Arrays & Expression

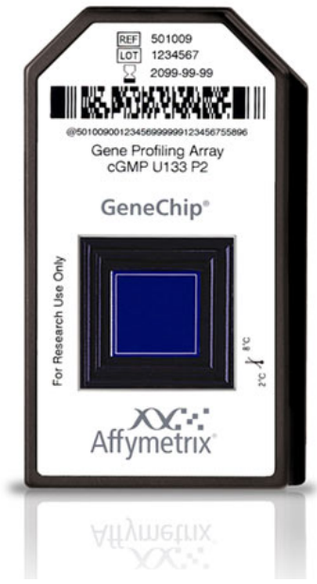
MMG 835, SPRING 2016
Eukaryotic Molecular Genetics

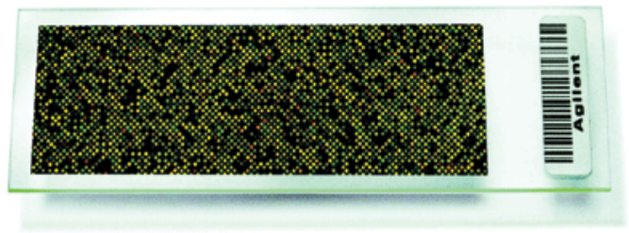
George I. Mias

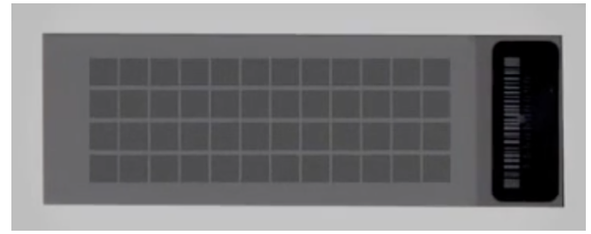
Department of Biochemistry and Molecular Biology
gmias@msu.edu

Arrays

- Spotted DNA arrays
  - Fluorescent (Cy3, Cy5)
- Long/short oligonucleotide arrays
- Glass slides
- Bead Arrays



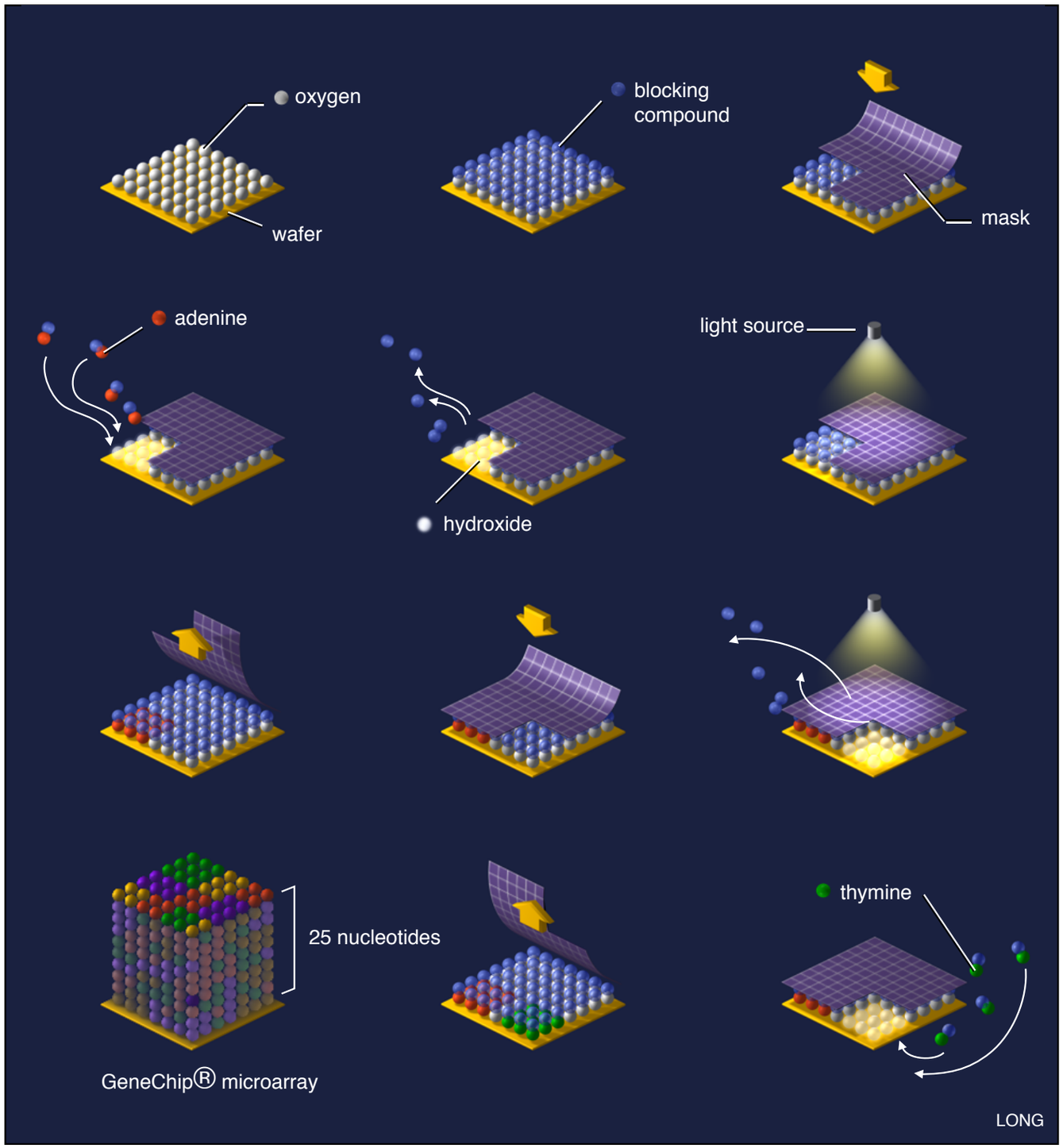




affymetrix.com
agilent.com
illumina.com
thermofisher.com

Arrays

- Affymetrix Genechip production
- 5-inch square quartz wafer
- 18-20 micron square windows
- nucleotide masks
- repeat for 25 nucleotides



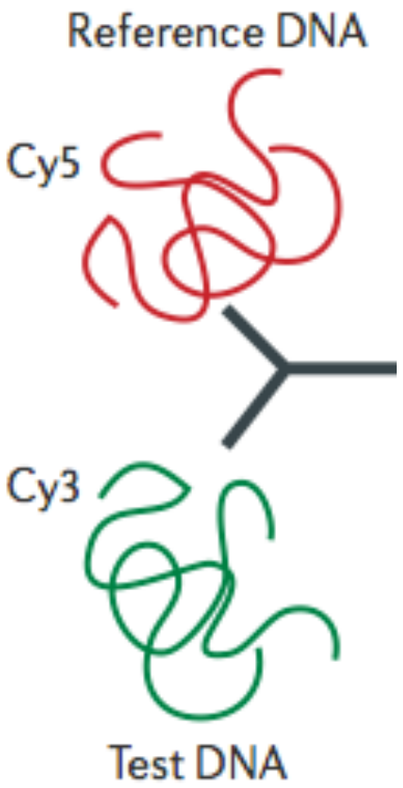
The Science Creative Quarterly
http://www.scq.ubc.ca/spot-your-genes-an-overview-of-the-microarray/ (Jiang Long)

Array CGH

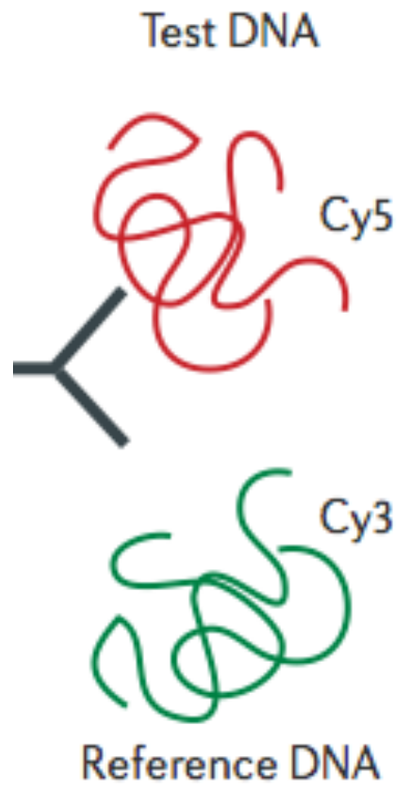
Array-based Comparative Genome Hybridization, **CGH**.

- genomic DNA, used as probe
- DNA copy-number variations
- Measuring relative copy number in the genome rather than in cellular mRNA.
- CGH resolution ~ 10-20 Mb

Array CGH

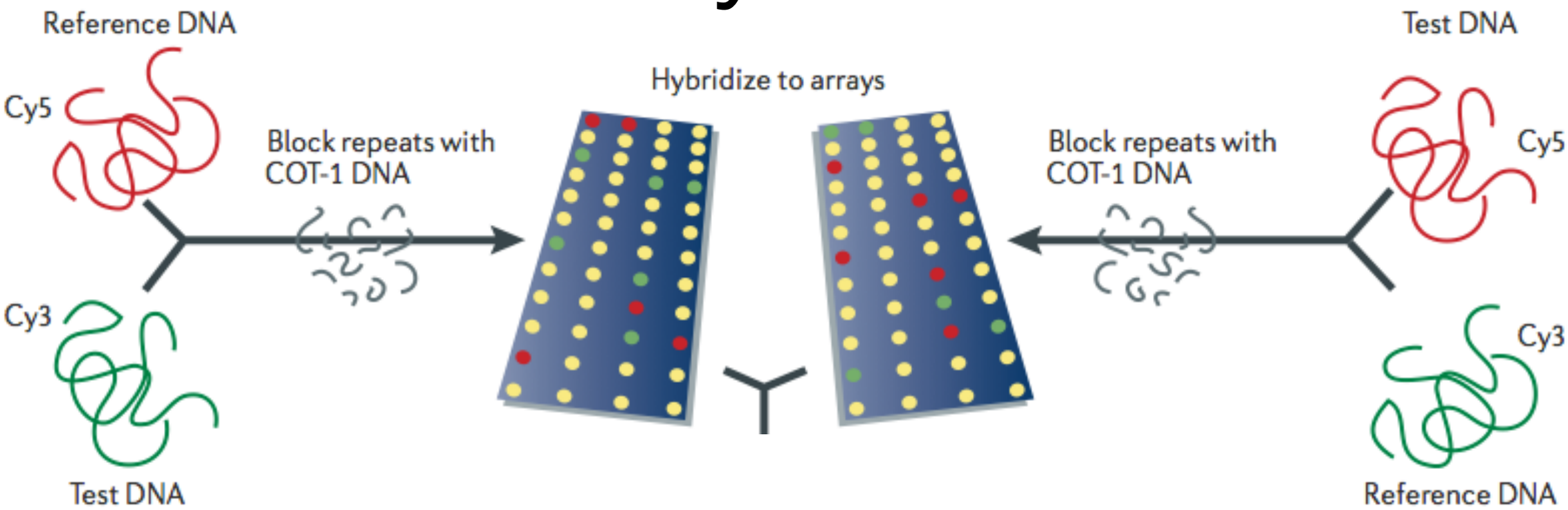


Reference and test DNA differentially Labelled with fluorescent tags (Cy5 and Cy3, respectively)



Feuk, Nature Reviews Genetics 7, 85-97 (2006)

