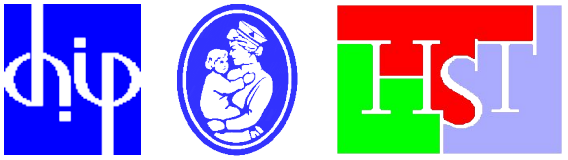


Gene expression and high-throughput molecular measurement technologies

September 29, 2005

Alvin T. Kho
Children's Hospital Boston
Dana-Farber Cancer Institute

Lecture Outline

- Revisit Central Dogma (CD)
- From DNA to Gene
- The concept of a Gene and its Expression.
- Gene Identification
- Quantifying Gene Expression
- High-Throughput Gene Expression Quantification Technologies
    - SAGE & Microarrays
    - How is CD important in this situation?
- User-friendly References
- Next time, analytic methods for analyzing high-throughput gene expression data

Central dogma of molecular biology (CD)

- Original CD (Crick, Nature 1958)
  - *The [CD] deals with detailed residue-by-residue transfer of sequential information … such information cannot be transferred from protein to either protein or nucleic acids.*
- Over-simplified (mis-interpreted) CD
  - DNA to RNA to Protein
  - DNA: C, G, A, **T** double strand
  - RNA: C, G, A, **U** single strand

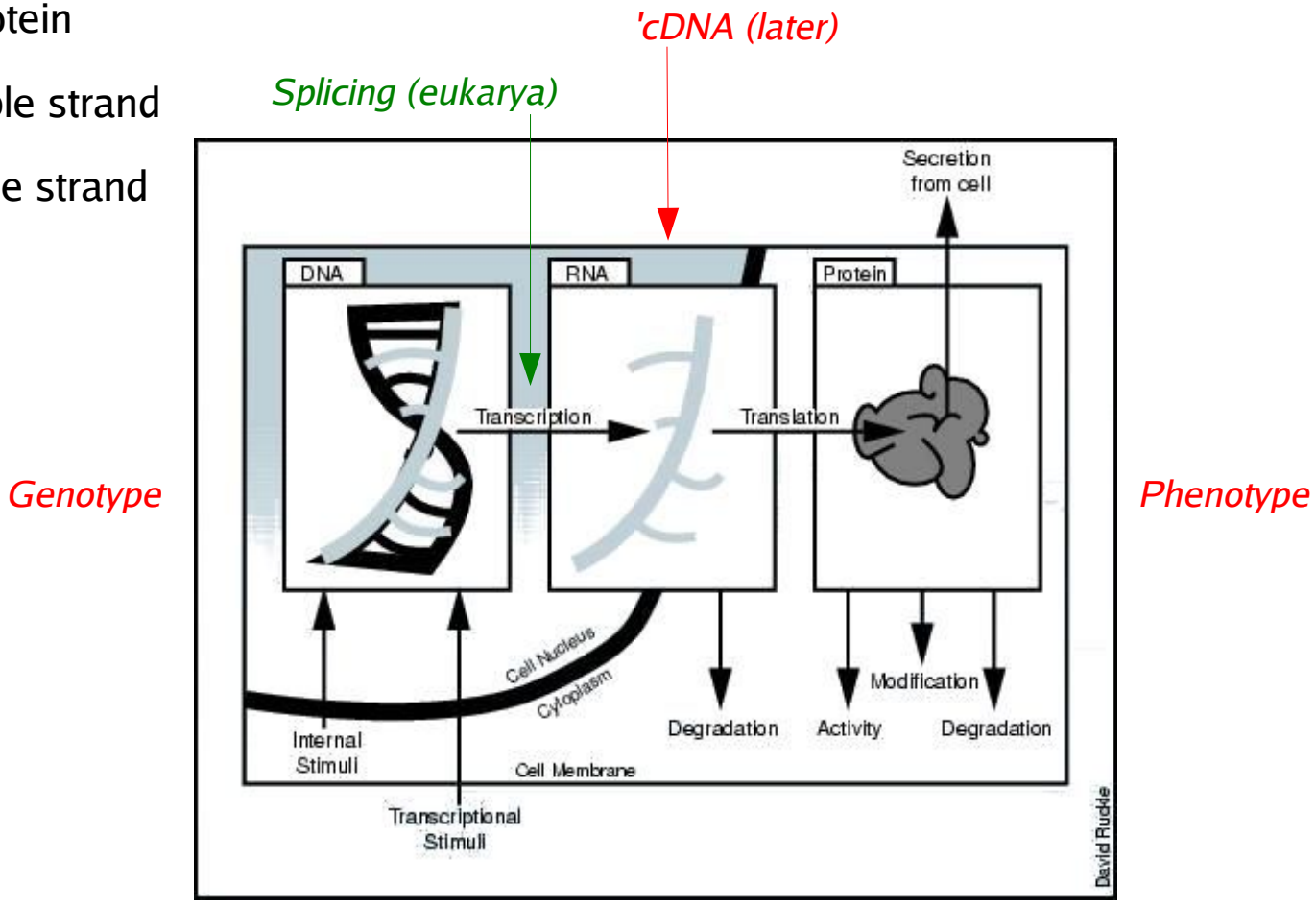


Figure 1.7 in Isaac S. Kohane, Alvin T. Kho, and Atul J. Butte. *From Microarrays for an Integrative Genomics.* Cambridge, MA: MIT Press, 2003, p. 28. ISBN: 026211271X Courtesy of MIT Press. Copyright 2003. Used with permission.

