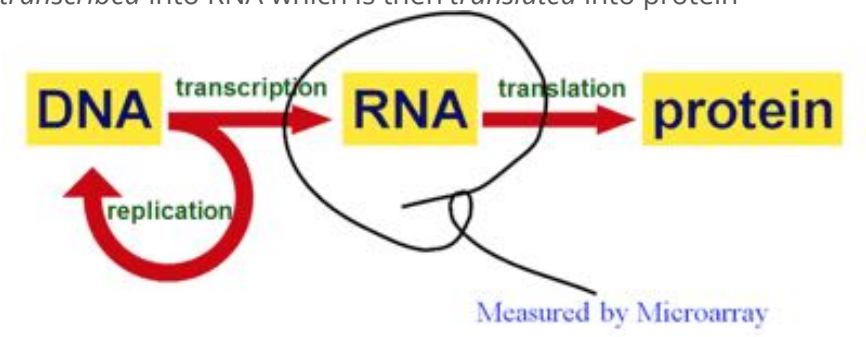
Microarray technology

Mikhail Dozmorov
Fall 2017

The Central Dogma of Molecular Biology

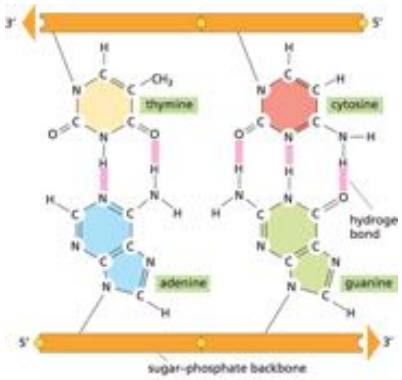
DNA is *transcribed* into RNA which is then *translated* into protein



Expression of all genes define phenotypes

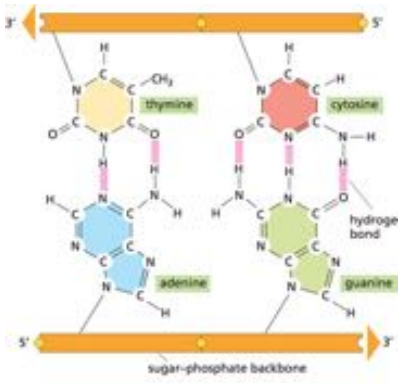
- Mendelian genetics explains transmission of genetic information, but sheds no light on how genes create cellular and organismic phenotypes.
- Assays have been developed to more formally study **the association between genes and phenotypes**.

Rules of base pairing



- All genetic code is spelled out with just four chemical letters, or bases: adenine (A), thymine (T), cytosine (C) and guanine (G)
- These pair up, A with T and C with G

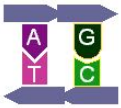
Complementary hybridization



- Two single-stranded DNA molecules whose sequences are complementary to each other will exhibit a tendency to bind together to form a single double-stranded DNA molecule. This process is called **hybridization**.

Complementary hybridization

- Sequence fully complementary to a target will hybridize with much higher efficiency than partially complementary.



- Even when the sequences on the two strands do not match perfectly, as long as there is sufficient **overall** similarity, it is likely that some base pairing will occur.

Complementary hybridization

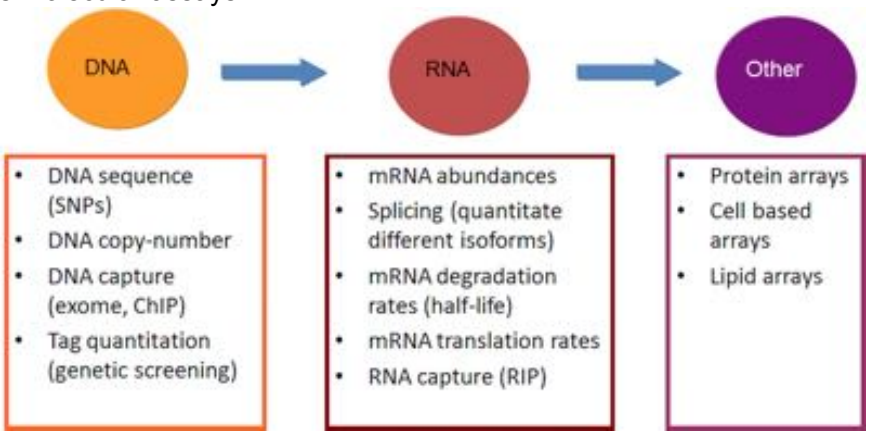
- The tendency of DNA strands of complementary sequences to hybridize is exploited in hybridization assays.
- A **probe** consisting of a homogeneous sample of single-stranded DNA molecules, whose sequence is known, is prepared and **labeled with a reporter** fluorescent chemical
- An **immobilized target**, usually a single-stranded DNA molecule, is challenged by the probe.
- As the probes will hybridize preferentially to sequences complementary to the targets, they can be identified by the presence of fluorescence.
- Location of the targets, and the amount of fluorescence, defines which genes, and how much, are expressed.

What is a Microarray?

- "A DNA microarray is a multiplex technology consisting of thousands of oligonucleotide spots, each containing picomoles of a specific DNA sequence."
- An *oligonucleotide* (from Greek prefix *oligo*-, "having few, having little") is a short nucleic acid polymer.

