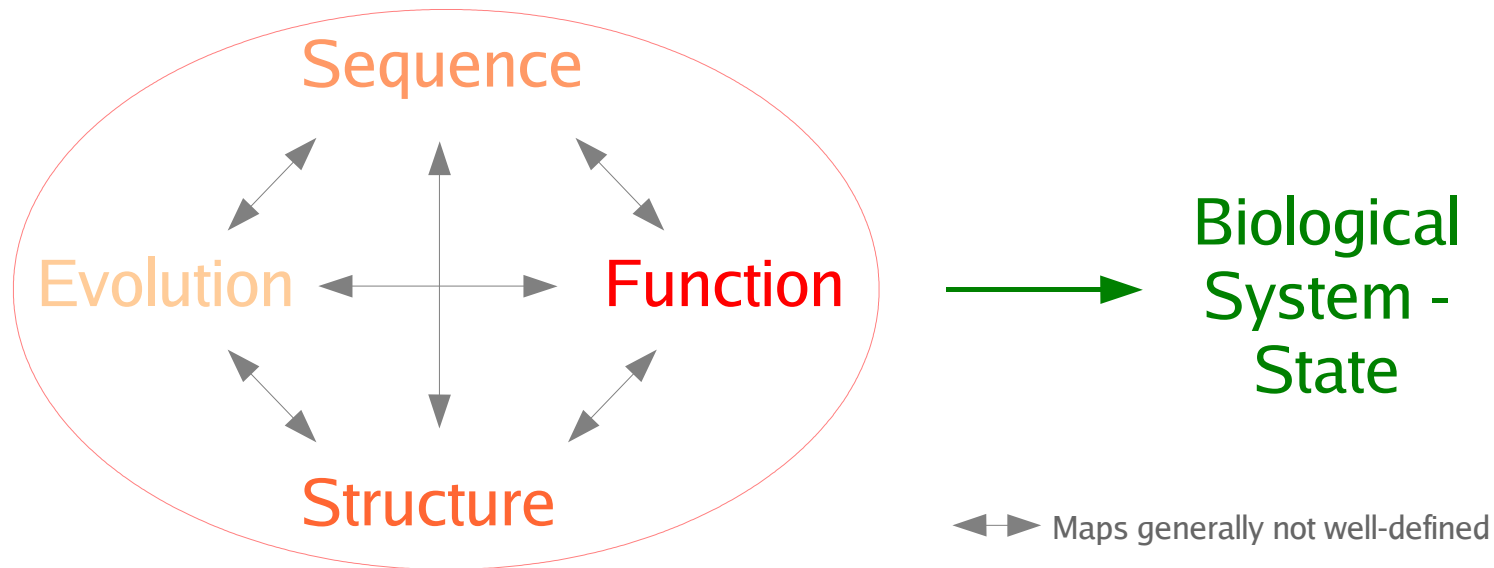


Prologue: HST 508 tetralogy



- Different scales of ***function*** for a given bio molecule X
 - **Chemical / physical** (microscopic scale): binds another molecule, catalyzes a molecular reaction, etc.
 - **Biological** (macroscopic scale): leads to a phenomenologic / phenotypic transformation
 - All scales in between the above (mesoscopic)
 - X may have >1 function, across / within these scales
 - A general / naïve test for ***function***: Perturb X in native system and observe what happens at all scales

Prologue: example of *functions* of a gene (product), 2 archetypal questions



- Eg. mutation (frameshift, mis-sense / non-synonymous) of methyl-CpG-binding protein 2 (**MECP2**, Xq28) -> Rett syndrome, a progressive, X dominant neurologic developmental disorder. Phenotype incl. autism, dystonia, short, etc. Typ. fatal in males (major encephalopathy). Females -> somatic X mosaicism.
 - *MECP2 chemical function*: binds methylated DNA -> repress transcription from methylated gene promoters
 - *MECP2 biological function*: embryonic development
 - Mutation (truncating frameshift, mis-sense) of cyclin dependent kinase-like 5 (**CDKL5**, Xp22) leads to almost similar phenotype. *CDKL5 chemical functions*: ATP binding, protein serine / threonine kinase activity, nucleotide binding
- 2 archetypal questions in functional genomics
 - What function does a given molecule X have in a specific biological system - state?
 - Which molecules (their interactions) “associate” with / “underwrite” a given biological system - state?

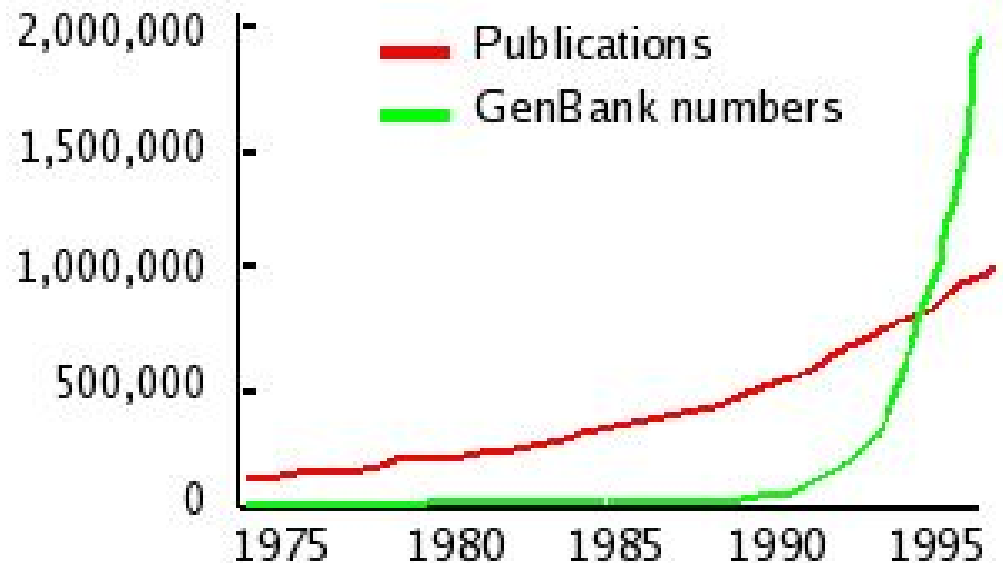
Module 4: Functional Genomics (FG) lecture 1 outline