Exceptions to over-simplified CD

- Retroviruses
  - RNA into DNA via reverse transcriptase: E.g., avian sarcoma/leukosis viruses, mouse leukemia viruses, human immunodeficiency virus (HIV)
  - RNA -> (sometimes host) DNA -> RNA -> Protein
- Primitive RNA viruses
  - Error-prone RNA replication. E.g., hepatitis B, rabies, Dengue, Ebola, flu
  - Genetic RNA -> Intermediate RNA -> Protein
- Prions
  - Self-replicating proteins. E.g., Creuzfeldt-Jakob, kuru
  - Protein -> Protein
- DNA-modifying proteins
  - DNA-repair proteins: MCM family
- Retrotransposons (not really)
  - Mobile DNA (specifically *genetic*) segments in eukarya. Esp. plants, >90% wheat genome.
  - Retrotransposon DNA -> RNA -> DNA

DNA ↔ Gene ?

DNA is a molecule
but
what is a gene ?

Concept of a Gene

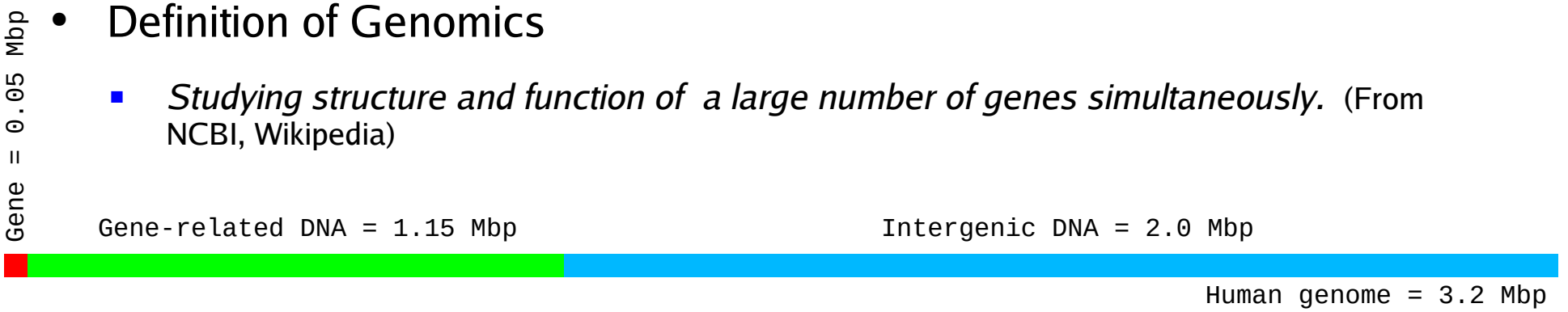
- Chronology of discoveries

  - 1854-65 *"Unit factors" of inheritance*, Gregor Mendel (Brno), Origin of Species 1859
  - 1869 *Nucleic acid / DNA isolated*, Johann Miescher (Tübingen)
  - 1952 *DNA (not protein) might be genetic material / agent*, Alfred Hershey & Martha Chase (Cold Spring Harbor)
  - 1953 *DNA is genetic material / agent (structurally makes sense)*, James Watson, Francis Crick & Rosalind Franklin (Cambridge, UK)
  - Aside: HOTHEAD / cress overturns Mendelian inheritance law? Lolle et al. Nature 2005 March 23 issue

- Definition of a Gene
  - *A fundamental physical and functional unit of heredity that is a DNA sequence located on a specific site on a chromosome which encodes a specific functional product (RNA, protein).* (From NCBI, Wikipedia)
  - A basic and complete unit of genetic information
  - Given an arbitrary (say human) DNA sequence X of length k-base pairs (k > 1 an integer), does this definition suffice to decide if X contains a gene segment?

Concept of a Gene

- Definition of Genome / Genotype
  - *Genome - the total genetic material in a living organism. Genotype – total genetic information in a living organism.* (From NCBI, Wikipedia)
- Content of Genome
  - Genes (~1.5% genome), gene-related DNA (~36% genome), intergenic DNA (~62.5% genome)
  - Exons (coding), introns (non-coding, eukarya). Coding = transmission into mRNA.
  - Pseudogene
  - Microsatellites
  - Genome-wide repeats. E.g., transposons, long/short interspersed nuclear elements
- Eukaryote vs. Prokaryote genomes
- Definition of Genomics
  - *Studying structure and function of a large number of genes simultaneously.* (From NCBI, Wikipedia)



Concept of a Gene

- Content of Genome Example

Figure removed due to copyright reasons.
Please see:

Figure 1.14 in Brown, Terence A., ed. *Genomes*. 2nd ed. New York, NY: Wiley-Liss, 2002. ISBN: 0471250465.

Intergenic DNA = Junk ? Probably not.
Muotri et al. (Nature 2005 16 June issue). L1 retrotransposon gene-hopping -> neuronal cell fate for rat neural stem cells.