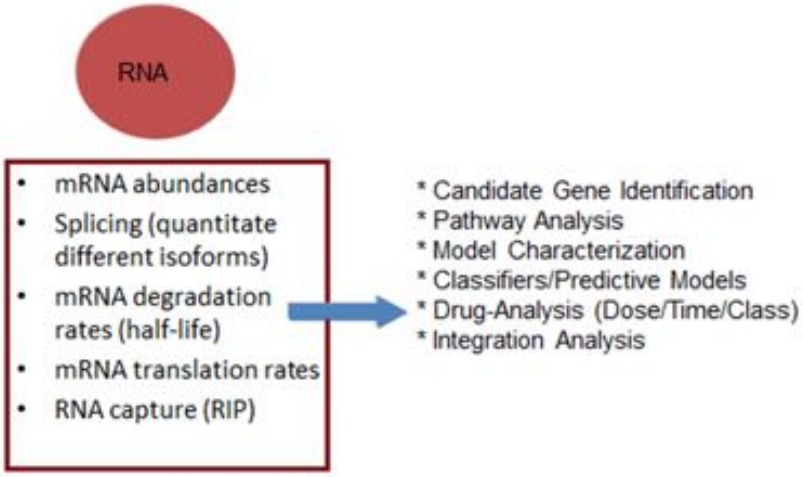
What Are Microarrays Used For?

- Various molecular assays



What Are Microarrays Used For?

- Biological insights



Microarrays measure expression of all genes

- Traditional molecular biology research followed a "*one gene per experiment*" paradigm
- With the advent of microarrays, research practice has moved from a "one gene at a time" mode to "thousands of genes per experiment"
- Allows for the study of how genes function *en masse*

Basic Design of Expression Arrays

- For **each gene** that is a target for the array, we have a **known DNA sequence**
- Microarrays are composed of short DNA sequences complementary to the target genes
- These sequences are attached to a slide at high density

Basic Design of Expression Arrays

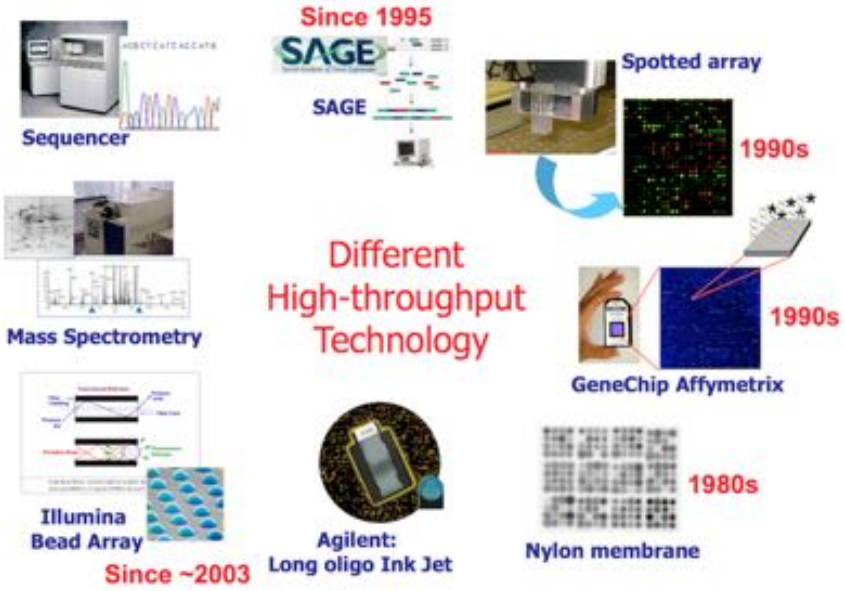
- mRNA is reverse transcribed to cRNA, and if a complementary sequence is on the on a chip, the cRNA will be more likely to hybridize to it
- The cRNA is labeled with a dye that will fluoresce and generate a signal that is monotonic with the amount of the mRNA sample
- The amount of hybridization can be **quantitatively** measured by the amount of fluorescence

Microarray types

- Two major types of microarrays

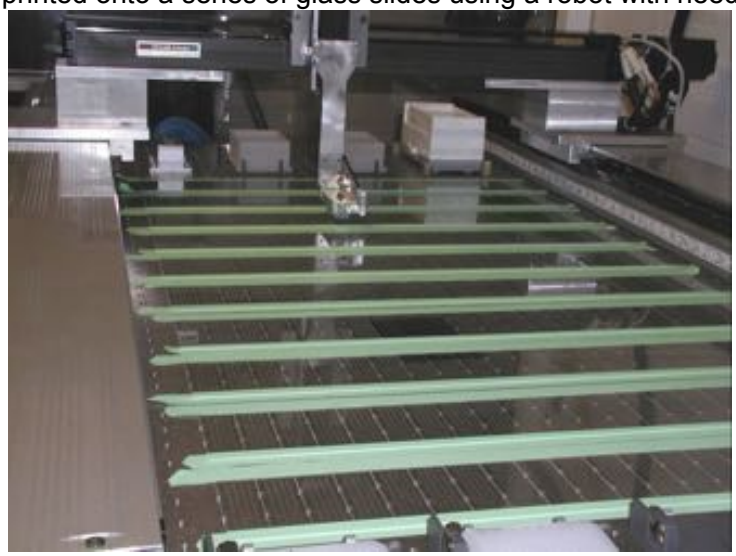
1. **Spotted arrays**, typically *two-channel*
2. **Oligonucleotide arrays**, typically *single-channel*

Different High-throughput Technology



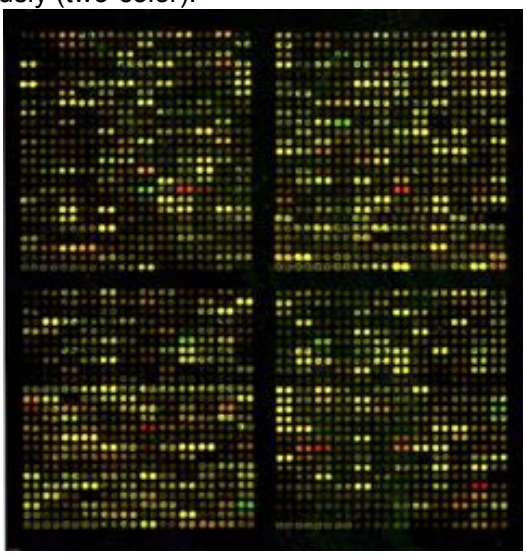
Spotted Arrays

- Robotically printed onto a series of glass slides using a robot with needle-heads.