- FG definition. 2 iconic studies
- Review basic functional concepts: gene, expression, epigenetics, uni-directional transfer of *genetic* information (“central dogma”)
- Survey of parallel high-throughput bio quantification technologies
 - Scalable detection principles: sequencing short oligomers, nucleotide complementarity
 - Representative technologies: SAGE, microarrays. Assumptions / Pros / Cons.
 - Technical generalizations
- Transcriptome studies: basic caveats / assumptions
- Shift in perspective / way to think about biological problems

FG definition

- Functional genomics is the deconstruction of the genome to
 - Ascribe *function* to genes & non (protein) coding genomic elements / NCGE's – different levels of *function*
 - Characterize interactions between genes & NCGE's
- Using the vast genomic evolutionary, sequential, structural, expression information.
- Leverage on synergy of diverse data modalities

Proxies for
"Function" 
Raw genomic info 



Graph adapted from Figure 1 of:
Ermolaeva, et al. "Data management and analysis for
gene expression arrays." *Nat Genet* 20, no. 1
(Sep 20, 1998): 19-23.

FG: Iconic study #1 (Alizadeh et al, Nature 2000)

- Classical lymphoma histopath classification unify different morphologic subtypes into 1 group, eg. diffuse large B cell lymphoma (DLBCL)
- Transcriptome-scale profiles of 96 lymphatic malignancies (mostly DLBCL, CLL, FL) and normal tissue. All DLBCL patient *de novo* and biopsy samples obtained pre-treatment. Questions:
 - Identify distinct molecular portraits for DLBCL malignancies
 - Identify DLBCL malignancy subtypes new to current classification system
 - Relate each malignancy to distinct stage of normal B cell development
- Hierarchical clustering using full transcriptome features reveal heterogeneity within DLBCL subgroup

FG: Iconic study #1 (Alizadeh et al, Nature 2000)

Hierarchical clustering of entire dataset using full transcriptome features – reveals ordered heterogeneity among samples

Figure removed due to copyright reasons.

Please see figure 1 from:

Alizadeh, et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling."

Nature 403, no. 6769 (Feb 3, 2000): 503-11.

GC B-cells genes relevant, specific to B-cell development, use this subset of features to re-cluster DLBCL

FG: Iconic study #1 (Alizadeh et al, Nature 2000)

Hierarchical clustering of DLBCL
using B-cell level relevant subsets
of transcriptome



Feature-induced
Regularities
within DLBCL



Figure removed due to copyright reasons.
Please see figures 3, 5a, and 5c from:
Alizadeh, et al. "Distinct types of diffuse large B-cell lymphoma identified
by gene expression profiling." *Nature* 403, no. 6769 (Feb 3, 2000): 503-11.



Reveal 2 distinct molecular subgroups
with significantly different survival
outcomes

FG: Iconic study #1 (Alizadeh et al, Nature 2000)

- *“It is important to note that considerable gene expression heterogeneity exist within each subgroup, and no single gene in either of these clusters was absolutely correlated in expression with the DLBCL subgroup taxonomy.”*
- Talking points from study #1:
 - New diagnostic subcategory of DLBCL missed by morphologic + immunohistochemical (with a few markers) analyses marred by irreproducibility.
 - New prognostic tool and corresponding therapeutic opportunities.
 - Hypothesis generation for basic biology of DLBCL. [1] Now we have a clearer sense of the granularity of DLBCL cases. [2] Mechanistic differences between these 2 DLBCL subgroups?
 - Genomic data can be fruitfully exploited without mechanistic functional assignment

FG: Iconic study #2 (Arkin et al, Science 1997)

- Reverse engineer the reaction network architecture of early glycolysis from metabolite in(2) / out(8) put time series – using time-lagged correlation + multi-dimensional scaling

Classical ly determined pathway of early glycolysis

In / Out put time series

Figures removed due to copyright reasons. Please see figures 1 and 2 from:
Arkin, Adam, Peidong Shen, and John Ross. "A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements." *Science* 277 (August 29, 1997): 1275-1279.

FG: Iconic study #2 (Arkin et al, Science 1997)

CMC/MDS predicted pathway of early glycolysis

Causality arrow inferred
from temporal ordering

Classically determined pathway of early glycolysis

Figures removed due to copyright reasons. Please see figures 1 and 5 from:
Arkin, Adam, Peidong Shen, and John Ross. "A Test Case of Correlation Metric Construction
of a Reaction Pathway from Measurements." *Science* 277 (August 29, 1997): 1275-1279.

FG: Iconic study #2 (Arkin et al, Science 1997)

- “[Metabolite input] ranges represent the extreme “physiological” concentration attained by these species.”
- Talking points from study #2:
 - Not all metabolites known to be involved / produced in the process were measured
 - Certain interactions between species that were measured were not resolved
 - Analysis is sensitive to initial conditions, eg. pH, temperature, initial inflow species concentrations.