Concept of Gene Expression

- Definition of Gene Expression
  - *The process by which information encoded in a gene is transcribed into RNA, and then typically into protein.* (From NCBI, Wikipedia)
- Gene expression is a function of:
  - Time and developmental stage
  - Space or location. E.g., brain vs. muscle
  - Response to (micro / global) environmental cues
  - Disease / cell state

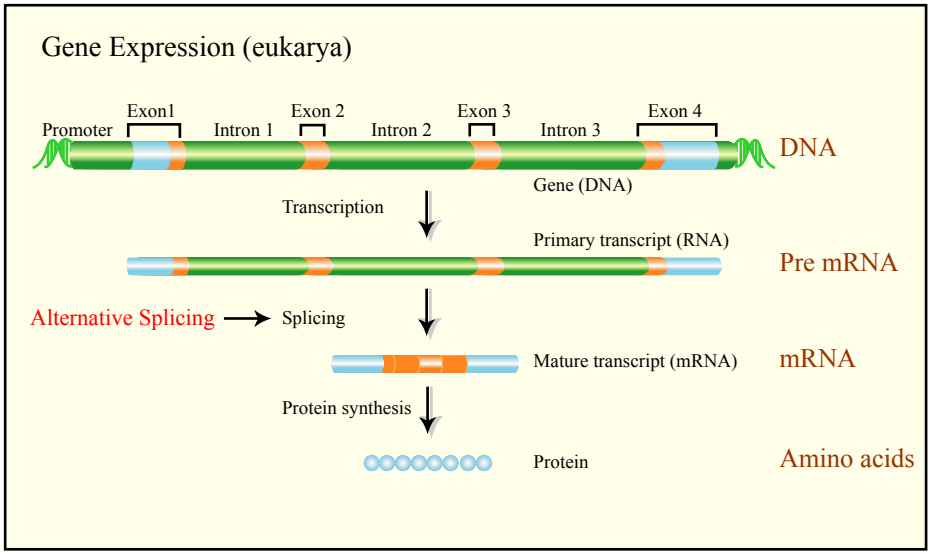


Figure by MIT OCW.

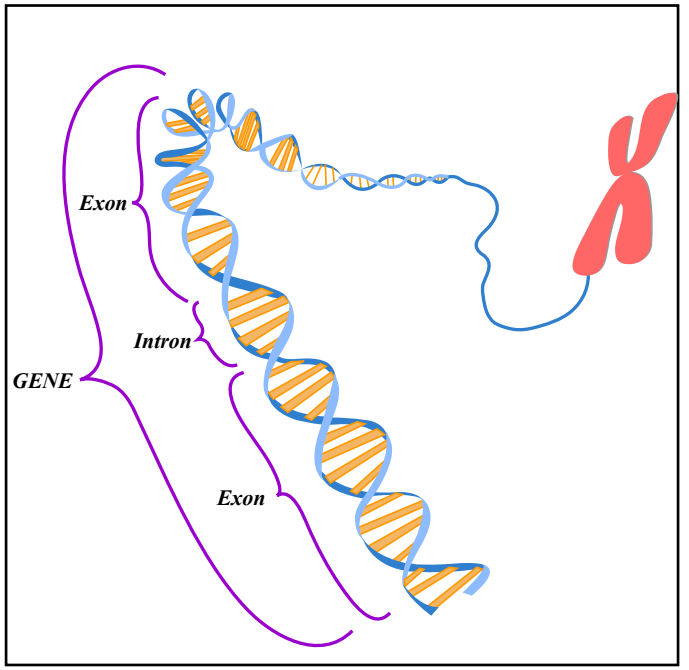
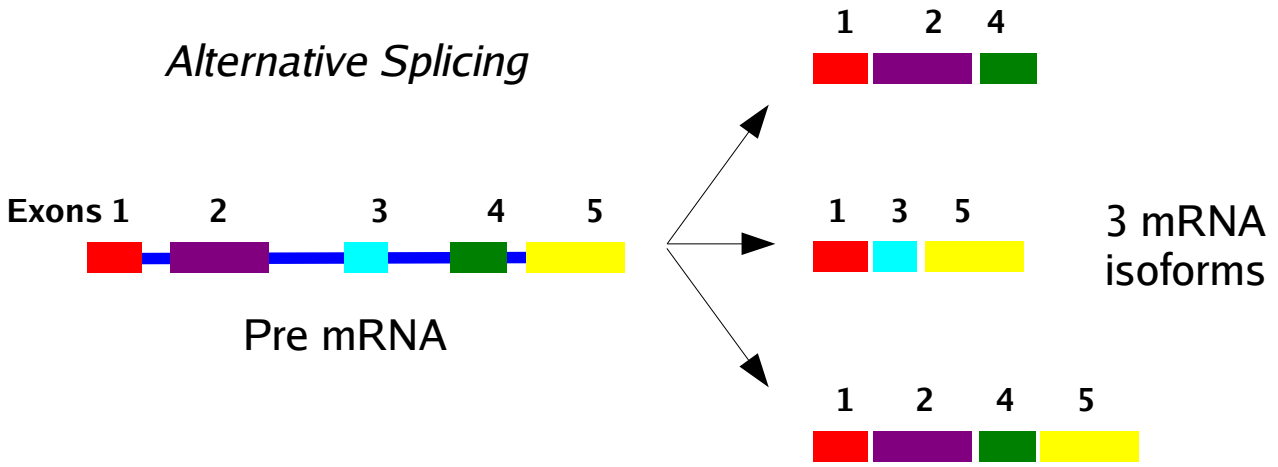


Figure by MIT OCW.