Spotted Arrays

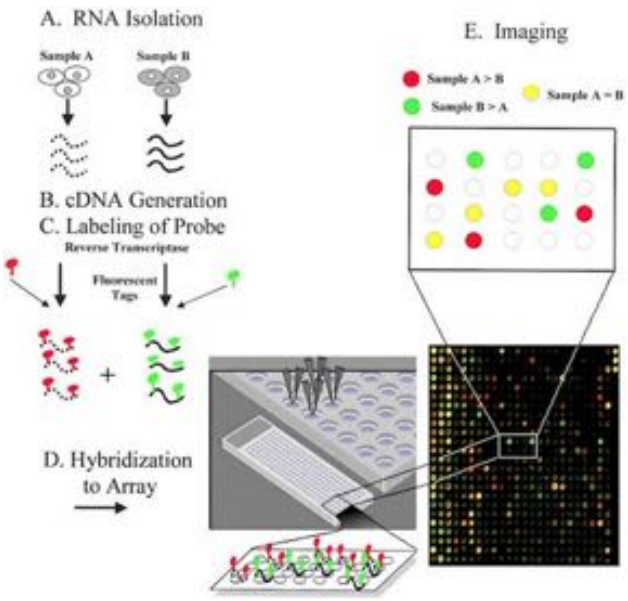
- Printing produce a characteristic gridding pattern and almost always use two samples simultaneously (two-color).



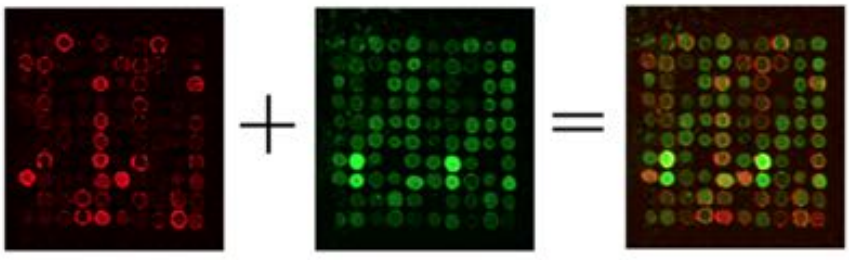
How two-channel arrays work?

- Use two samples, control (reference) and test, e.g., tumor and normal cells. Take identical amounts of mRNA and convert to cDNA
- Incorporate **GREEN** fluorescent dye into one cDNA (e.g. control)
- Incorporate **RED** fluorescent dye into the other (e.g. test)
- Hybridize mixture of both onto array. **GREEN** spot indicates mRNA expression only in control sample, **RED** - in test sample, **ORANGE** - in both

Two-channel (two-color) arrays



Combine scans for Red & Green



- False color image is made from digitized fluorescence data, not by superimposing scanned images.

http://www.bio.davidson.edu/courses/genomics/chip/chip.html

Two-channel arrays

- Advantages

1. Assessment of gene expression in two samples on a single array
2. Two samples have the same background variability on the array
3. Typically, longer molecules are used, so non-specific binding is not much of a problem

Two-channel arrays

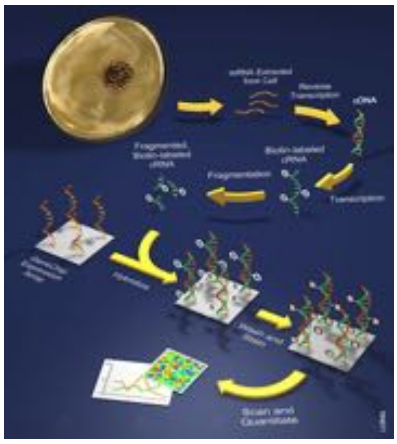
- Disadvantages

1. More laborious, need to handle two samples
2. Each channel may behave differently
3. Typically, one spot per gene - optical noise is a concern
4. Normalization of microarray data WITHIN and BETWEEN the arrays is still needed

Single-channel arrays

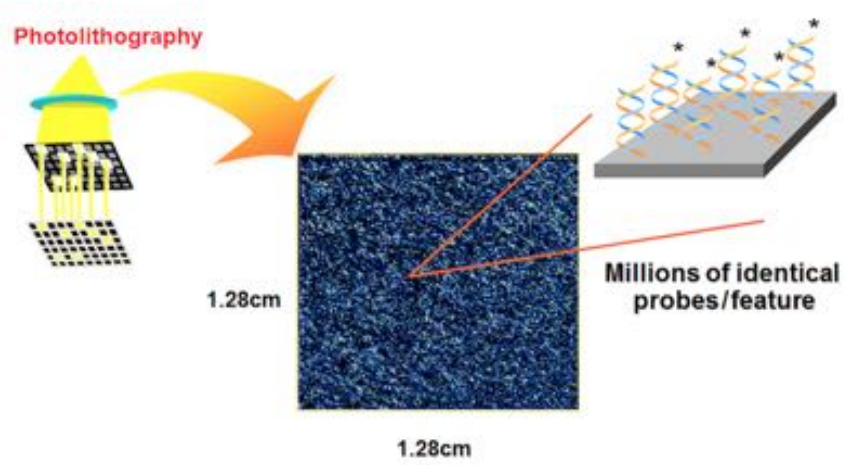
Single-channel arrays

- mRNA extraction from one sample
- cRNA synthesis and fluorescent dye-labeling
- cRNA hybridization onto array
- Scanning and quantification of fluorescence of each spot