FG: Review concept of a gene

- DNA is a physical molecule. **Genome** = total cellular DNA. What is a gene?
 - 1854-65 “Unit factors” of inheritance, Gregor Mendel (Brno)
 - 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)
 - 2005 Lolle et al. (Nature March 23 issue) epigenetic (non-Mendelian) recovery of HOTHEAD gene in cress
- Definition of a gene (NCBI)
 - *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 (eg. RNA, protein)*

Gene = 0.05 Mbp

Gene-related DNA = 1.15 Mbp

Intergenic DNA = 2.0 Mbp

Human genome = 3.2 Mbp

FG: Review concept of a gene

- Example zoom into a contiguous subset of the genome

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. Example 1: Muotri et al. (Nature 2005 16 June issue). L1 retrotransposon gene-hopping -> neuronal cell fate for rat neural stem cells.

FG: Review concept of a gene

- What's the non-gene stuff in the genome? (eukarya)
 - **Genes** (~1.5% genome. Eg. protein coding **exons**), **gene-related DNA** (~36% genome. Eg. non-coding **introns** – eukarya, **pseudogenes**), **intergenic DNA** (~62.5% genome. Eg. **microsatellites**, **genome-wide repeats**). Coding = transmission into mRNA.
 - **Genome-wide repeats**. E.g., transposons, long/short interspersed nuclear elements
- Eukaryote vs. Prokaryote (operons, no introns) genomes
 - C

FG: Review concept of gene expression

■ Definition of gene expression (NCBI)

- *The process by which information encoded in a gene is transcribed into RNA, and then typically into protein.*

■ Gene expression is a function of cellular *state*

- Time – eg, developmental stage
- Space – eg, organ. tissue
- Other state variables – eg, disease, environmental cues

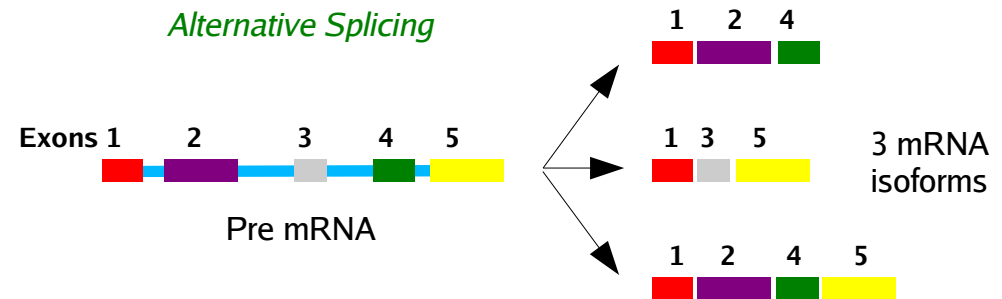
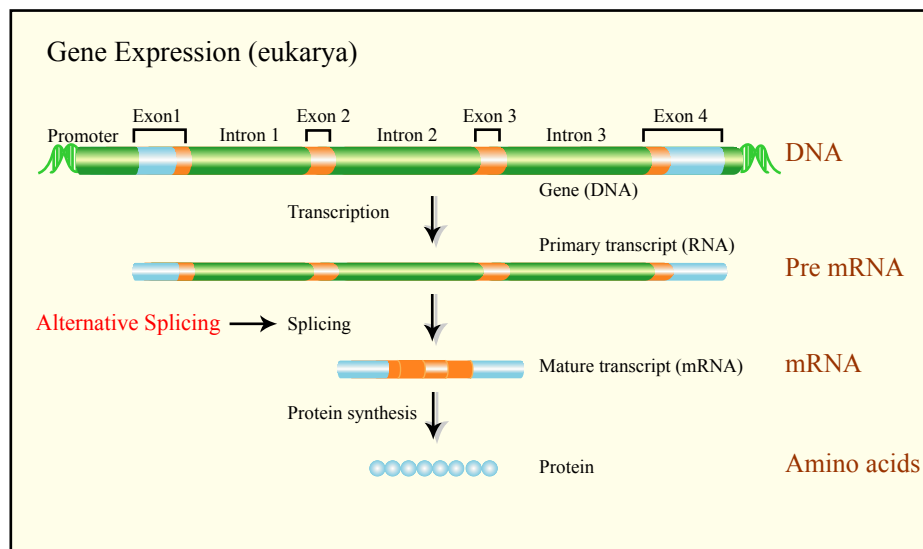
Sequence-wise

Transcriptome \subset Genome

“Function”-wise

Transcriptome $\not\subset$ Genome

■ **Transcriptome** = all mRNA present in a cell at a particular state



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

Figure by MIT OCW.

FG: *Functional* / expressed elements of the genome

- Recall different levels of *function* (chemical, biological, etc). (protein)

Coding and non-coding RNA categories:

- Messenger RNA = protein coding transcripts, typically high degradation rate
- Transfer RNA = transfer aa to polypeptide chain during translation
- Ribosomal RNA = primary (structural) constituent of ribosomes
- Small nuclear RNA = RNA splicing, telomere maintenance, form snRNP
- Small nucleolar RNA = chemical modification (eg. methylation) of rRNA
- Guide RNA = RNA editing in protozoa
- Micro RNA = RNA interference at post/pre-transcription

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.