Concept of Gene Expression

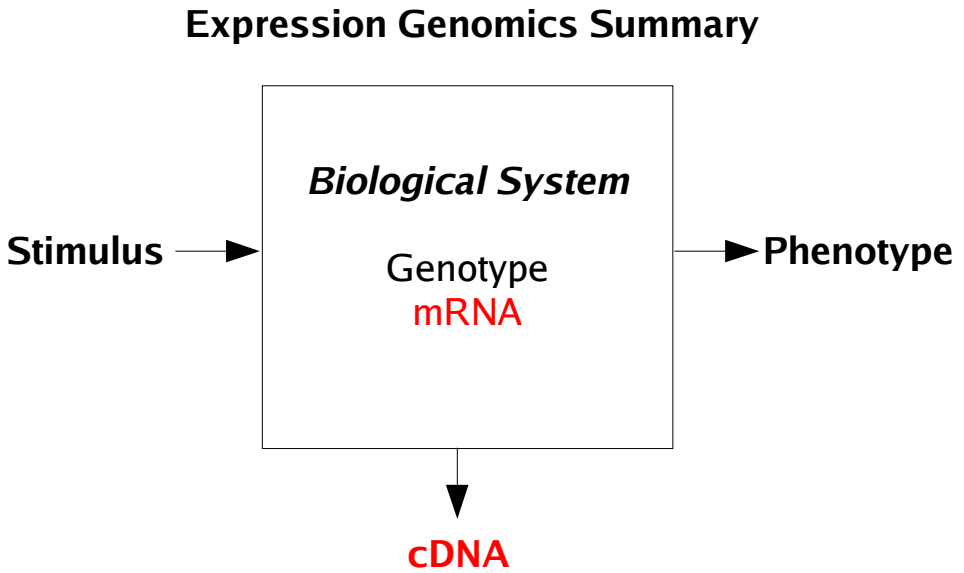
- Definition of Transcriptome
  - *All mRNA present in a cell at a particular time.*
  - Definition is organism and state specific (space, time, etc).



Different isoforms -> different function, i.e., different proteins translated.

Expression Genomics

- (Postulate) A biological system is characterized by its transcriptome profile (i.e., whole mRNA profile as cDNA)
  - Necessarily depends upon Central Dogma
- Use transcriptome profile to unravel the interactions of stimulus + genotype that engender phenotype
  - Limitations? Reductionism. Would proteins more accurately characterize a biological system?

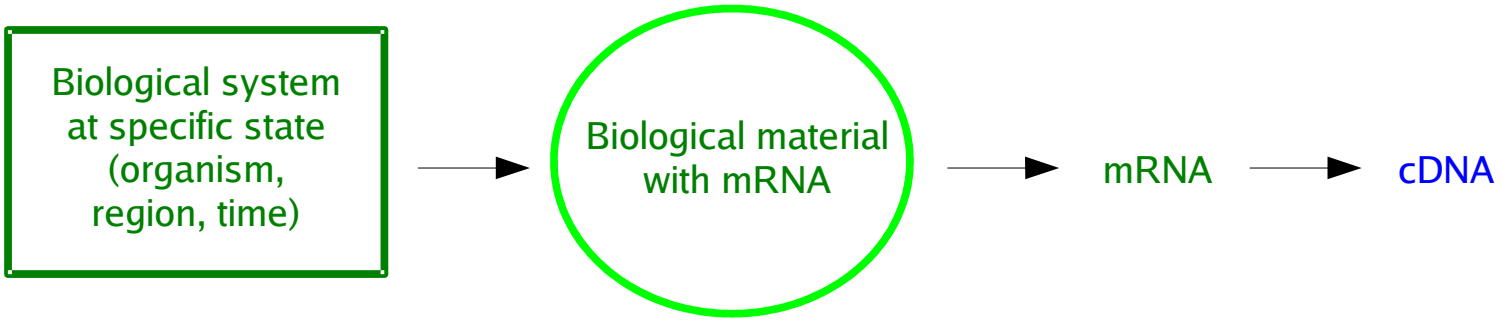


Expression Genomics

- Ways to measure gene expression (mRNA levels)
  - IDEA: Identify molecule. Then, Quantify molecule
  - (Identify) DNA libraries: Genomic, cDNA.
  - (Quantify) Low throughput: northern blot, RT-PCR.
  - (Quantify) High throughput: SAGE, microarrays. Our focus.

DNA libraries (Identification)

- DNA libraries are collections of clones DNA fragments: *Genomic and cDNA*
- Genomic libraries
  - Genomic DNA fragments representing (almost) entire genome of an organism.
- 'cDNA libraries
  - From mRNA, obtain cDNA
  - Contains only coding regions of genome (introns are gone) -> gives all possible expressed proteins
  - Sequence cDNA -> Expressed Sequence Tags (EST). Each EST is assigned a Genbank ID.
  - A gene on a chromosome may be "covered" by >1 EST's. Redundancy. Human genome -> over 4 million EST's. Estimate # of genes in human genome ~30K.
  - EST's screened. Set of EST's associated with a particular gene forms an EST cluster for that gene This cluster is assigned a Unigene ID.



cDNA libraries (Identification)

- cDNA example of a Unigene cluster of >1 ESTs. The human FoxP2 gene has 52 EST's in it's Unigene cluster (Hs.282787)