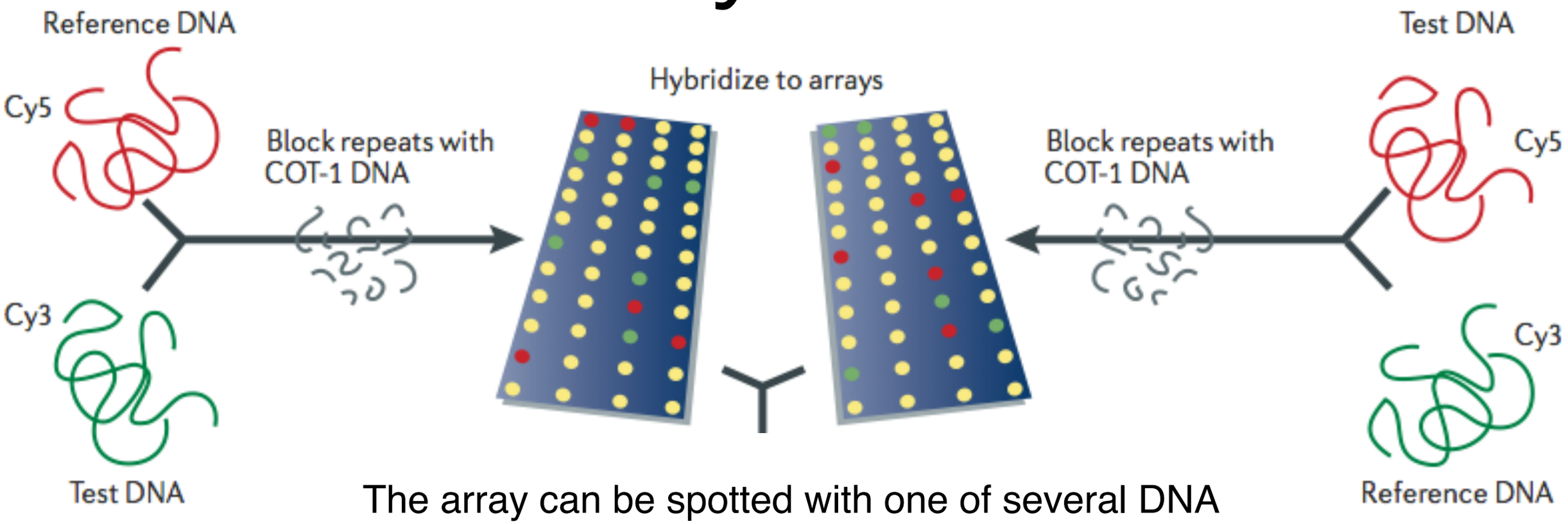
Array CGH



- Repetitive-element binding is blocked using COT-1 DNA.
- Samples are then hybridized to genomic arrays

Feuk, Nature Reviews Genetics 7, 85-97 (2006)

Array CGH

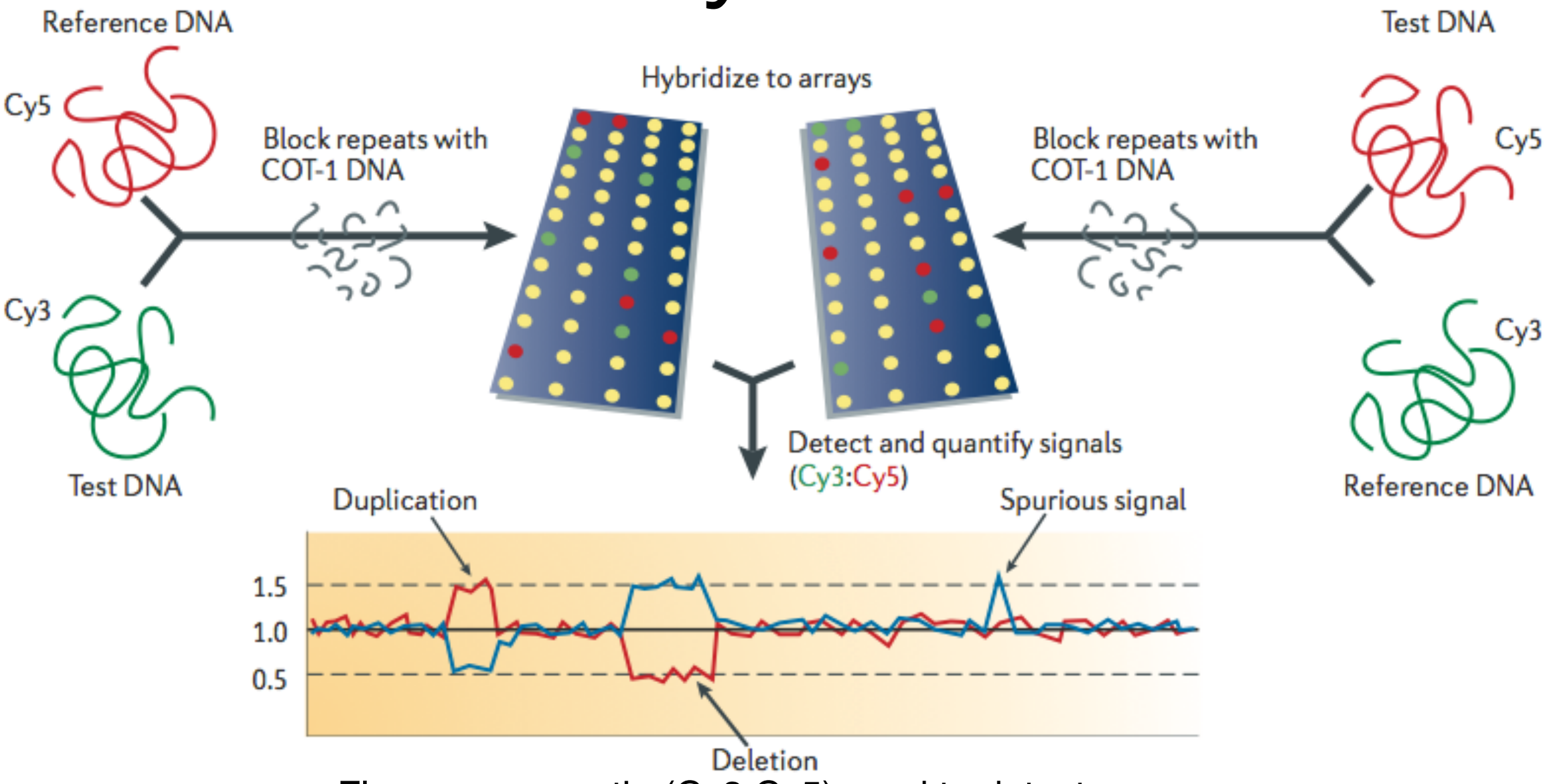


The array can be spotted with one of several DNA sources:

- including BAC clones
- PCR fragments
- oligonucleotides

Feuk, Nature Reviews Genetics 7, 85-97 (2006)

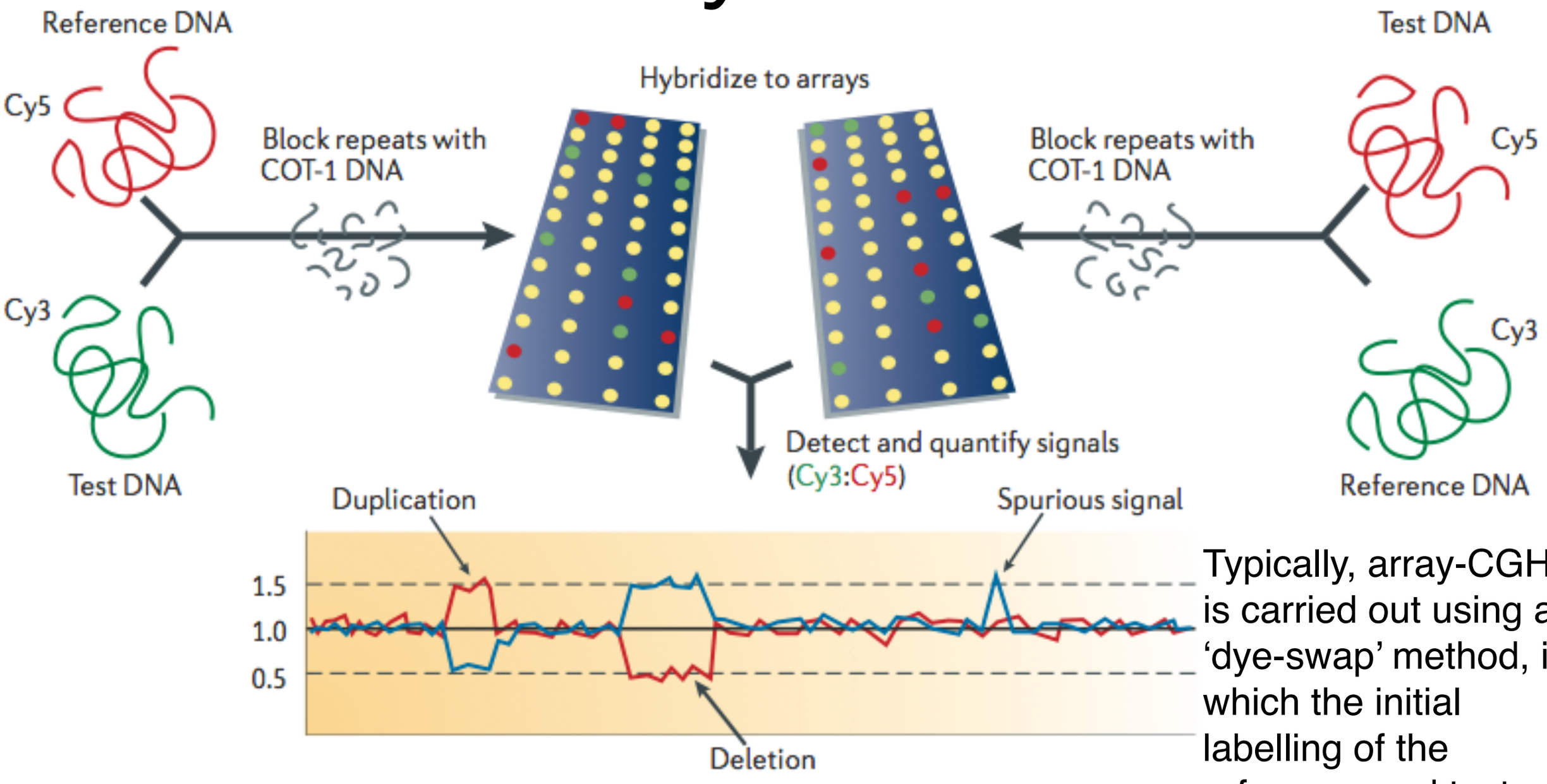
Array CGH



Fluorescence ratio (Cy3:Cy5) used to detect copy-number differences between the 2 DNA samples.