FG: Epigenetic processes affecting *function*

■ Epigenetic definition

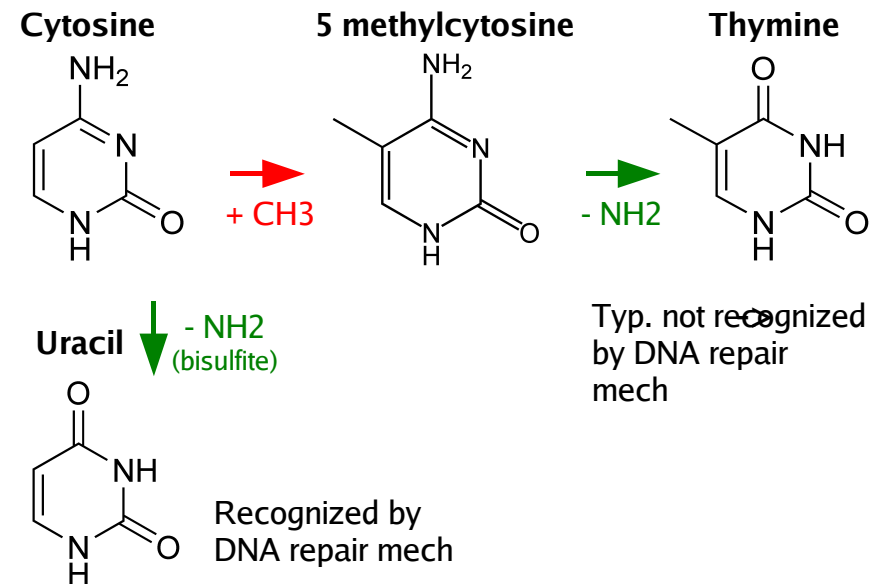
- Heritable change in gene function *without* nuclear DNA sequence change
- Selective gene in/activation within an organism. Eg. X mosaicism, imprinting, cell fate determination

■ A key process: methylation (H → CH₃) of DNA or protein

- **DNA methylation** typically on CpG sites, catalyz. by DNA methyltransferases (DNMT's)
- CpG sites (5'CG3') < expected prob. given uniform random base dist. due to DNMT's.
- CpG islands: >200bp, GC% > 50%, obs/exp CpG ratio > 0.6, high density near eukarya 5' gene promoters. Tend to be un-CH₃.

- +CH₃ @ CpG islands modulates correp gene's expression.
- Protein methylation typically on arginine/R, lysine/K, catalyz by protein CH₃-transfs. Esp. **histones** (post translational modif.) modulate local gene activity

- Bisulfite sequencing assay: bisulfite -NH₂ but not 5-methylcytosine



FG: Epigenetic processes affecting *function*

■ Functional expression of methylation

- Embryonic development: 1 to 8 cell stage, eukaryotic genome de-CH₃. 8-cell to morula (~32 cell blastomere), *de novo* +CH₃. By blastula stage, +CH₃ complete. DNA CH₃-transf knockout embryos die at morula stage.
- Environment factors (cellular stress -> polyamines) modulate CH₃ pattern postnatal development
- Imprinting: Prader-Willi/Angelman (chr 15q11.2-11.3)

- CH₃ state as cancer, neoplastic biomarkers

Gene	+CH ₃ of CpG islands in promoter
v-abl Abelson murine leukemia viral oncogene homolog 1	50-100% chronic myelogenous leukemia
chondroitin sulfate proteoglycan 2	Colorectal cancer
deleted in bladder cancer chromosome region candidate 1	50% bladder cancer
endothelin receptor B	60-70% prostate cancer
Wilm's tumor 1	90% breast cancer
pi-class glutathione S-transferase	80-100% prostate cancer