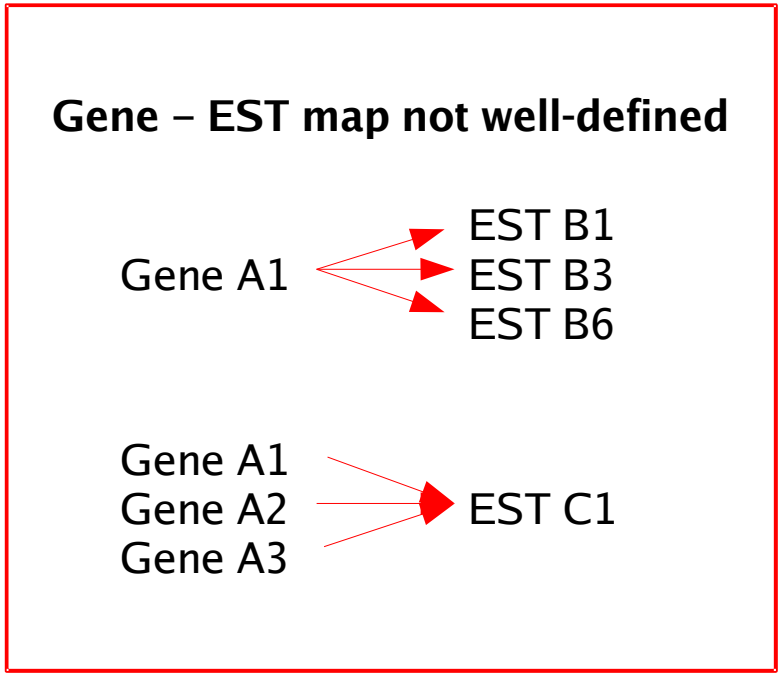
http://www.ncbi.nlm.nih.gov/UniGene

Genbank ID	Description	Tissue of Origin
BF700673.1	Clone IMAGE:4285527, 5' read	brain
T97069.1	Clone IMAGE:121181, 5' read	mixed
T96957.1	Clone IMAGE:121181, 3' read	mixed
BU521502.1	Clone IMAGE:6527367, 5' read	uterus
BQ948273.1	Clone IMAGE:6473507, 5' read	uterus
AL711700.1	Clone DKFZp686E0284, 5' read	muscle
BM725479.1	Clone UI-E-EJ0-aie-p-18-0-UI, 5' read	other
BM701645.1	Clone UI-E-EJ0-ahl-h-24-0-UI, 5' read	other
BI752226.1	Clone IMAGE:5192788, 5' read	brain
N31133.1	Clone IMAGE:265380, 5' read	skin
N21118.1	Clone IMAGE:265380, 3' read	skin
DN990126.1	Clone TC100653, 5' read	Whole brain
AV658847.1	Clone GLCFQG08, 3' read	liver
AV658824.1	Clone GLCFQE09, 3' read	liver
CV573230.1	Clone od33g10, 5' read	eye
CR738014.1	Clone IMAGp998F084735_;_IMAGE:1929991, 5' read	lung
BP871788.1	Clone HKR01979	embryonal kidney
BE068078.1	Clone (no-name)	mammary gland
CD637513.1	Clone (no-name)	other
CD637512.1	Clone (no-name)	other
CD637511.1	Clone (no-name)	other
CD637510.1	Clone (no-name)	other
CD637509.1	Clone (no-name)	other
CD637508.1	Clone (no-name)	other
BX481950.1	Clone DKFZp686D03228, 5' read	muscle
CD001942.1	Clone (no-name)	other
CB410738.1	Clone (no-name)	other
CB410682.1	Clone (no-name)	other
CB410681.1	Clone (no-name)	other
BX280996.1	Clone IMAGp998G13581_;_IMAGE:265380	skin
CB118125.1	Clone B1T694954-5-A03, 5' read	brain
T06261.1	Clone HFBDR02	brain
AI459612.1	Clone IMAGE:2152081, 3' read	colon
AI624789.1	Clone IMAGE:2231455, 3' read	uterus
AI798932.1	Clone IMAGE:2348762, 3' read	mixed
CK430225.1	Clone oj46f12, 5' read	eye
CV569620.1	Clone od07e09, 5' read	eye
BF678535.1	Clone IMAGE:4250207, 5' read	prostate
BG722650.1	Clone IMAGE:4826916, 5' read	testis
BI495413.1	Clone IMAGE:2539657, 3' read	other

...

cDNA libraries (Identification)

- Unigene cluster sizes
  - Not every Unigene cluster is mapped to a known gene. Different similarity levels (sequence matching)
  - Estimated # of genes in human genome ~30K
  - Human example



Cluster Size	# of Unigene Clusters
1	~10,400
2	7,100
3-4	6,800
5-8	5,300
9-16	3,800
17-32	3,100
~500-1,000	1,500
~2,000-4,000	130
~8,000-16,000	12
~16,000-30,000	3

From J. Pevsner, Bioinformatics & Functional Genomics, 2005

Comparing between cDNA libraries

- Binary (present / absent) comparison between cDNA libraries derived from various tissue systems
  - Digital Differential Display (DDD). Statistical significance (p value) is assessed via Fisher exact test (non parametric). Contingency table. Null hypothesis: # of sequences for a given gene X is the same in the two cDNA libraries.
  - Not quantitative
  - Limitations: (1) Sequencing error (2) Depth of sequencing (3) tissue of origin contamination (4) library construction bias.