Feuk, Nature Reviews Genetics 7, 85-97 (2006)

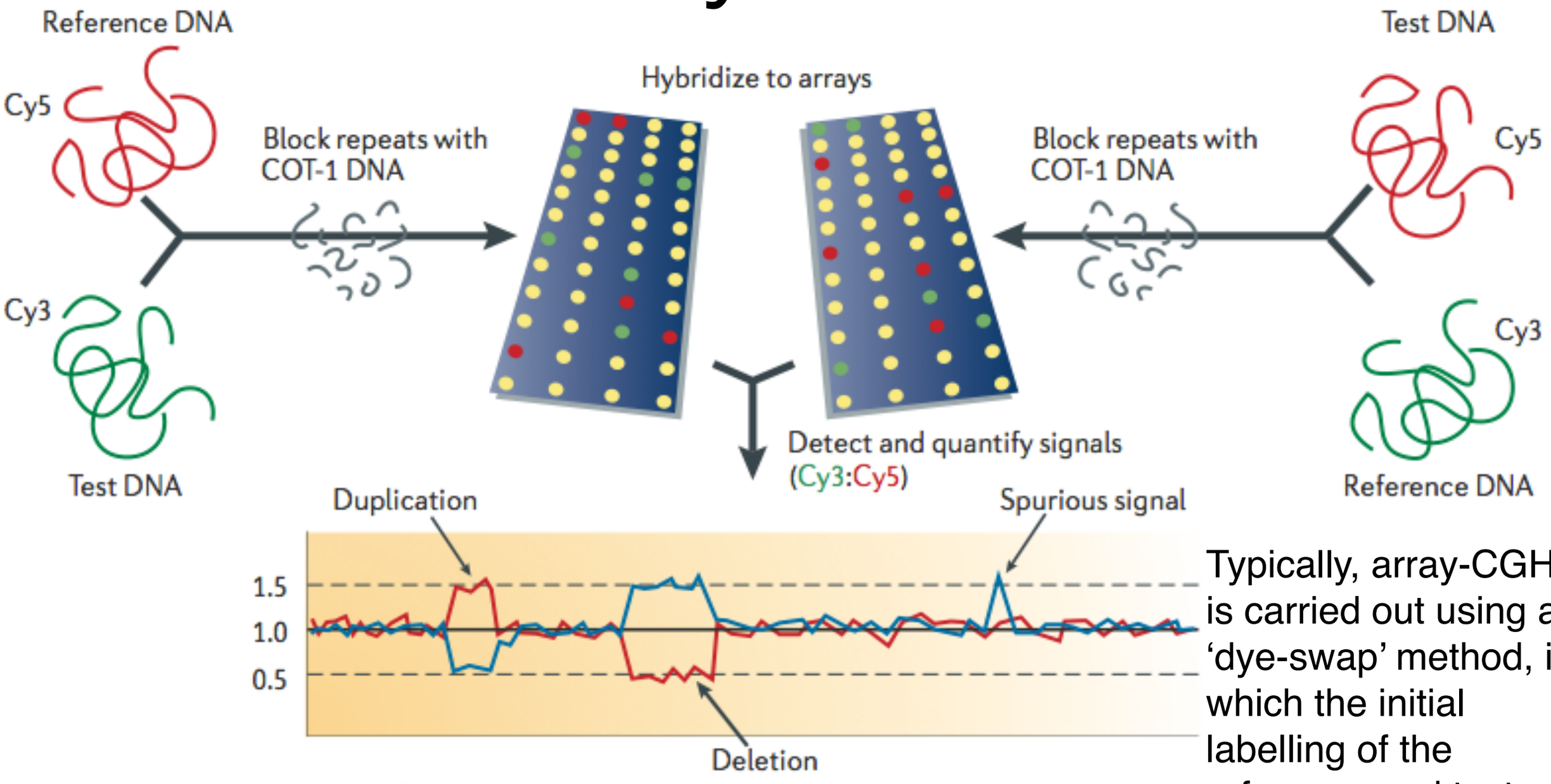
Array CGH



Typically, array-CGH is carried out using a ‘dye-swap’ method, in which the initial labelling of the reference and test DNA samples is reversed for a second hybridization

Feuk, Nature Reviews Genetics 7, 85-97 (2006)

Array CGH



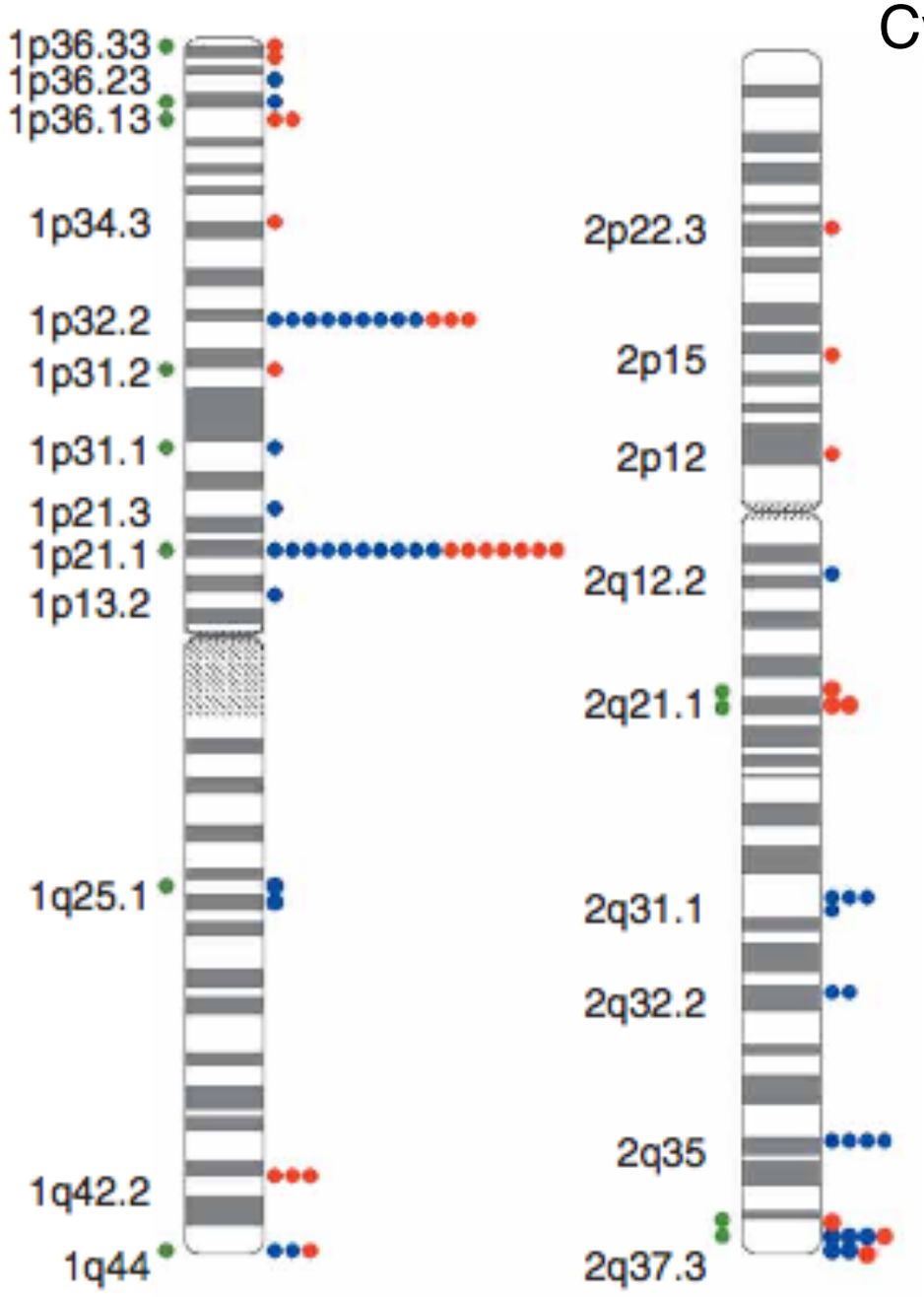
Spurious signals if the reciprocal ratio is not observed.

Typically, array-CGH is carried out using a ‘dye-swap’ method, in which the initial labelling of the reference and test DNA samples is reversed for a second hybridization

Feuk, Nature Reviews Genetics 7, 85-97 (2006)

Array CGH

- CNVs in 39 unrelated healthy people and 16 individuals with known chromosomal abnormalities
- Control: pooled 55 normal individuals).
- Detected all the previously known abnormalities + 255 other CNVs (~12/person)
- 102 CNV occurred in more than one person
- 24 were found in >10%.



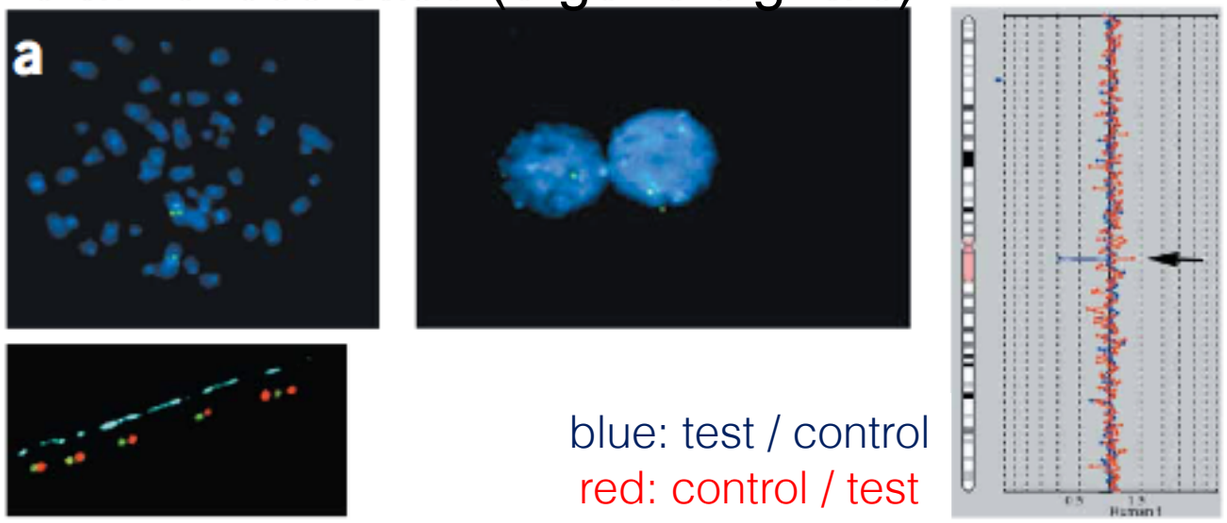
Cytogenetic bands

Dots indicate Individuals with copy
gains
losses

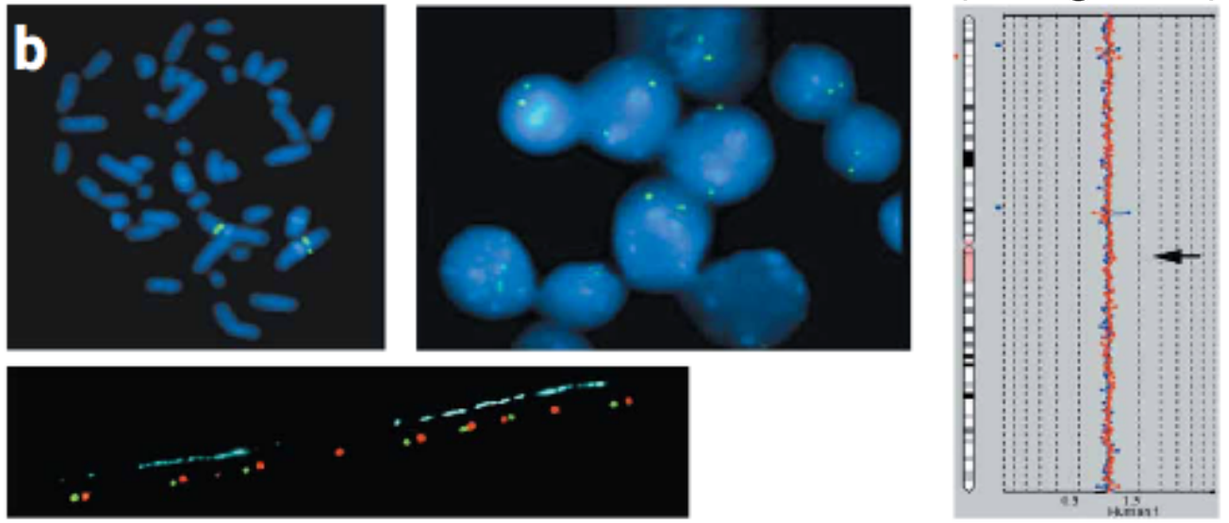
Iafrate et al.,Nature Genetics 36, 949 - 951 (2004)