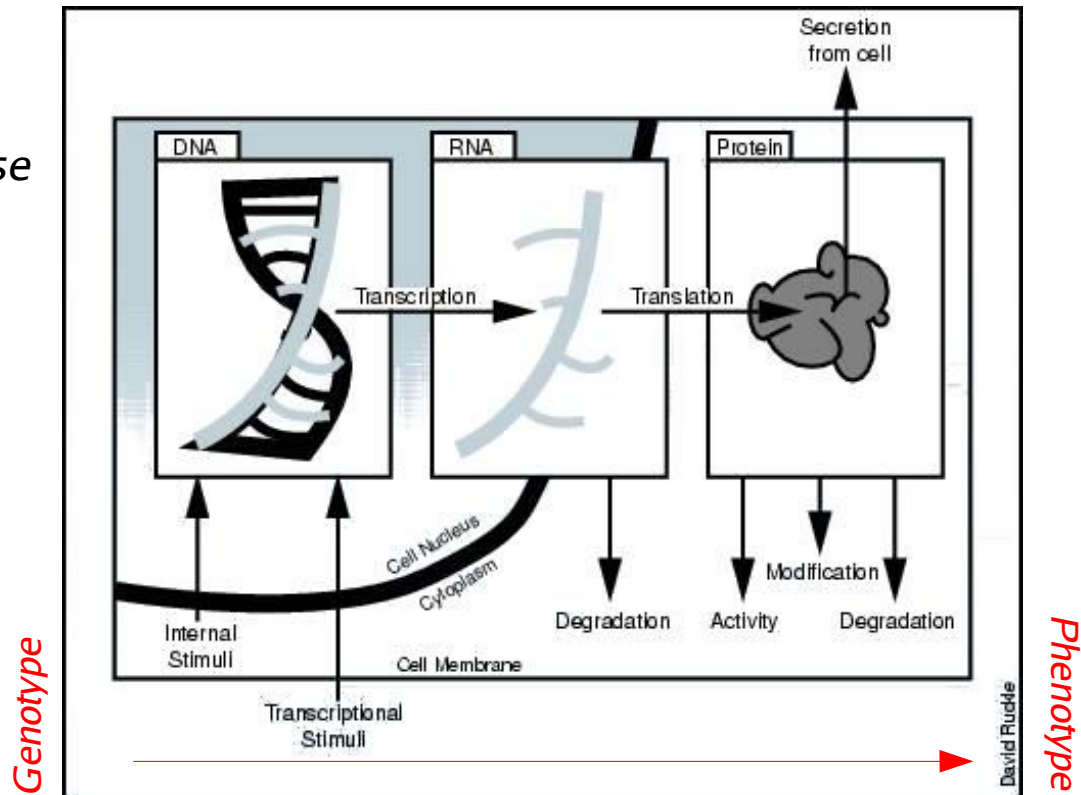
■ Other epigenetic disorders

- Mitochondrial diseases (matrilineal) – typically metabolic

FG: Uni-directional transfer of *genetic* information

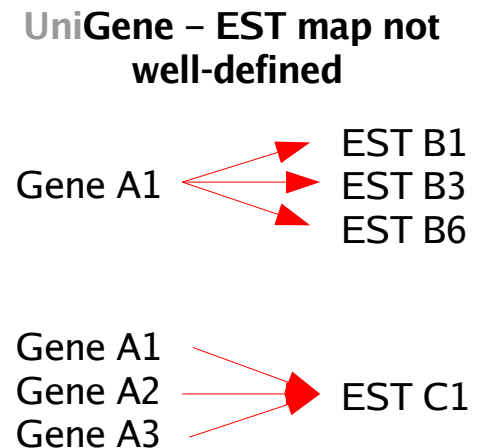
- Original statement, “central dogma” (CD) of molecular biology
 - *The [CD] deals with detailed residue-by-residue transfer of sequential information ... such information cannot be transferred from protein to either protein or nucleic acids.* [Crick, 1958]
- Over-simplified (mis-interpreted) CD
 - *DNA to RNA to Protein*
 - *(faux) Exceptions: Retroviruses (by reverse transcriptase), DNA modifying proteins,*

Figure from:
Kohane, Isaac S., Alvin T. Kho, and Atul J. Butte.
Microarrays for an integrative genomics.
Cambridge, Mass: MIT Press, 2003. ISBN: 026211271X.



FG: Expressed sequence tags (EST), cataloging the transcriptome

- Recall, ***transcriptome*** = all mRNA present in a cell at a particular state, organism-space-time specific
- Identification / characterization
 - Genomic libraries: DNA fragments of (near) total genome @ *specific state*
 - cDNA libraries: mRNA fragments (no intron) -> cDNA fragments -> sequence -> expressed sequence tags (EST's), GenBank ID#
 - 1 gene “covered” by >1 EST's. Eg. human genome >4M EST's, ~30K genes
 - Screen EST's -> EST's assoc with a particular gene form a Unigene cluster
 - Differential comparison between cDNA libraries: Binary analysis (present/absent). H_0 : # of seq for a given gene X is the same in two libraries. Prob test: Fisher exact. Limitations: sequencing error + depth, tissue of origin contamination, library construction bias



FG: Concept of Unigene cluster

- Human FoxP2 gene has 52 EST's in it's Unigene cluster (Hs.282787)

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

Human Unigene clusters Oct 2005

#	Category
155,852	mRNAs
4,514	Models
48,605	HTC
1,574,398	EST, 3'reads
2,431,310	EST, 5'reads
1,114,365	EST, other/unknown
5,329,044	Total sequences in clusters

Approx # human genes ~30K

FG: Parallel high-throughput transcriptome profiling technologies

- Low throughput (1 RNA species at a time): northern blot, RT-PCR
- 2 scalable principles for detecting / quantifying gene transcription products, and their representative technologies
 - Sequencing short representative sub-sequence (unsupervised): serial analysis of gene expression (**SAGE**). Sequence frequency \propto abundance
 - Nucleotide base pair complementation of short representative sub-sequences (supervised): cDNA / RNA **microarray**. Fluorescent intensity \propto abundance
 - Unsupervised = the universe of measurable entities is not constrained by the assaying platform. However, mapping these entities to known RNA species depends upon reference sequence library.
 - Supervised = universe of RNA species which are measurable is constrained by the assaying platform
- Pre-assay steps: From a biological system at specific state -> extract mRNA -> form cDNA (more stable), fragmented. Amplify? Bias towards 3' end targets, other non-linear artifact.

FG: SAGE (sequencing short representative sub-sequences)

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

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

FG: SAGE

- SAGE example result: 3 transcript **types** relative to a SAGE library

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

FG: SAGE Pros / Cons

■ Pros

- Discover new genes, or old genes with new role (function / tissue-time specificity)
- Abundance of a transcript = simple counting

■ Cons

- 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 / Reference** system for comparison? This is a more general problem that we will see as we progress in functional genomics.

FG: Microarray (nucleotide complementation)

- Definition RNA / cDNA microarray (chip)
 - Single-stranded DNA (gene / EST sub-sequences) anchored at one end on a substrate, eg. gridded array or bead surface. Different species placed on separate grid coordinates / beads. ssDNA fragment (called **probes**), not entire gene sequence is placed. Why?
 - Evolved from southern blots (DNA). Exploits parallelism
 - Mechanistic principle: Nucleotide complementarity $A \leftrightarrow T, G \leftrightarrow C$
 - ssDNA on chip will hybridize to complementary strand in solution (cDNA's derived from a biological system, called **targets**). Complementary strand is fluorescent labeled.
 - Basic assumption: Fluorescence is proportional to RNA abundance (thus gene expression level)

FG: Microarray categories

■ 2 categories of microarrays (by manufacturing process)

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

Spotted

Figure removed due to copyright reasons.