	Gene X	**Genes other than Gene X**
cDNA library A	# EST's mapped to Gene X	# EST's not mapped to Gene X
cDNA library B	# EST's mapped to Gene X	# EST's not mapped to Gene X

http://www.ncbi.nlm.nih.gov/UniGene/fisher.shtml

SAGE (Quantification)

- Serial Analysis of Gene Expression (SAGE)
  - Have a SAGE library. Essentially a (bijective) map between SAGE tags and genes & EST's
  - Obtain mRNA to construct corresponding cDNA.
  - From each cDNA transcript, cut a short sequence tag (SAGE tag) of 10-14 bps from a specific position (3'-end typically) that will uniquely identifies that transcript.
  - Tags have uniform length.
  - Concatenate all SAGE tags into one concatamer -> clone and sequence.
  - # of times a particular tag is observed = expression level of particular gene (mRNA)
- *Details@ www.bioteach.ubc.ca/MolecularBiology/PainlessGeneExpressionProfiling*

Figures removed due to copyright reasons.
Please see www.sagenet.com.

SAGE (Quantification)

Example of SAGE result: 3 types of transcripts relative to SAGE library

Table removed due to copyright reasons. Please see: www.embl-heidelberg.de/info/sage

SAGE (Quantification)

- Limitations of SAGE:
  - Tag specificity. Short SAGE tag size may lead to identification problems. 1 tag mapping to >1 genes is a problem.
  - Restriction enzyme action variability. Each SAGE tag must have constant length, otherwise problems arise in sequencing concatamer. Restriction enzyme may not yield tags of uniform length. Not all mRNA species have the same enzyme recognition sequence, plus temperature dependent.
  - What is appropriate **control** or **reference** system for comparison? This is really a more general problem that we will see as we explore microarrays and other high-throughput assaying technologies.

DNA microarrays (Quantification)

- What is a DNA microarray (chip)
  - A collection of single-stranded DNA (known sequences of genes / EST's) anchored at one end onto a substrate, typically in the form of a gridded array. Different DNA species placed on separate grids. ssDNA fragments (called probes), not entire gene sequence is placed. Why?
  - Evolved from southern blots for DNA. Exploits parallelism
  - Mechanistic principle: Nucleic acid complementarity – i.e., hybridization of complement partners, A $\leftrightarrow$T, G $\leftrightarrow$ C
  - 'ssDNA on chip will hybridize to complementary strand in solution (derived from biological system under investigation). Complementary strand is fluorescent labeled.
  - Assumption: Fluorescence is proportional to gene expression level
- Microarray oligo probe design technicalities
  - GC content: Hybridization (binding) energy for GC > AT. Introduces non-linearity in hybridization rate for cDNA species with different %CG content. General problem.
  - Distance from 3' end. General problem.

DNA microarrays (Quantification)

- Primarily 2 types of DNA microarrays
  - *Spotted*: (Pat Brown, Stanford). Robot attaches down previously prepared ssDNA probes of order $10^{2\text{-}3}$ bp long on substrate. Customizable -> heterogeneous (noisy)
  - *Oligonucleotide*: (e.g., Affymetrix). Photolithography. Typically standardized manufacturing and shorter (relative to spotted microarrays) length oligos placed.

Figure removed due to copyright reasons.
Please see:

Southern, et al. "The Chipping Forcast." *Nature Genetics Supplement* 21, no. 1 (January 1999.)

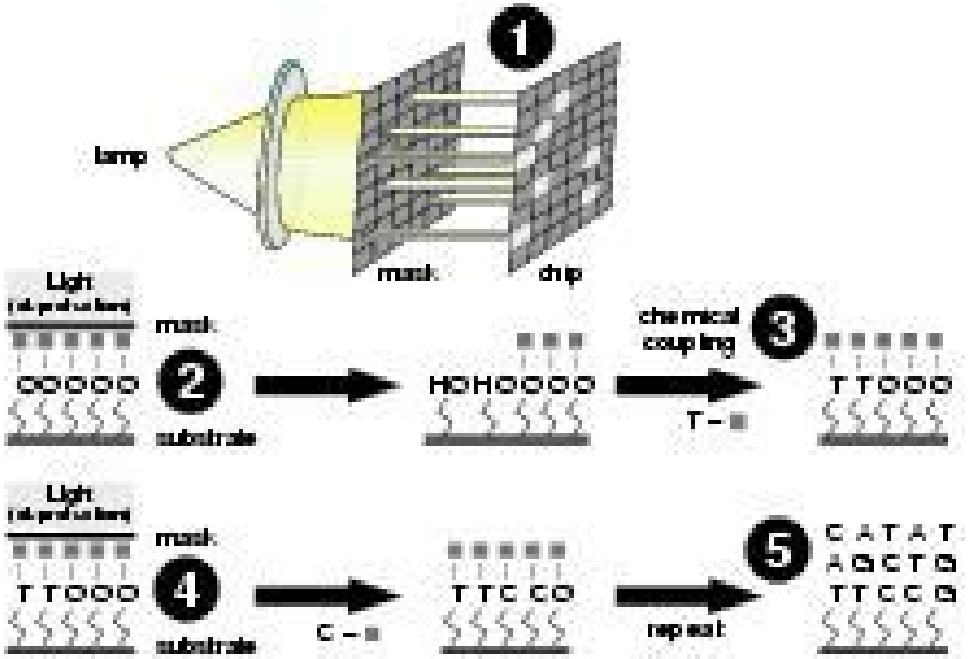


Figure 3.3 in Kohane, Isaac S., Alvin T. Kho, and Atul J. Butte. *From Microarrays for an Integrative Genomics*. Cambridge, MA: MIT Press, 2003, p. 79. ISBN: 026211271X. Courtesy of MIT Press. Copyright 2003. Used with permission.