Array CGH

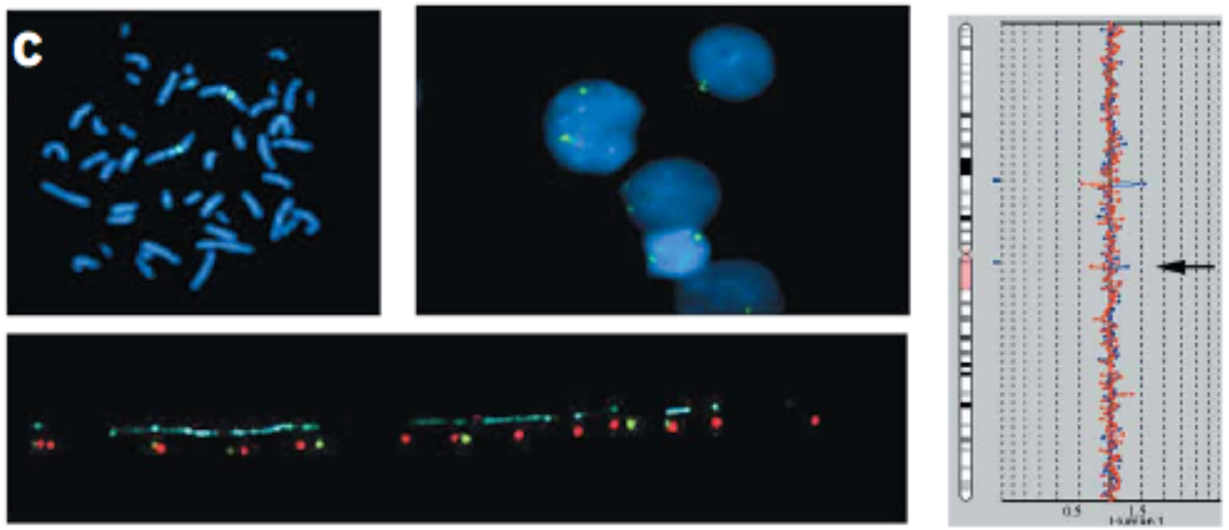
relative loss case (6 gene signals)

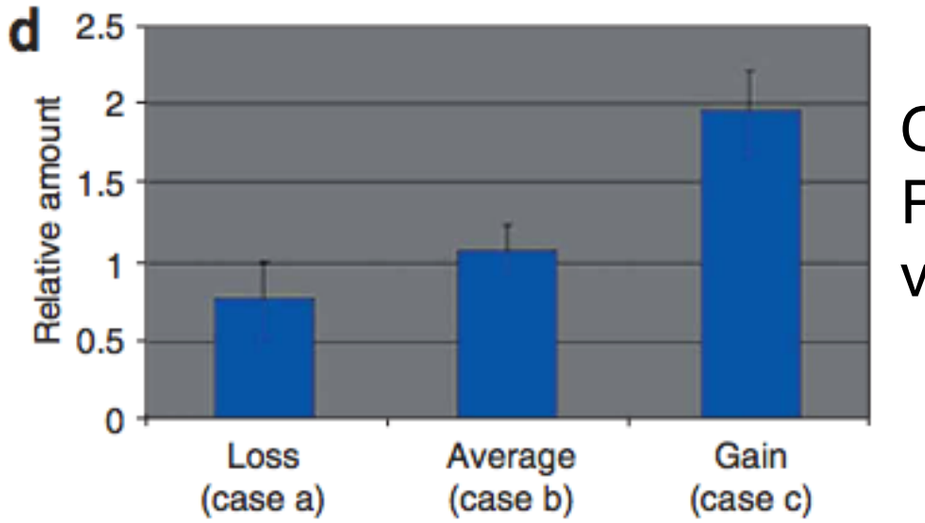


normal case (9 signals)



blue: test / control
red: control / test





Quantitative PCR validation

case with gain (12 signals)

Most common CNV at amylase gene cluster 1p13.3.
Length of the polymorphic region varied from

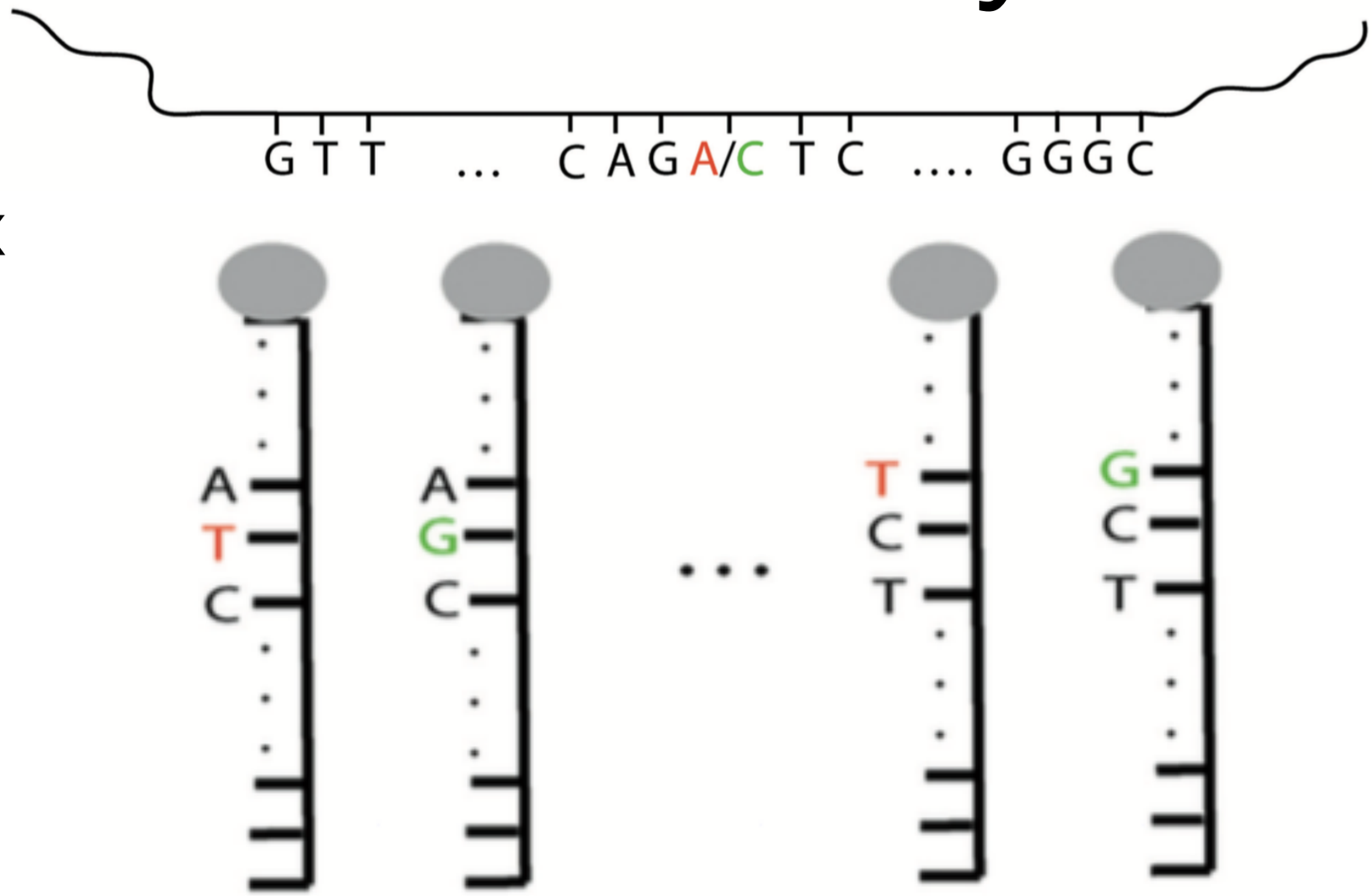
~ 150 kb (a) to

~ 425 kb (c).

Iafrate et al.,Nature Genetics 36, 949 - 951 (2004)