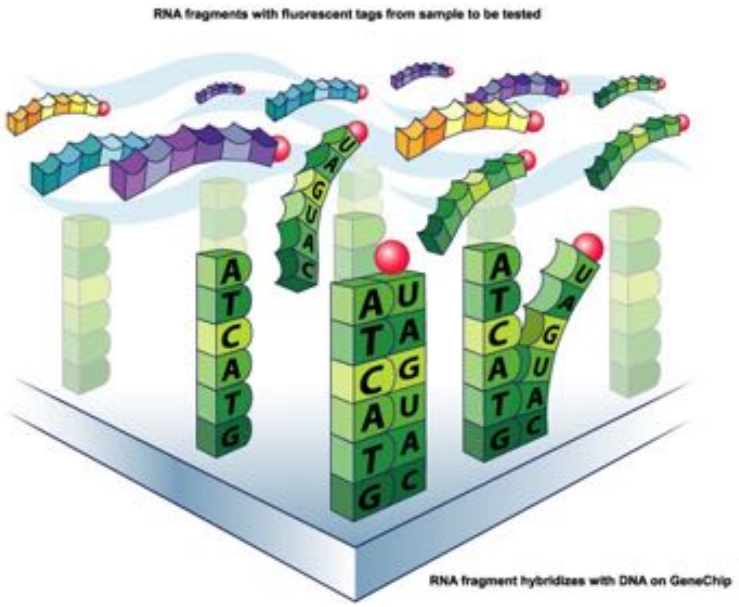
Oligonucleotide Arrays

Affymetrix arrays

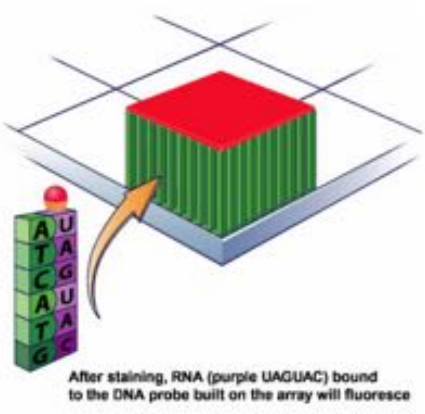


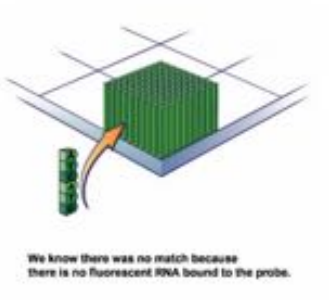
https://youtu.be/MRmpeBTwwWw

RNA Wash



RNA Wash



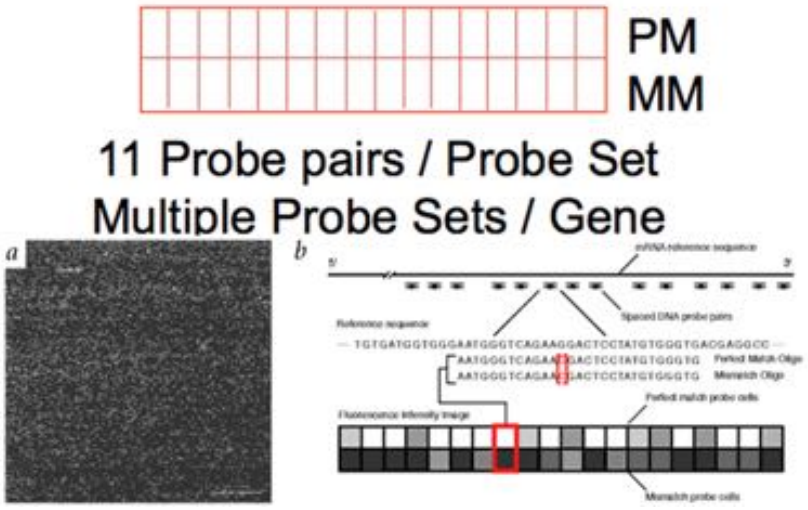


Affymetrix array design

- Rather then an entire gene being placed in a Affymetrix Genechip is an oligonucleotide array consisting of a several **perfect match** (PM) and their corresponding **mismatch** (MM) probes that interrogate for a single gene.

1. PM probe is the exact complementary sequence of the target genetic sequence, composed of 25 base pairs
2. MM probe, which has the same sequence with exception that the middle base (13th) position has been reversed
3. There are roughly 11-20 PM/MM probe pairs that interrogate for each gene, called a probe set

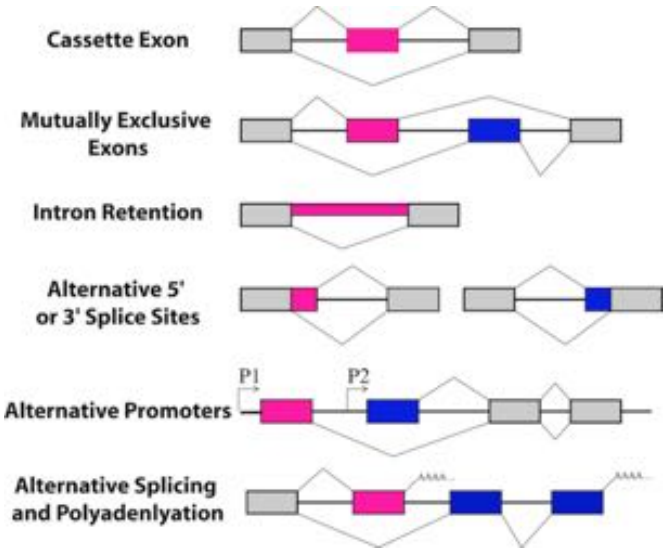
Affymetrix array design



Affymetrix files

- **DAT** file: Image file, 10^7 pixels, ~50 MB.
- **CEL** file: Cell intensity file, probe level PM and MM values.
- **CDF** file: Chip Description File. Describes which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs).