DNA microarrays (Quantification)

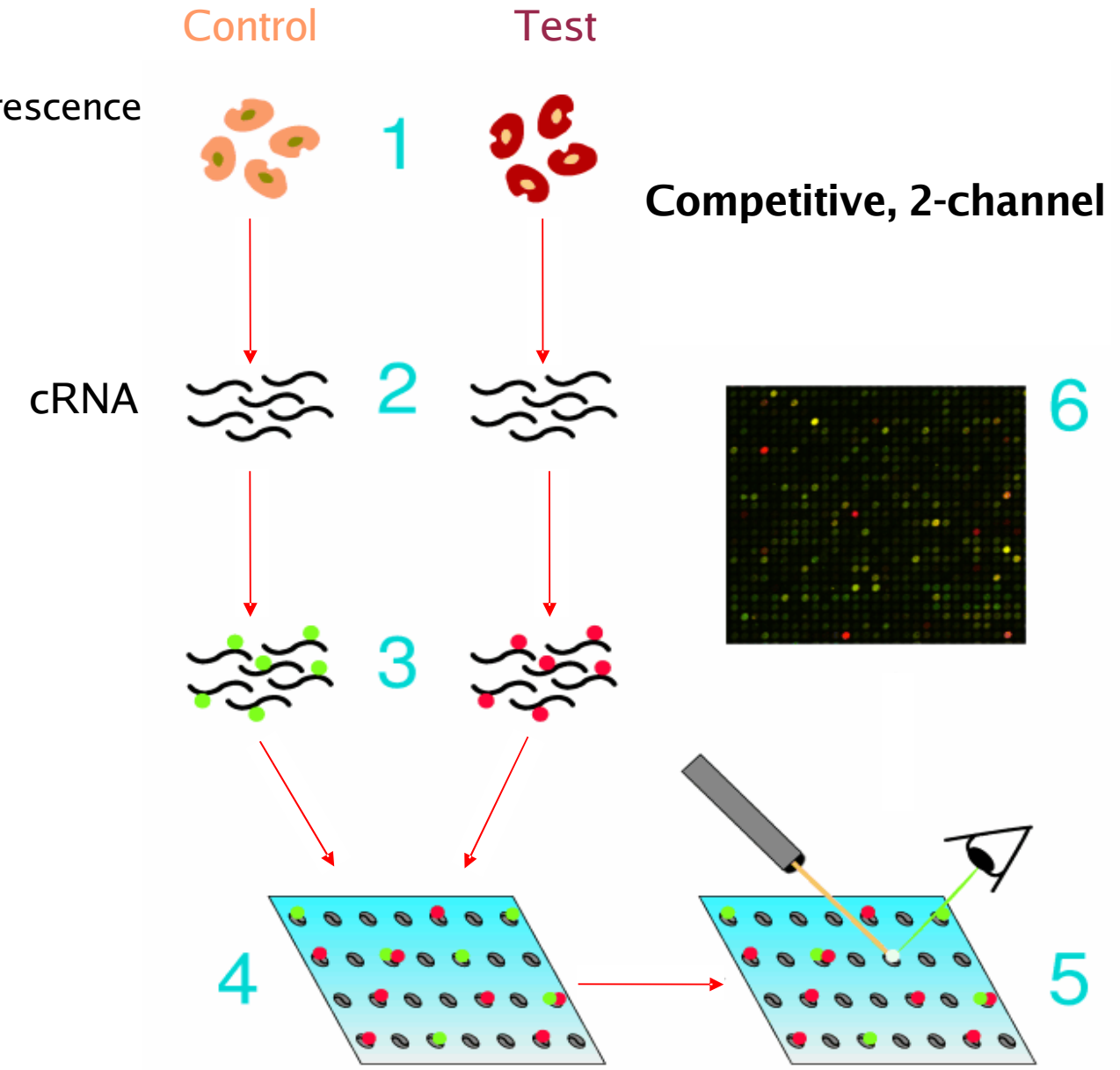
- Stages of a typical microarray experiment
  - Experimental design involving biological system under investigation. Replicates – biological and measurement / technical
  - RNA and (target) probe preparation: Extract mRNA. Convert (to ss cDNA typically). Label with fluorescence.
  - Probe hybridization.
  - Fluorescence image analysis
  - Microarray data analysis (post image)

Figures removed due to copyright reasons.
Please see:

Pevsner, Jonathan. *Bioinformatics and Functional Genomics*. Hoboken, NJ: Wiley-Liss, Inc., 2003. ISBN: 0471210048.