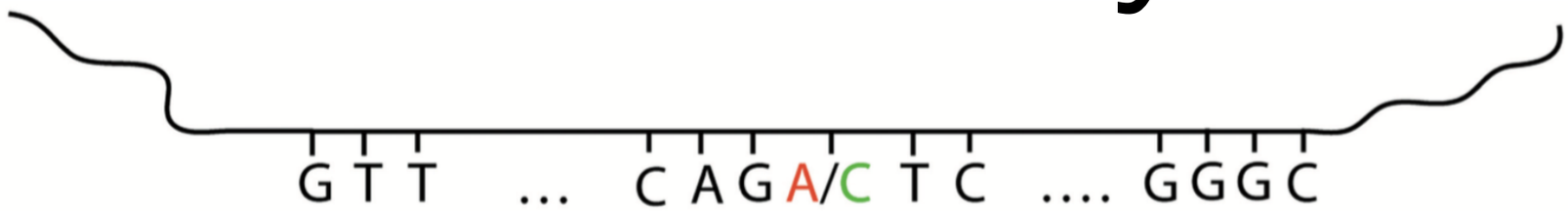
SNP Arrays

Affymetrix



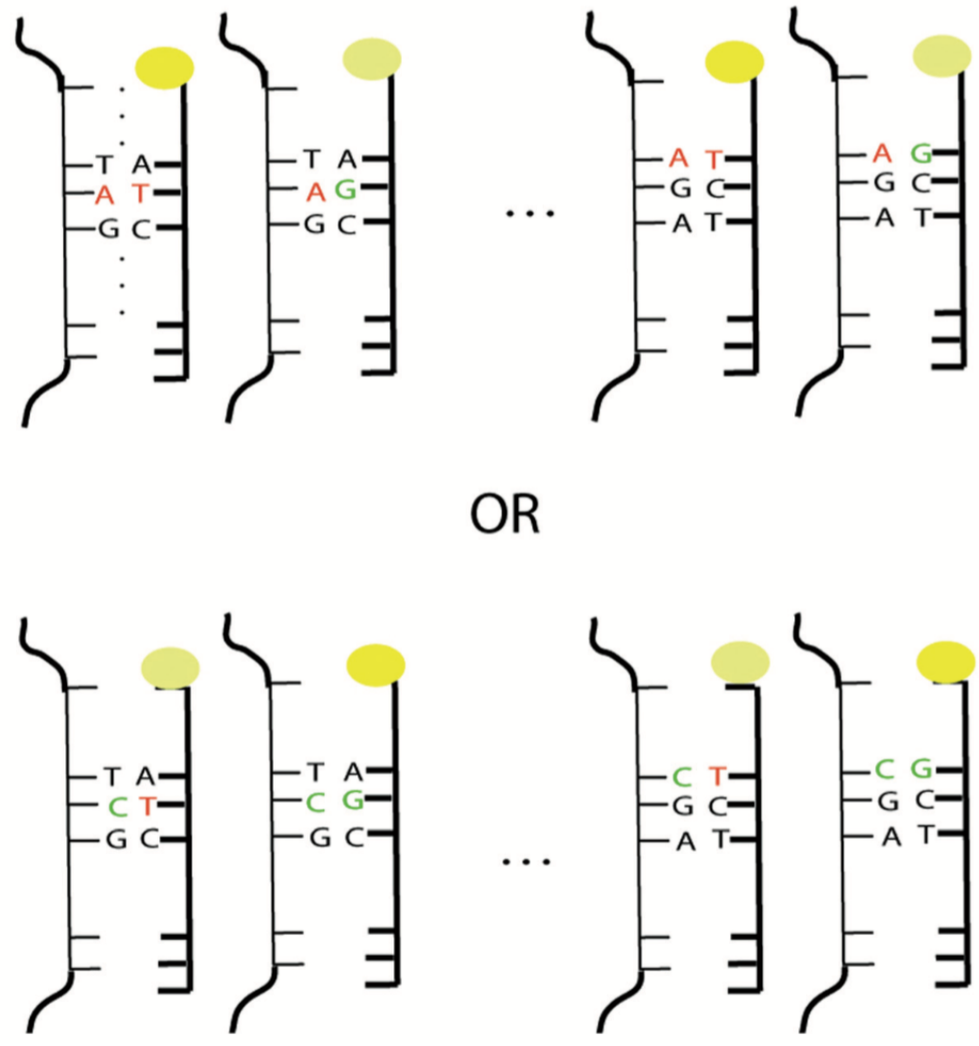
LaFramboise, Nucleic Acids Research,37(13) p 4181 (2009)

SNP Arrays



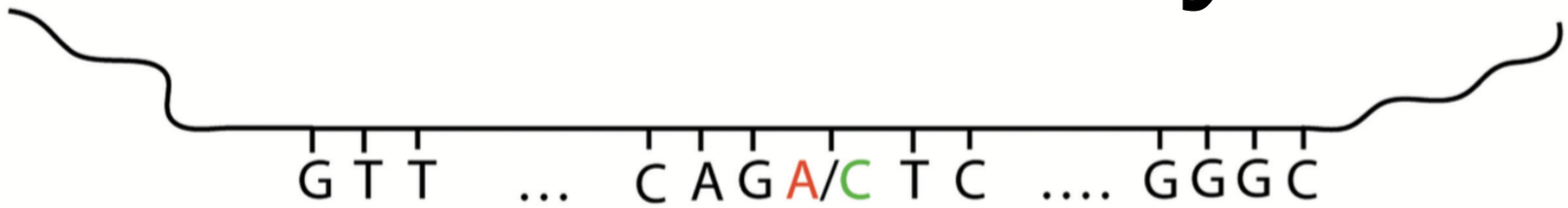
Affymetrix

- 25-mer probes for both alleles
- Location of the SNP locus varies from probe to probe.
- Binding to both probes regardless of the allele.
- More efficient when all 25 bases are complementary
- Dimmer mismatched signal

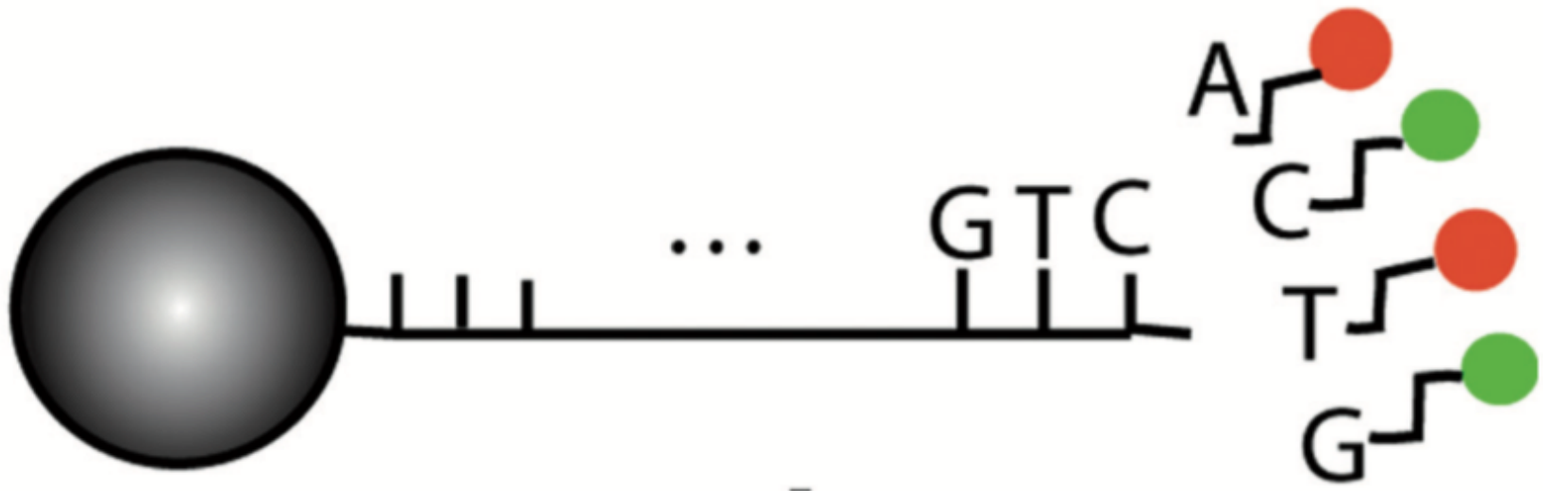


LaFramboise, Nucleic Acids Research,37(13) p 4181 (2009)

BeadArray



Illumina

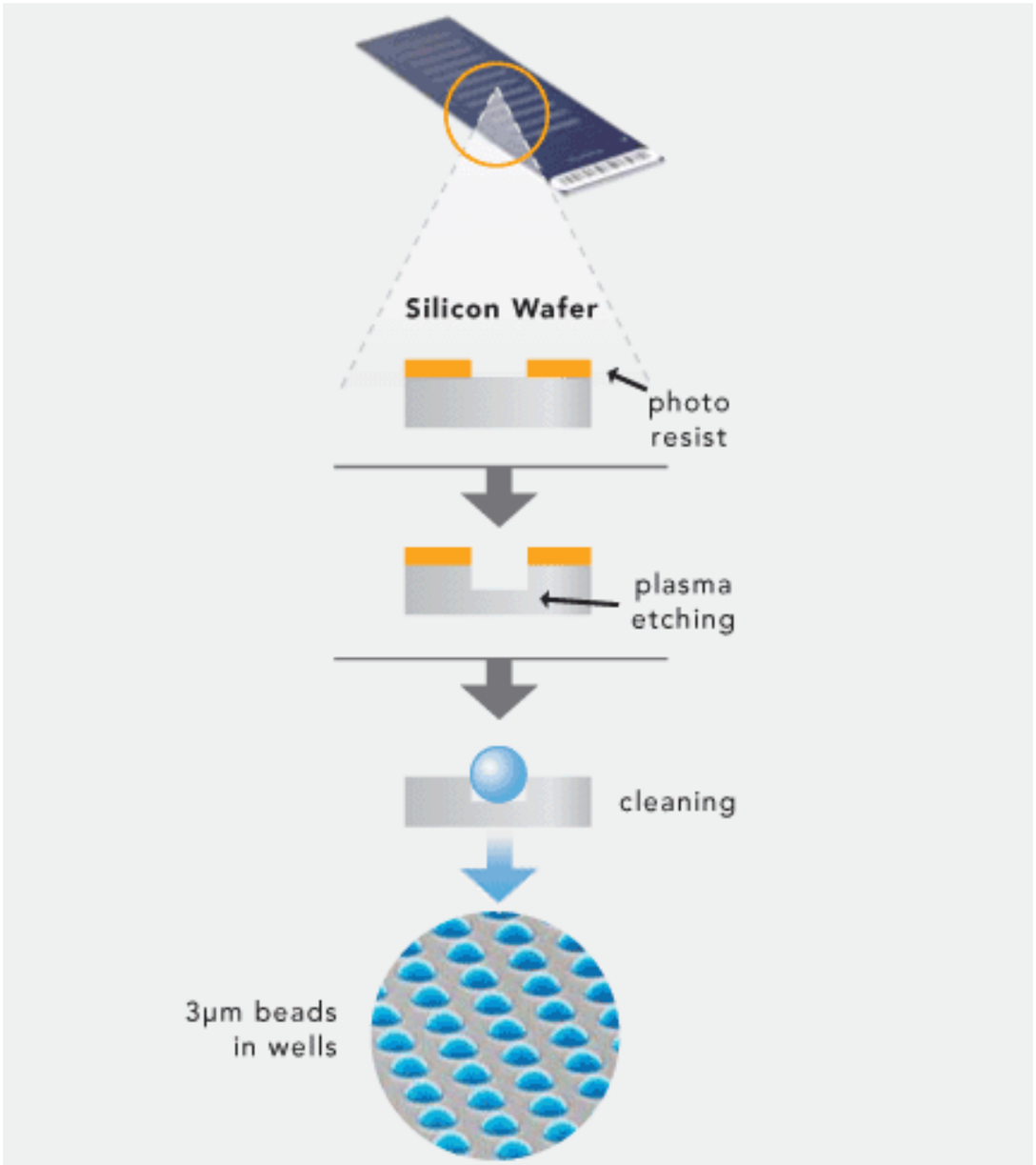


Bead has 50-mer sequence attached complementary to sequence adjacent to SNP site.