Please see:

Southern, et al. "The Chipping Forecast."

Nature Genetics Supplement 21, no. 1 (January 1999).

Oligo

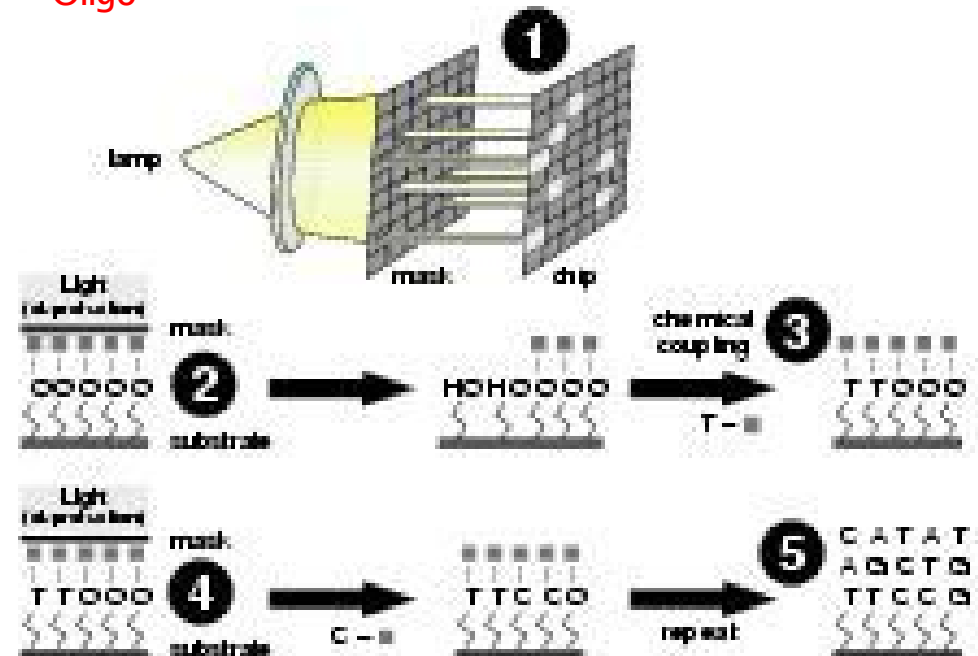


Figure from:

Kohane, Isaac S., Alvin T. Kho, and Atul J. Butte.

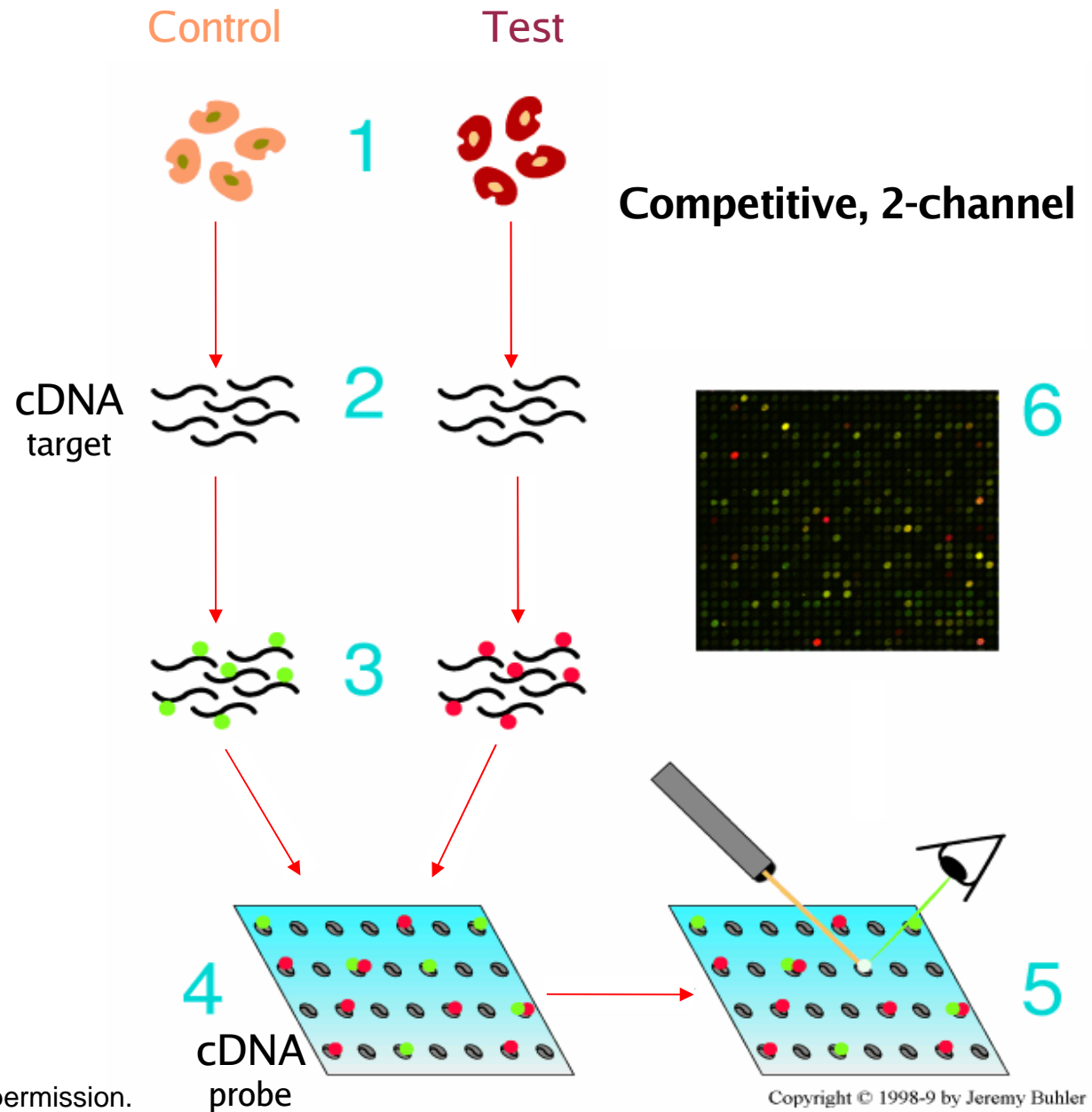
Microarrays for an integrative genomics.