DNA microarrays (Quantification)

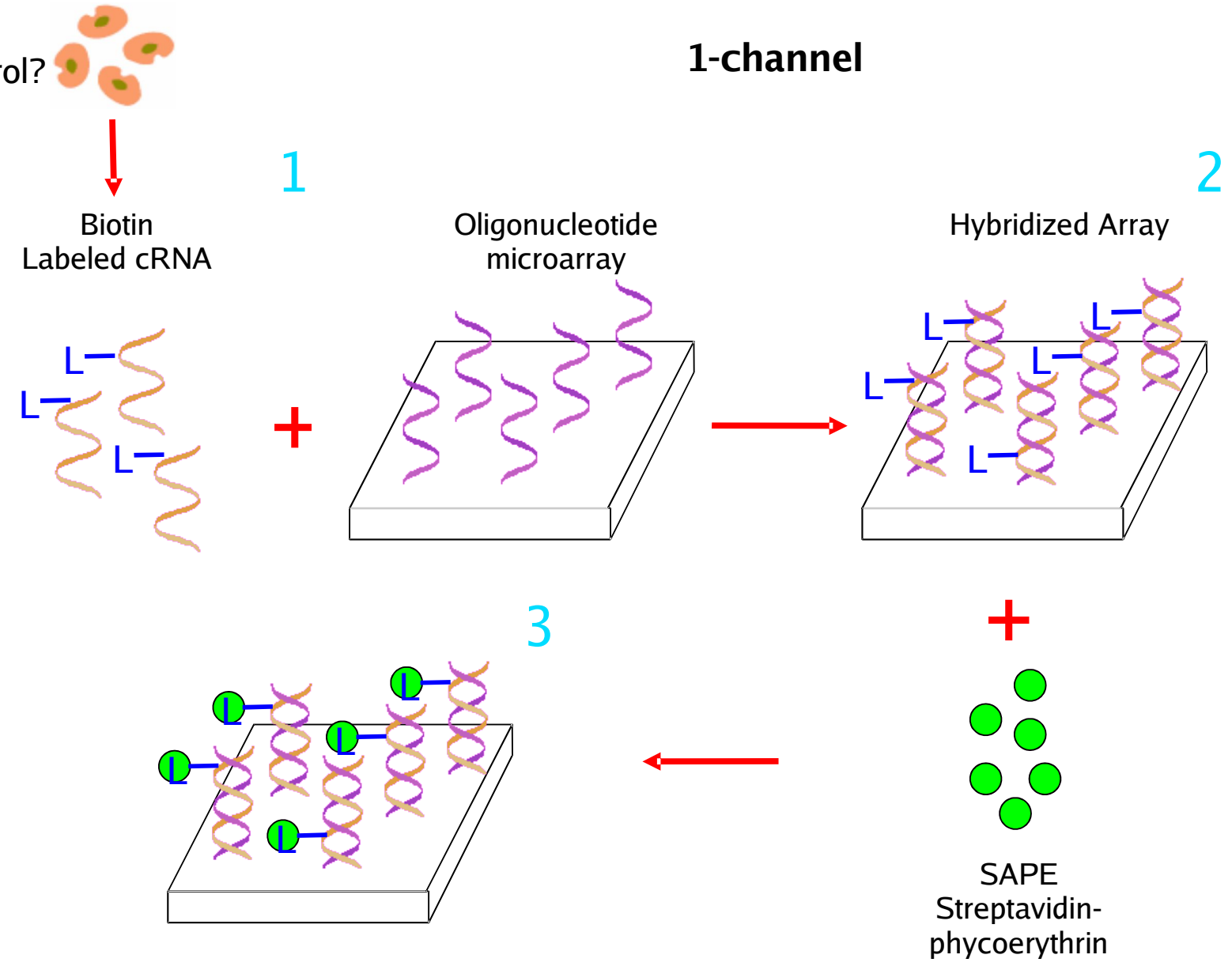
- 1 channel vs. **2 channel** microarray usage
  - 2 channel
  - Internal control for fluorescence



Courtesy of Jeremy Buhler. Used with permission.

DNA microarrays (Quantification)

- **1 channel** vs. 2 channel microarray usage
  - 1 channel
  - Internal control?



DNA microarrays (Quantification)

- Typical usage and prototypical experimental designs
  - Comparing two groups. E.g., tumor/cancer vs. “normal” tissue
  - Time course (dosage-level) profiling
  - Suitable reference state is a challenge. General problem
  - Typical “statistically-sound” experimental design principles apply. Sample pooling mRNA? For 2-channel experiments: Swap dye.
- Microarray experiment (biological) assumptions
  - Central Dogma holds. Specifically that mRNA transcription is proportional to its associated protein translation
  - All mRNA have identical lifespans. Uniform degradation rate.
  - All cellular activities are entirely characterized by the transcriptome.

DNA microarrays (Quantification)

- Generalization of chip parallelism / complementarity principle
  - Protein microarrays. Identify protein targets, e.g, other proteins (protein-protein interaction), mRNA, other bio-active small molecules.
  - Tissue microarrays. Paraffin blocks of distinct biological tissue cores. Simultaneous histologic analysis, immunohistochemical (protein) & in situ (mRNA) analyses.
  - Reverse transfection microarrays. 'cDNA probes on grid with a cell culture on top. Cells assimilate probes.
- Limitations
  - Probe specificity. Cross (RNA) species hybridization, promiscuous probes.
  - RNA degradation
  - “Noise” . Next time.

DNA microarrays (Quantification)

- Reader-Friendly References
  - T.A. Brown. **Genomes**. 2nd edition. Wiley-Liss, 2002. This is a comprehensive, highly-readable biologically oriented text on genome. This book is FREE online at www.ncbi.nlm.nih.gov/books
  - And of course, the NCBI has a great many user-friendly and FREE primers on genomic biology: **www.ncbi.nlm.nih.gov/**. Look at their Coffee Break section, and library of text-searchable books www.ncbi.nlm.nih.gov/books
  - IS Kohane, AT Kho, AJ Butte. **Microarrays for an Integrative Genomics**. MIT Press 2003. (shameless self-promotion)
  - Good luck.