Alternative splicing

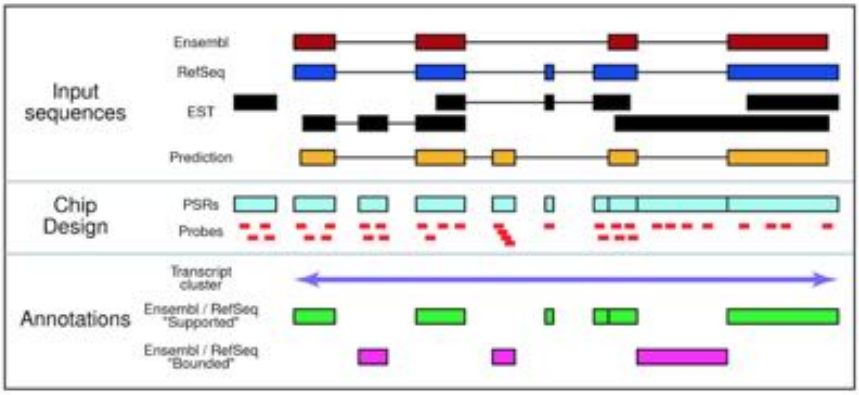


Exon array principles

- Gene-level and exon-level detection of expression.
- Allow detection of alternative splicing mRNA transcripts.

Exon array design

- PSR - Probe Selection Region



Affymetrix exon arrays

- Affymetrix GeneChip **Exon** 1.0 ST

1. Wide coverage
2. Well annotated genes plus gene prediction sets
3. Over 1.4 million probe sets

The use of exon array

- Advantages

1. Allow detection of alternative splicing.
2. Cost is about the same as for regular microarrays.

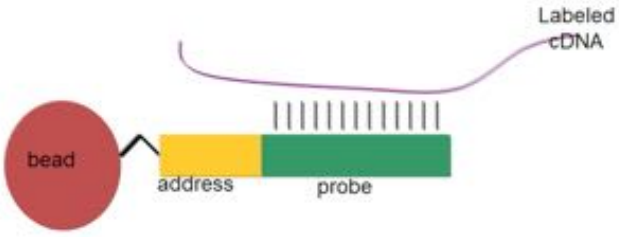
The use of exon array

- Disadvantages

1. Careful probe design is imperative.
2. Methods for analysis are not well developed.

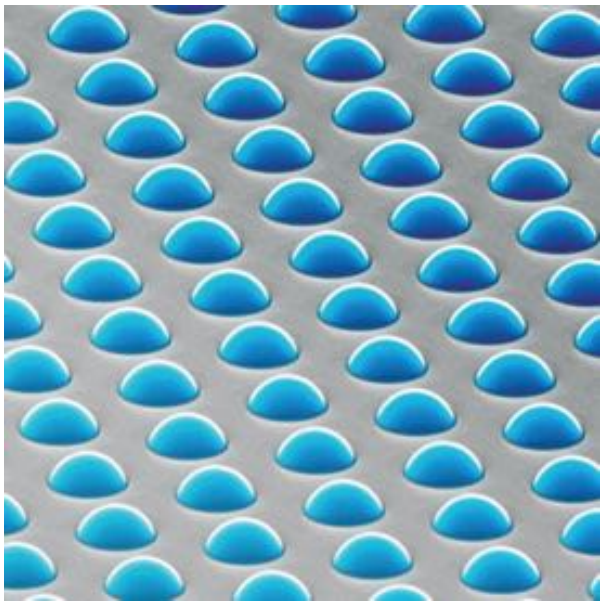
Self-Assembling Bead-Arrays

- Made by Illumina
- 3 μm silicon beads, randomly spread across the surface of the chip
- Each bead coated with ~105 identical 50bp probes
- Each probe has identifying barcode (address) sequences
- ~30 beads per gene



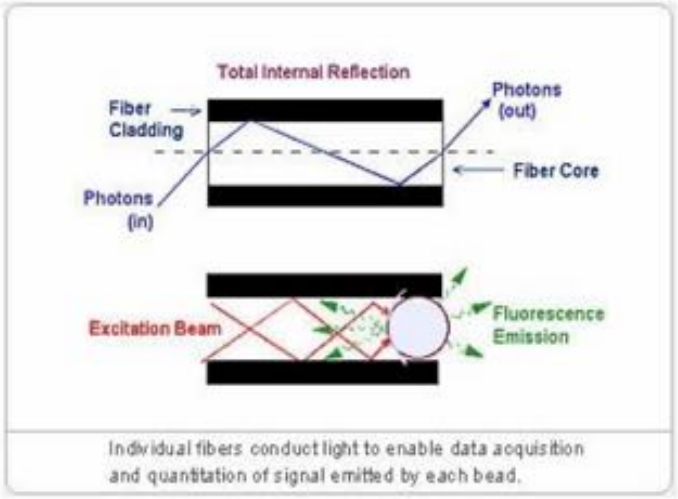
Illumina Bead Arrays

Beads form array on light fibers.



Illumina Bead Arrays

Illumination from below excites fluorescence – quantifies probe bound.



Illumina Bead Arrays

- Each chip of the Ref-8 contains 8 arrays with ~ 25,000 targets, plus controls
- Each chip of the WG-6 contains 6 arrays with ~ 50,000 targets, plus controls
- Each chip of the HT-12 chip contains 12 arrays with ~ 50,000 targets and controls

The use of single-channel arrays

- Advantages

1. Analysis of ONE sample per array
2. Straightforward approach - more fluorescence = more RNA

The use of single-channel arrays

- Disadvantages

1. Need to use another array(s) for comparative analysis
2. Careful normalization of one microarray data to the other is a must

InkJet Arrays

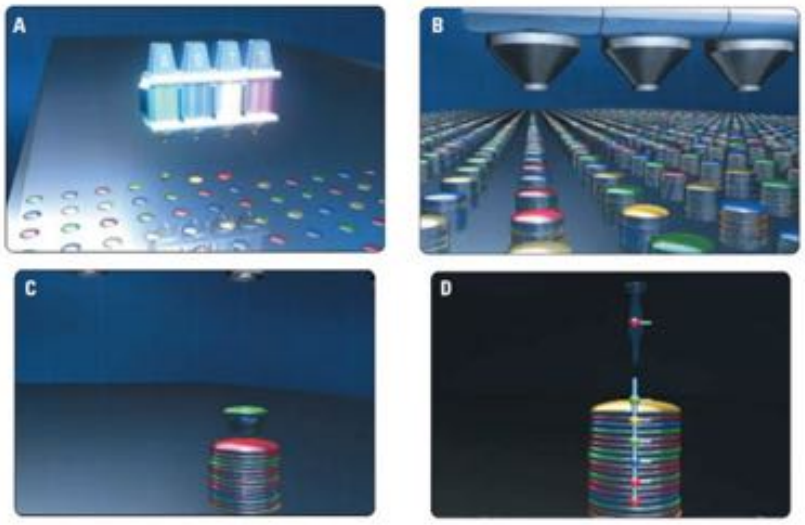
In 1999, HP spun off its life-science and measurement division into Agilent Technologies.

