	0.25	0.33	0.5
gene K	1	1	1	1	3	3
gene J	1	2	1	2	1	2
gene E	1	4	8	8	8	8
gene C	1	8	12	16	12	8
gene L	1	2	3	4	3	2
gene G	1	2	3	4	3	2
gene D	1	3	4	4	3	2
gene I	1	4	8	4	1	0.5
gene F	1	1	1	0.25	0.25	0.1

Gene Expression Clusters (I)



- When gene expression ratios are clustered and converted to colors, trends are easier to see than when the data were unorganized ratios.
- When all 6,200 genes are clustered this way, the boxes are reduced to thin lines.

Gene Expression Clusters (II)



Each row represents a different **gene**, and each column represents a different **time point**. Green indicates repression and red indicates induction of genes. Genes have been clustered so they are near other genes with **similar expression patterns**. Notice that the genes at the bottom were repressed for the first few time points.

"What if?" Questions

- What would happen if we delete a transcription factor gene?
- What would happen if we overexpress a transcription factor?
- These questions are "discovery science" and not hypothesis driven.
- Cell and molecular biology have been powered by hypothesis-driven for many years.
- With the advent of genomic methods, such as microarrays, people are asking different types of questions: "What if …?"

Anatomy Lesson of Dr. Nicolaes Tulp

1632 oil painting by Rembrandt Harmenszoon van Rijn

If Rembrandt was Around Today

Source: Carlos Cordon-Cardo, Columbia University