LaFramboise, Nucleic Acids Research,37(13) p 4181 (2009)

BeadArray

Illumina

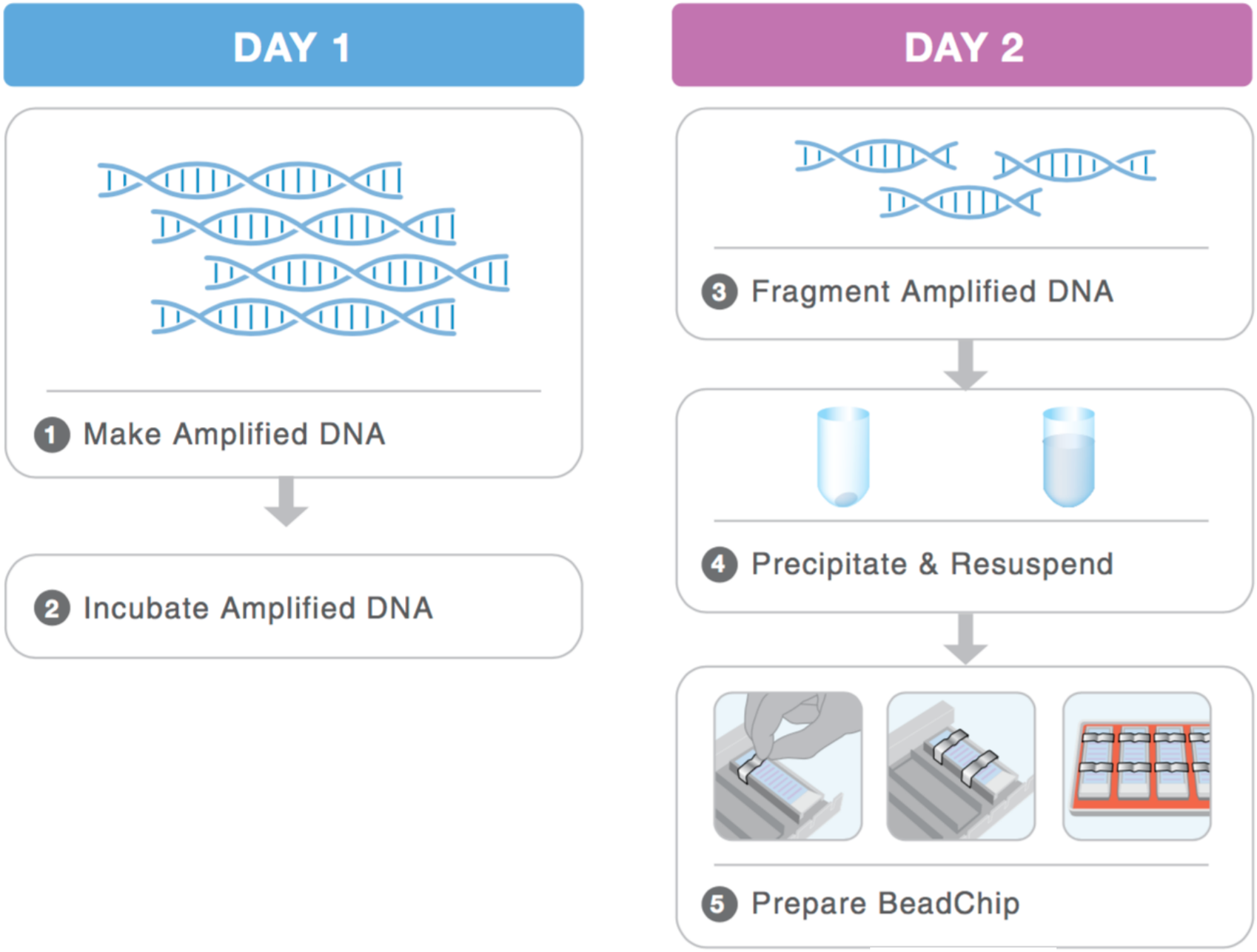


Illumina multi-sample array formats

illumina.com

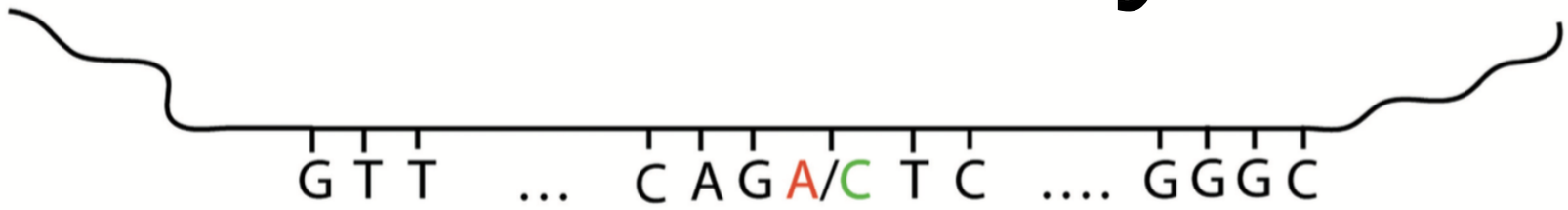
BeadArray

Illumina



illumina.com

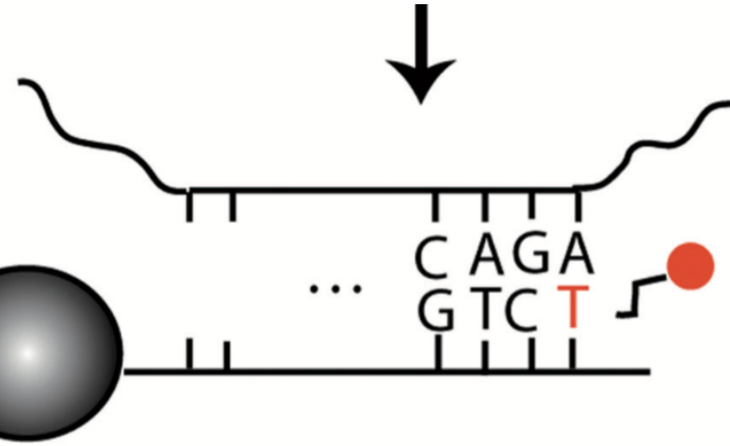
SNP Arrays



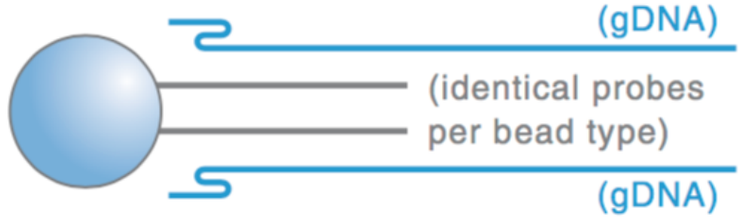
Illumina

Enzymatic single-base extension (T or G) complementary to the allele carried by the DNA (A or C).

Labeled/colored nucleotide signal.



OR

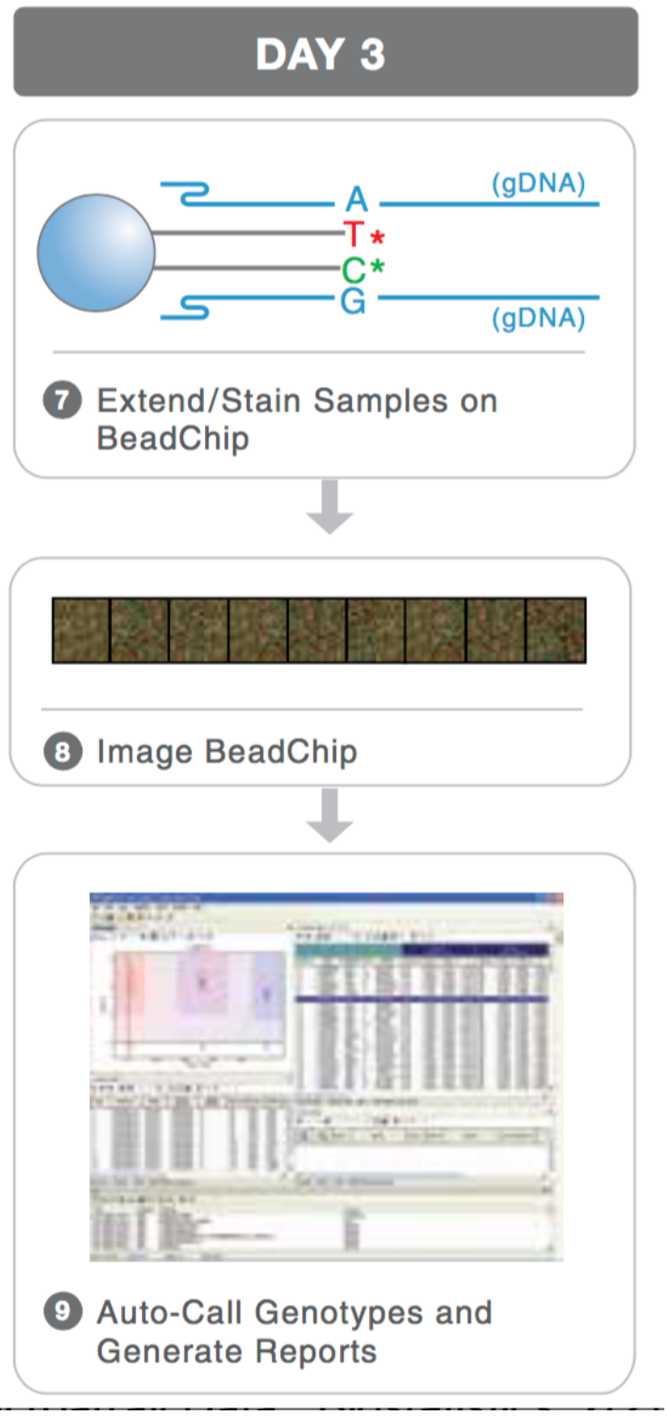


illumina.com

LaFramboise, Nucleic Acids Research,37(13) p 4181 (2009)

SNP Arrays

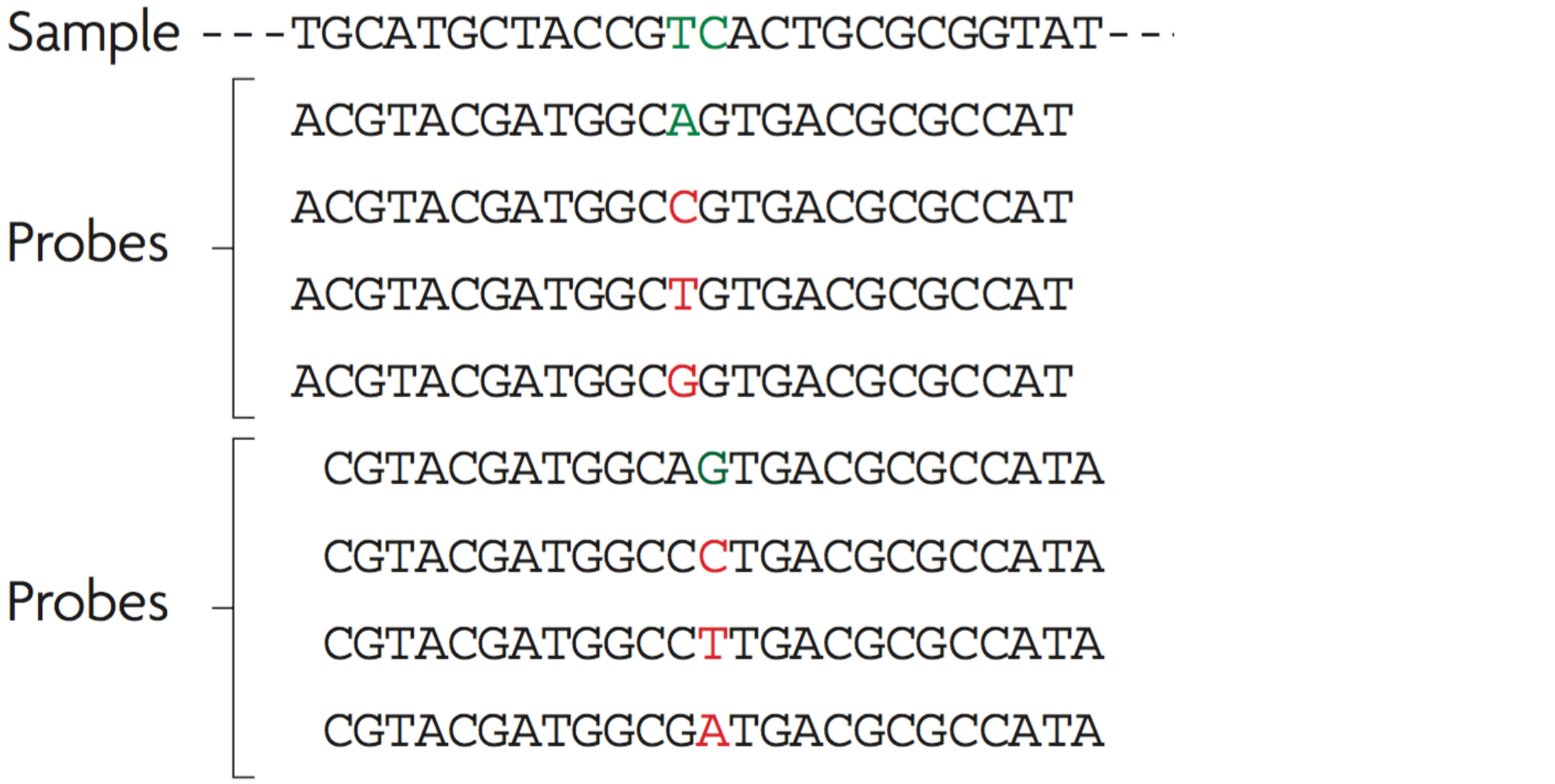
Illumina



illumina.com

Resequencing Microarrays

Resequencing microarrays: probe provided for every possible single-base-pair mismatch