Agilent Technologies

hp invent

The new company wanted to determine if **printer** technology could be harnessed to generate microarrays.

Inkjet Array Manufacture Involves Sequential Nucleotide Addition



Genome-wide genotyping panels

- 10,000-5 million variants
- Affymetrix, Illumina
- Random SNPs
- Selected haplotype tag variants
- Copy number probes
- More lower frequency variants
- Exome variants
- Some arrays allow variants to be added

SNP arrays

- Affymetrix SNP Array 5.0 (500,568 SNPs) and SNP Array 6.0 (934,968 SNPs)
- Illumina HumanHap300 (317,511 SNPs), HumanHap550 (555,352 SNPs), HumanHap650Y (660,917 SNPs), and Human1M (1,072,820 SNPs)

Table 2 Number of genes with 0% coverage by SNP chips

SNP chip	*CEU*	*CHB*	*JPT*	*YRI*
SNP Array 5.0	575	540	496	980
SNP Array 6.0	163	152	151	265
HumanHap300	106	209	236	1064
HumanHap550	46	50	56	225
HumanHap650Y	43	46	52	114
Human1M	8	8	9	16

Note: only gene regions containing with 5 HapMap common SNPs were considered, and coverage was evaluated at $r^2$0.8.

https://www.nature.com/ejhg/journal/v16/n5/full/5202007a.html

Tiling arrays: Biological motivations

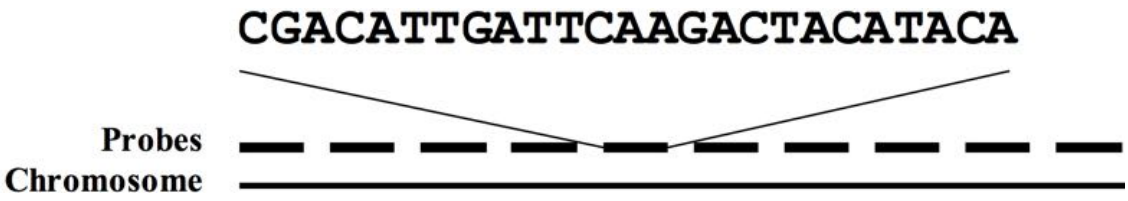
- There are many types of "events" happen at different locations on the genome. For example, protein bindings, epigenetic modifications (DNA methylation and histone modifications), copy number variations, etc.
- It is often of great interest to detect the genomic locations where a specific event happens, or quantify the events along the genome.

An example: transcription factor (TF) binding

- Transcription factors(TF): proteins that bind to specific DNA sequences and control the transcription from DNA to mRNA.
- There are many different types of TFs, each recognize different DNA sequences (motifs).
- The functions of the TFs are important for understanding gene regulatory mechanisms.
- The first step toward the understanding is to detect the TF binding sites (TFBS).

Tiling arrays

- The goal is to quantify the events of interests along the genomes, and/or detect the genomic coordinates for the events.
- Work the same as gene expression array (hybridization based), except that the probes are designed to tile up the genome at non-repeat regions.
- Data for probes in the location of interest often behave differently from backgrounds (e.g., bigger intensities).

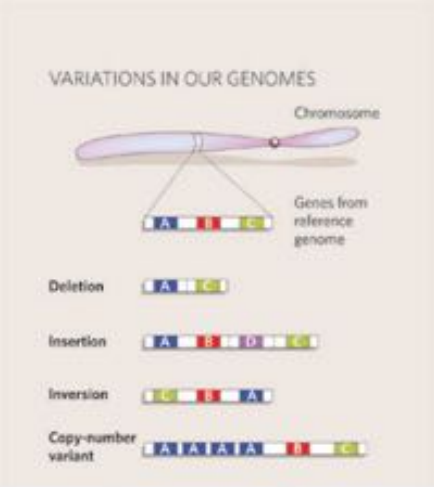


Types of tiling arrays

- ChIP-chip: Chromatin ImmunoPrecipitation (ChIP) + tiling array (chip) for detecting transcription factor binding sites or measuring histone modification levels.
- MeDIP-chip: Methyl-DNA ImmunoPrecipitation (MeDIP) + tiling array (chip) for measuring DNA methylation level.
- ArrayCGH (Comparative Genomic Hybridization) for detecting copy number variations.

There's no major differences in array designs. Difference are the ways to prepare biological samples.