Cambridge, Mass: MIT Press, 2003. ISBN: 026211271X.

FG: Microarray categories – 2 channel

- 2 categories of microarrays (by usage principle): 2 channel vs. 1 channel

- 2 channel
- Paired experiment design
- Can be treated as 1 channel if common control is used for all chip experiments
- Internal (measurement, possibly biological) control / reference for fluorescence

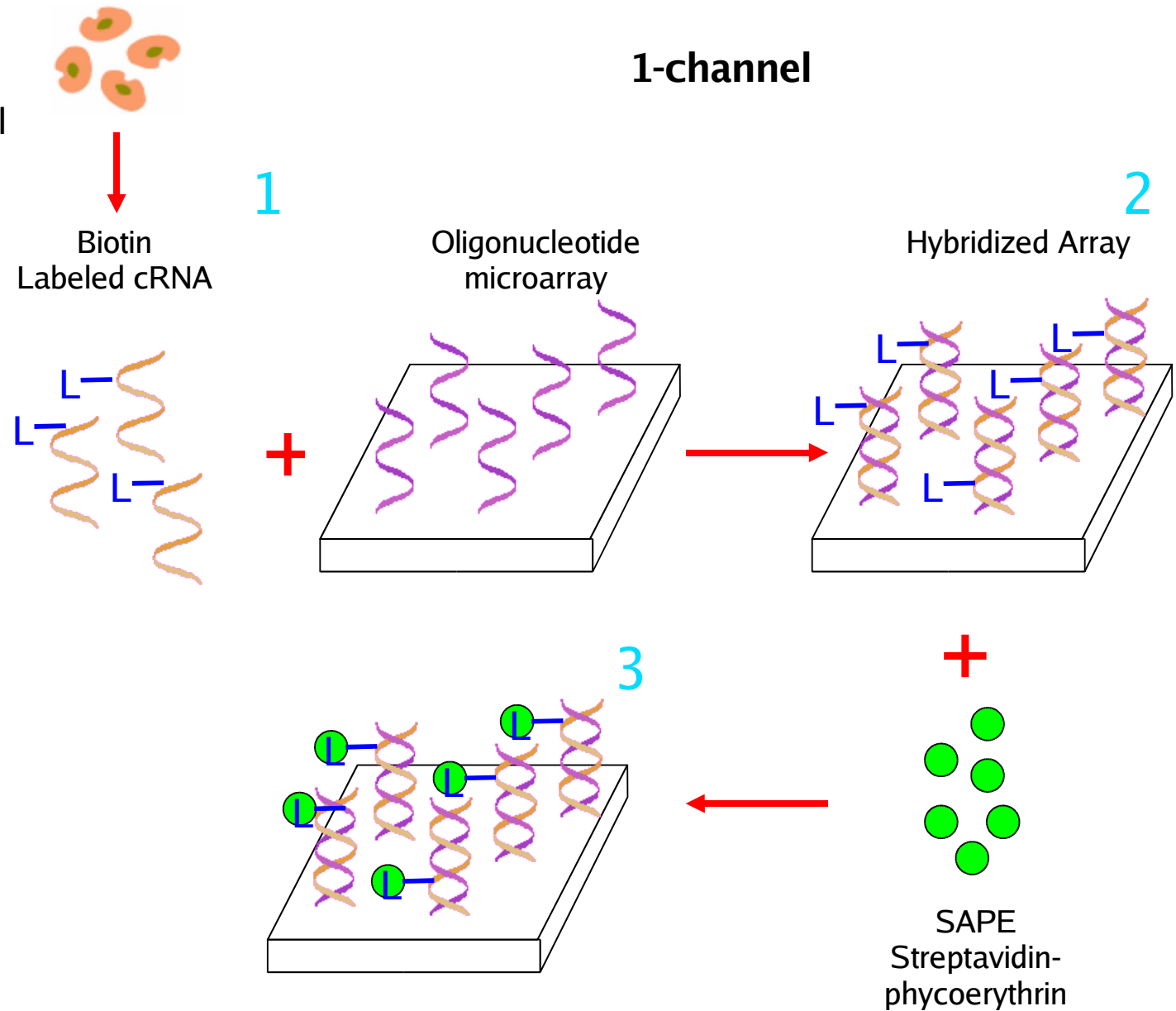


FG: Microarray categories – 1 channel

- 2 categories of microarrays (by usage principle): 2 channel vs. 1 channel

- 1 channel

- Different internal control housekeeping genes



FG: Generic microarray experiment stages

- Generic stages of a microarray experiment
 - Experimental design involving biological system under investigation. Replicates – both measurement / technical and biological
 - RNA and target preparation: Extract mRNA. Convert (to ss cDNA typically). Label with fluorescence.
 - Probe hybridization.
 - Fluorescence image analysis
 - **Microarray data analysis** (post image) – one lecture onto itself