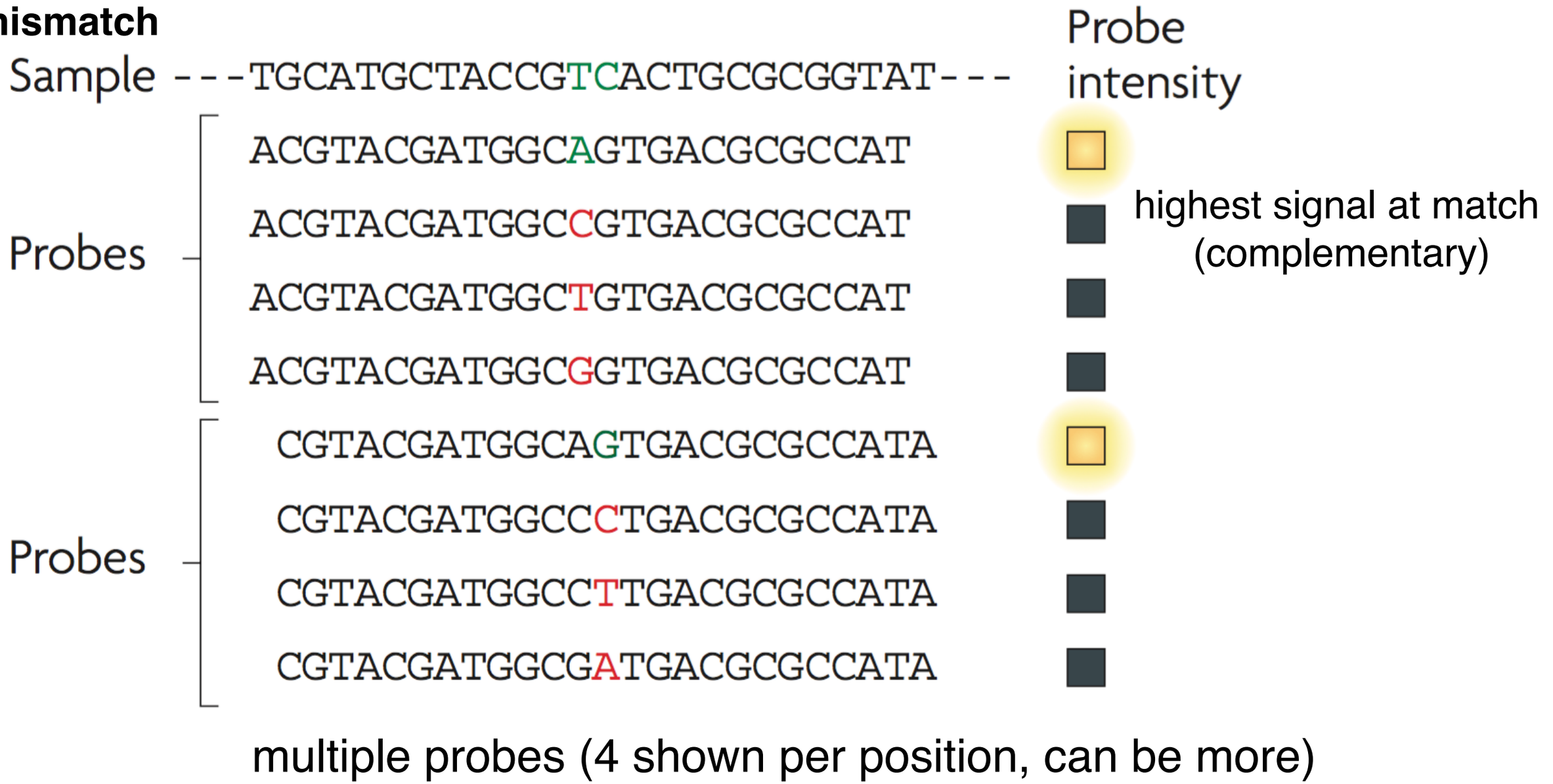


multiple probes (4 shown per position, can be more)

Gresham, Dunham & Botstein, Nature Reviews Genetics 9, 291-302 (2008)

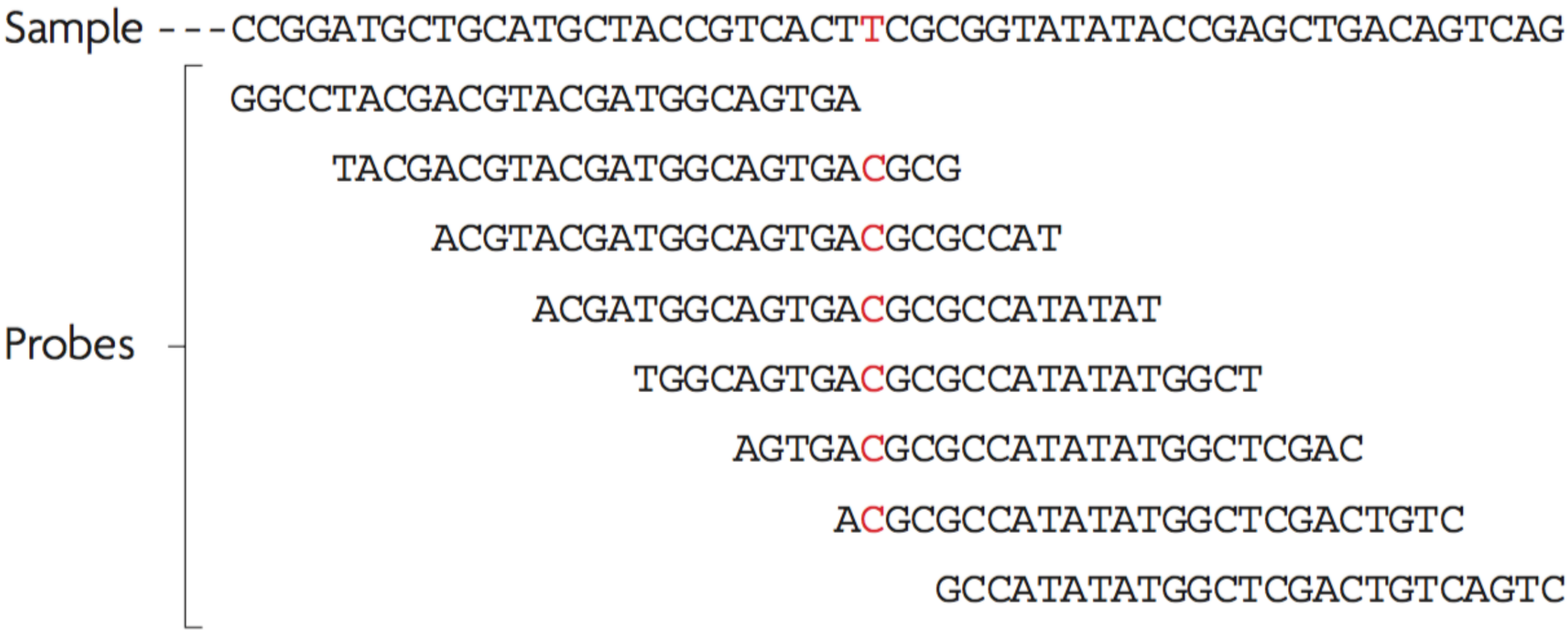
Resequencing Microarrays

Resequencing microarrays: probe provided for every possible single-base-pair mismatch



Gresham, Dunham & Botstein, Nature Reviews Genetics 9, 291-302 (2008)

Hybridization to Candidate Probes



In the absence of resequencing arrays, hybridization of the sample to candidate sequence probes can be used.

Gresham, Dunham & Botstein, Nature Reviews Genetics 9, 291-302 (2008)

Hybridization to Candidate Probes

Mismatches from mutations in sample DNA lead to lower hybridization efficiency (compared to complete complementary sequence matching).

Gresham, Dunham & Botstein, Nature Reviews Genetics 9, 291-302 (2008)

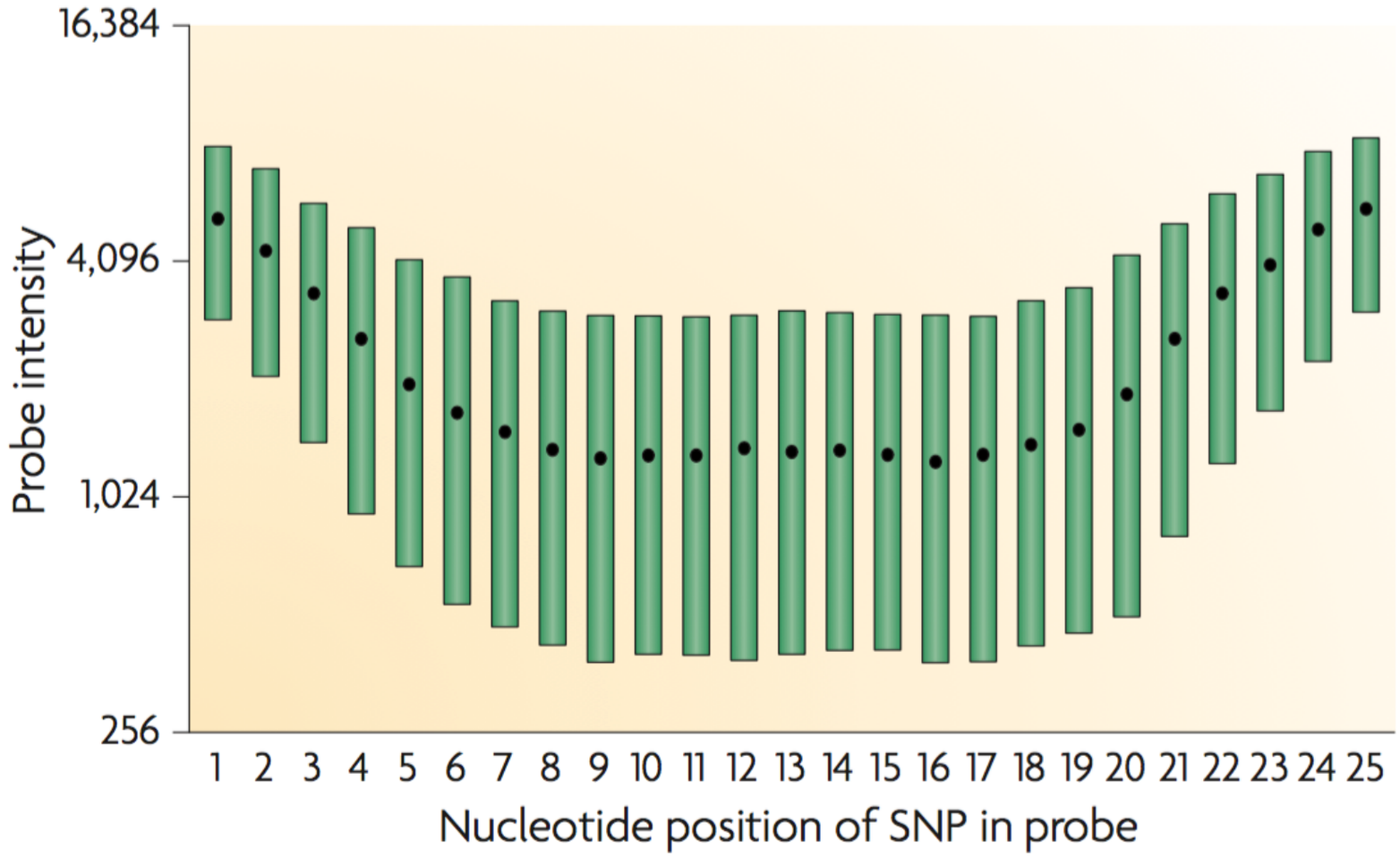
Hybridization to Candidate Probes

- Advantage: fewer probes may be enough to detect variation.
- If mismatch inferred need small-scale sequencing to identify the variant nucleotide.