Copy number alterations (CNA) can lead to disease



Nature **437**, 1084-1086

- CNAs are a hallmark of tumor genomes
- CNAs can lead to adverse expression

Array comparative genomic hybridization

- aCGH (Array CGH). A technique based on competitively hybridizing fluorescently labelled test and reference samples to a known target DNA sequence immobilized on a solid glass substrate and then interrogating the hybridization ratio.
- The signal ratio between a test and reference sample is normalized and converted to a log ratio, which acts as a proxy for copy number
- An increased log ratio 2 represents a gain in copy number in the test compared with the reference; conversely, a decrease indicates a loss in copy number

Measuring CNAs with array comparative genomic hybridization (aCGH)

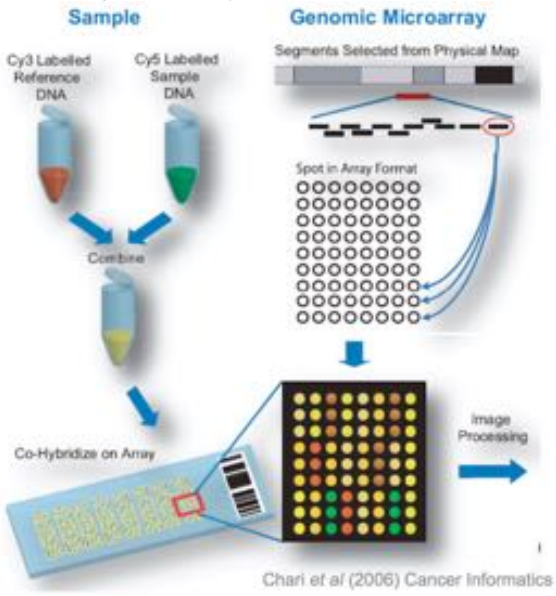
- Array hybridization - similar to two-color array studies:

1. Test DNA sample - Unknown DNA copy number
2. Reference DNA sample - normal karyotype DNA copy number
3. Label, mix, hybridize, scan

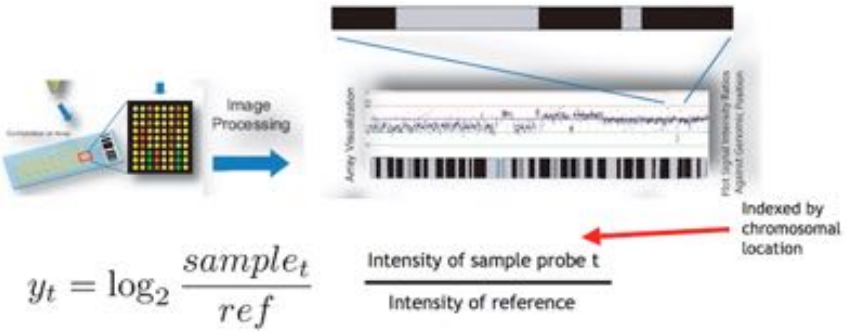
- Array analysis - resulting data are normalized, log test over reference intensities for genomic targets

Measuring CNAs with aCGH

- aCGH - array comparative genomic hybridization



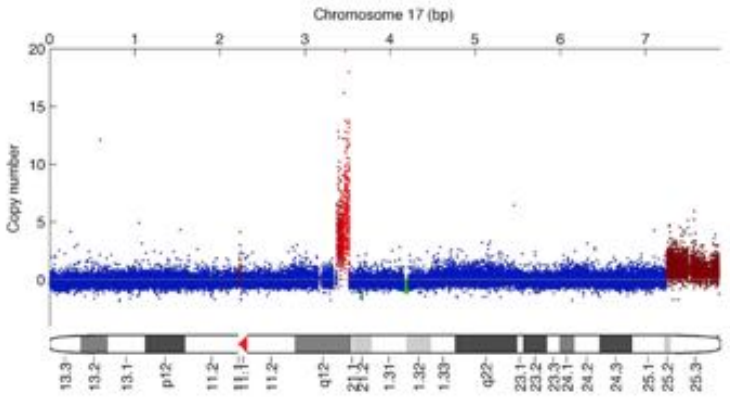
Detecting copy number alterations



$$y_t = \log_2 \frac{sample_t}{ref}$$

Intensity of sample probe t / Intensity of reference

Example copy number alterations in cancer

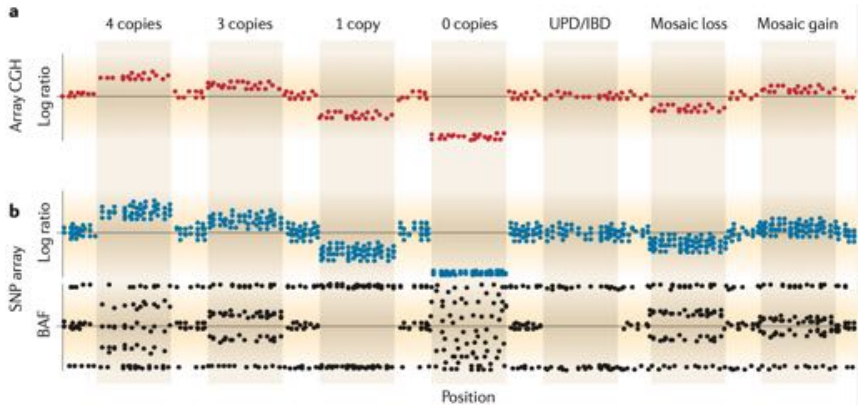


SNP microarrays

Hybridization-based assays in which the target DNA sequences are discriminated on the basis of a single base difference. Assays are processed with a single sample per array and perform both SNP genotyping and copy-number interrogation.

- B allele frequency (BAF) metric - the proportion of the total allele signal (A + B) explained by a single allele (A).
- The BAF has a significantly higher per-probe SNR than the log ratio data and can be interpreted as follows:
- a BAF of 0 represents the genotype (A/A or A/–), whereas 0.5 represents (A/B) and 1 represents (B/B or B/–).
- Different BAF values occur for AAB and ABB genotypes or more complex genotypes (for example, AAAB, AABB and BBBA). Homozygous deletions result in a failure of the BAF to cluster. Thus, the BAF may be used to accurately assign copy numbers from 0 to 4 in diploid regions of the genome.
- The BAF also allows detection of copy-neutral events such as segmental uniparental disomy (segmental UPD) or whole-chromosome UPD and identity by descent (IBD), which results when a segment of one chromosome is replaced by the other allele without a change in copy number (this is therefore not detectable by array CGH).

aCGH vs SNP arrays



Limitations of both - repetitive regions

Tiling array data analysis

- Goal: detect locations of interests (also called “peaks”) based on probe locations and signals.
- Normalization: remove technical artifacts.
- Detection of regions of interests:
  – Data from neighboring probes need to be combined to make inference, because the regions of interests often overlap many probes.
  – Easiest method: moving average, then use an arbitrary cutoff.
  – Many different methods.