Figures removed due to copyright reasons. Please see:

Pevsner, Jonathan. *Bioinformatics and Functional Genomics*. Hoboken, NJ: Wiley-Liss, Inc., 2003. p. 181. ISBN: 0471210048

FG: Microarray, transcriptome profiling caveats

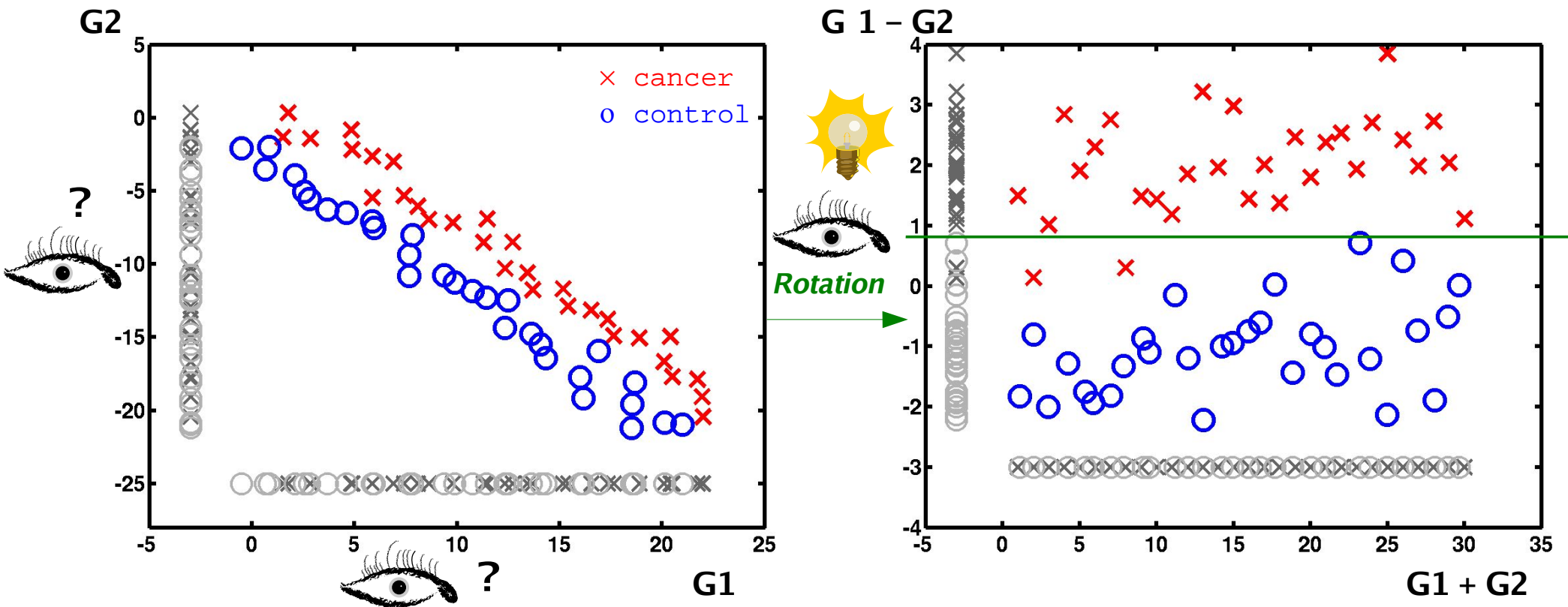
- **Microarray oligo probe design technical issues**
 - 3'-end target amplification bias (not strictly microarray problem). Assess by 3'-end probe-to-target / 5'-end probe-to-target intensity ratios of housekeeping probes eg. Gapdh, β -actin. Non-linear effect (with respect to diff RNA species for fixed time interval).
 - CG% content of probes: C \leftrightarrow G (3 H-bonds) vs. A \leftrightarrow T (2 H-bonds) -> diff bonding energy -> diff hybridization rate. Non-linear (wrt. diff RNA species for fixed time interval)
- **Cons:** Probe specificity. Cross RNA species hybridization, promiscuous probes
- **General caveats with transcriptome profiling studies**
 - Non-uniform RNA degradation – pre-assay step
 - “Noise”: Measurement / technical and biological variation. Choice of a Reference system. This is a more general problem.
 - Assumption: Central Dogma holds ($\text{mRNA} \propto \text{Protein}$). Bio-process of interest engages transcriptome machinery and state is characterized by transcriptome profile
 - Averaging / pooling of RNA across heterogeneous cell populations

FG: Generalizations of microarray

- 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.

FG: Parallel high-throughput tech changes our perspectives / questions?

- 2 views qualitatively-diff afforded by these technologies
 - View 1: Whole = Sum of individual parts. Purely an efficient way to screen many many molecules. Multi-plexing classical assaying techniques eg. northern blot
 - View 2: Whole > Sum of individual parts. As above, plus unraveling intrinsic regularities (correlations) between measured molecules. Eg. below, do G1 or G2 intensities alone distinguish between disease groups?



FG: next time in lecture 2

- ... mathematical reformulation of biological problems involving multi-variables (microarrays). Next class.