Popular software

For ChIP-chip:
– CisGenome, http://www.biostat.jhsph.edu/~hji/cisgenome/ – MAT, http://liulab.dfci.harvard.edu/MAT/

CNVarrays:
– Affymetrix:
- APT: uses a hidden Markov model, https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html
- R package VanillaICE: HMM based. https://www.bioconductor.org/packages/release/bioc/html/VanillaICE.html - R package DNAcopy: Circular Binary Segmentation. https://bioconductor.org/packages/release/bioc/html/DNAcopy.html
– Illumina:
- QuantiSNP: https://sites.google.com/site/quantisnp/
- PennCNV: http://penncnv.openbioinformatics.org/en/latest/

Tiling arrays Summary

- Tiling arrays are DNA microarrays for detecting locational modifications of genome.
- Probes tile up a part of whole genome.
- Still hybridization based (DNA segments stick to probes), same as gene expression arrays.
- Location of interests shows some patterns: peaks for TFBS, or plateau for CNV.
- Need to combine data from neighboring probes to make calls.
- Being replaced by sequencing (e.g., ChIP-seq).

Additional Microarray Platforms

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
Tissue	tissues	proteins	Histology; protein expression (immunohistochemistry)

Applications of microarrays

All areas of life sciences

- **Cancer research**: Molecular characterization of tumors on a genomic scale; more reliable diagnosis and effective treatment of cancer
- **Immunology**: Study of host genomic responses to bacterial infections
- **Model organisms**: Multifactorial experiments monitoring expression response to different treatments and doses, over time or in different cell types

Typical comparisons

- Compare mRNA transcript levels

1. different type of cells, tissues (e.g., liver vs. brain)
2. treatment (Drugs A, B, and C)
3. disease state (tumor vs. normal)
4. different organism (yeast, different strains) different timepoints

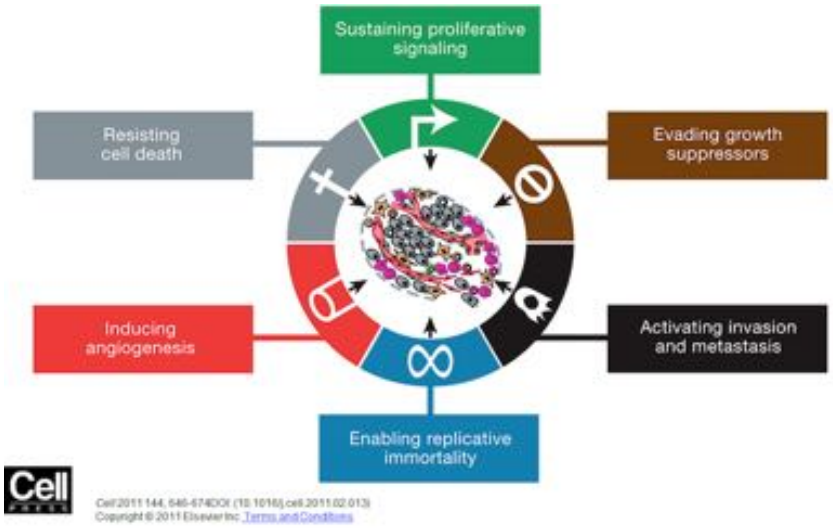
Normal vs. cancerous cells

- All cells in the body are the lineal descendants of a fertilized egg. Almost all of these cells carry genomes that are reasonable accurate copies of the genome that was initially present in the fertilized egg
- However, cells throughout the body are phenotypically distinct (e.g., skin cells versus brain cells) though genetically identical.
- **Differentiation** is the process whereby cells in different parts of the embryo begin to assume distinct phenotypes.
- The molecular mechanisms of differentiation can be understood by examining the sets of genes that are expressed (transcribed) in some cells but not others. These are tissue-specific genes.

Cancer

- Cancer is a disease in which cells escape the restraints on normal cell growth, and become less and less differentiated
- Once a cell has become cancerous, all of its descendant cells are cancerous
- Clonal expansion of cancer cells results in cancer progression

Hallmarks of cancer



Genetic abnormalities in cancer

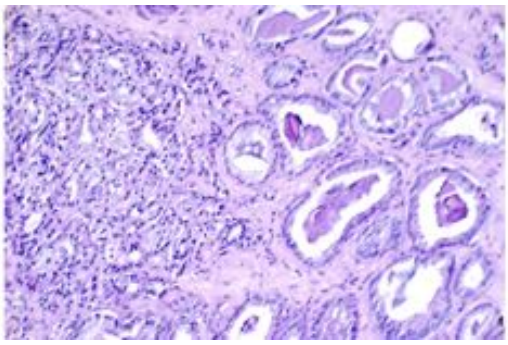
- Mechanisms whereby mutations and genetic alterations cause cancer

1. Gain of function (proto-oncogene)
2. Loss of function (tumor suppressor gene)
3. Translocations - creation of chimeric proteins with novel function
4. Aberrant gene expression
5. Epigenetic changes

Clinical cancer detection

- Pathologist makes an interpretation based upon a compendium of knowledge which may include

1. Morphological appearance of the tumor
2. Histochemistry
3. Immunophenotyping
4. Cytogenetic analysis
5. etc.



Microarrays in cancer detection

- Applications of microarrays

1. Characterize molecular variations among tumors by monitoring gene expression

2. Divide morphologically similar tumors into different groups based on gene expression.

- Goal: microarrays will lead to more reliable tumor classification and sub-classification (therefore, more appropriate treatments will be administered resulting in improved outcomes)