Gresham, Dunham & Botstein, Nature Reviews Genetics 9, 291-302 (2008)

Resequencing Microarrays

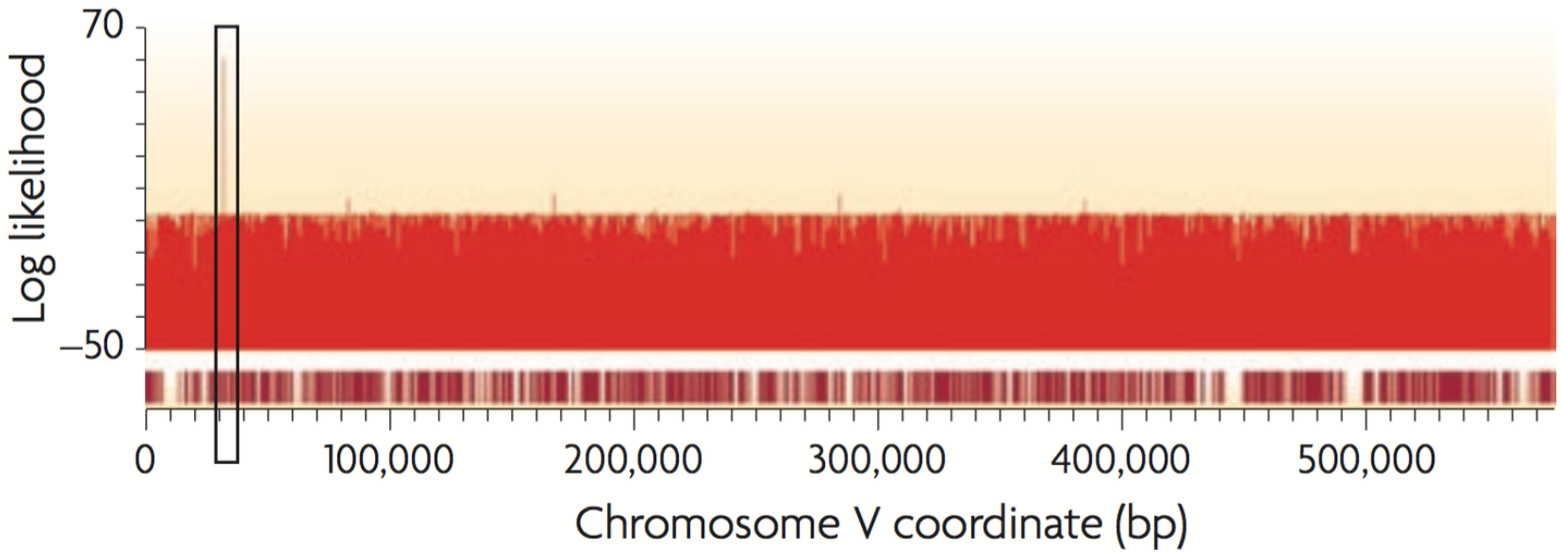
The effect of SNP hybridization as a function of its position in a probe.



- More central positions result in greatest decrease
- SNPs positioned at the end of probes much less likely to result in a significant decrease in hybridization.

Feuk, Nature Reviews Genetics 7, 85-97 (2006)

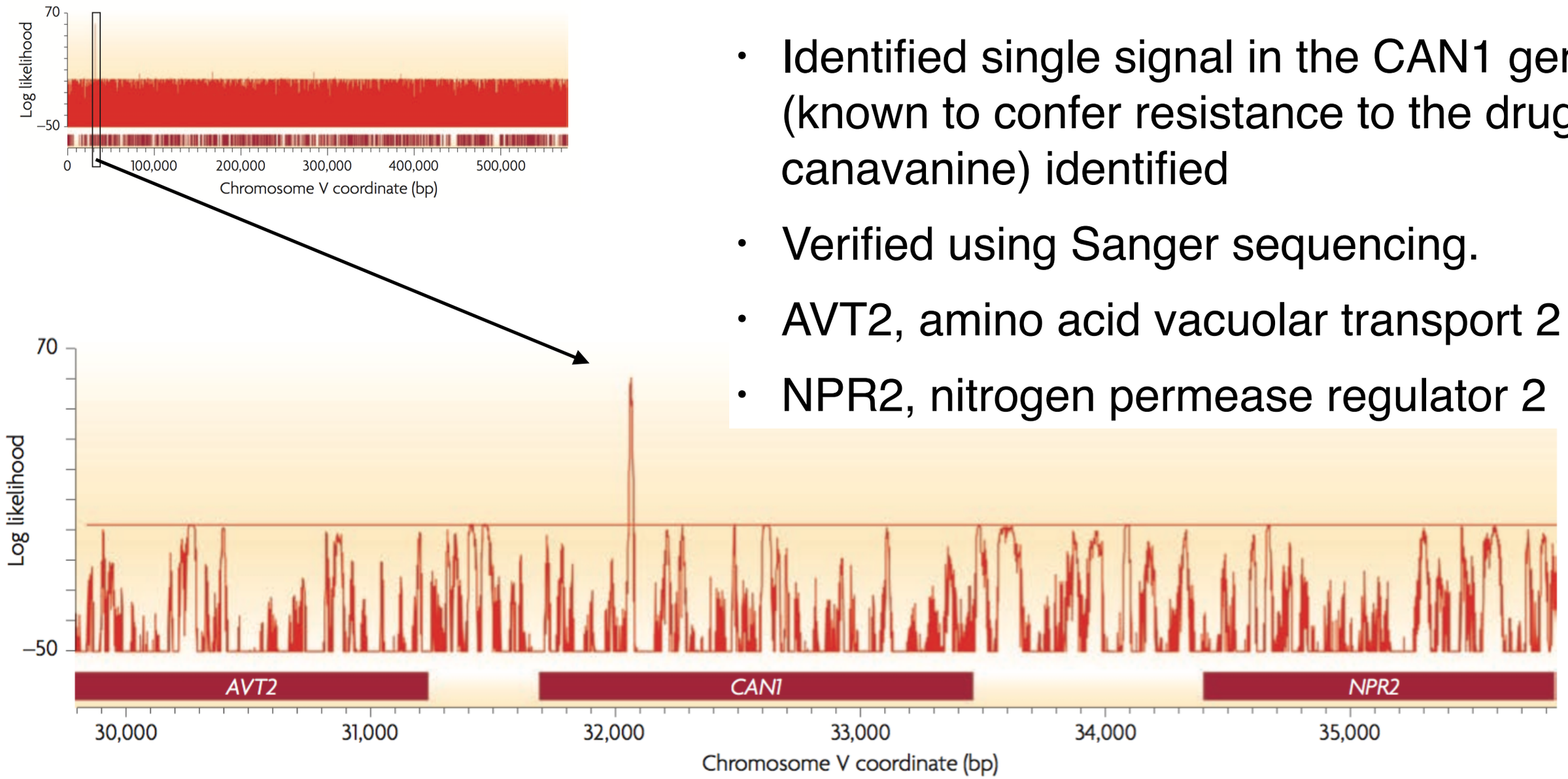
Mutation Detection



- Drug resistant Saccharomyces cerevisiae (budding yeast). Candidate SNPs are identified by a positive log likelihood value.
- Small number of candidate SNPs detected throughout the genome

Feuk, Nature Reviews Genetics 7, 85-97 (2006).
Gresham, D. et al., Science 311, 1932–1936 (2006).

Mutation Detection



- Here on chromosome V
- Identified single signal in the CAN1 gene (known to confer resistance to the drug canavanine) identified
- Verified using Sanger sequencing.
- AVT2, amino acid vacuolar transport 2
- NPR2, nitrogen permease regulator 2

Feuk, Nature Reviews Genetics 7, 85-97 (2006).
Gresham, D. et al., Science 311, 1932–1936 (2006).

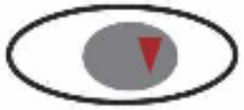
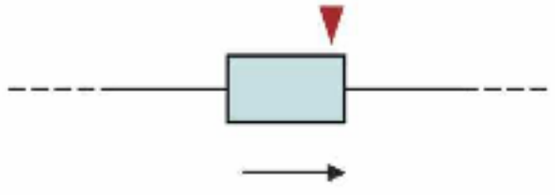
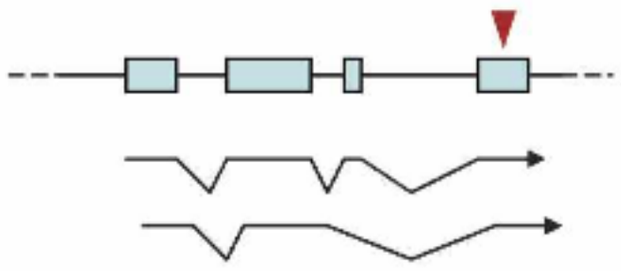
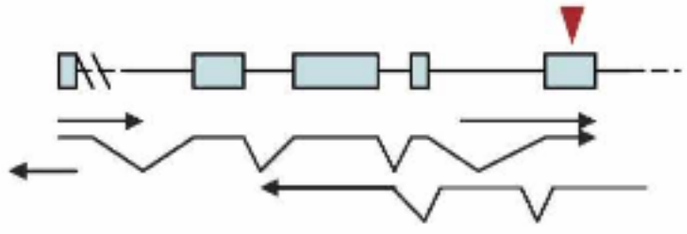
Gene Expression

Gene Expression

- Transcriptomics – Collective of all expressed transcripts in a given cell .
- Gene “expressed” when transcribed to RNA
- Level of RNA produced from a gene is controlled by:
  - Transcription
  - Stability/Degradation
- Dynamic (Vs. static genome)
- Tissue specific

Gene expression may also refer to levels of protein instead

Gene Expression

Time Period	Key Facts	Gene Model	Phenotype
1900s	Cells contain genes. (*location, composition and structure unknown*)		
1950-1960s	Gene located on specific regions of chromosomes. (*composed of DNA transcribed into RNA*)		
1970-2000s	Genes clustered together in gene-rich regions separated by large intergenic regions of "junk DNA". (*mostly transcribed into protein-coding RNAs which can be spliced into multiple isoforms*)	 Spliced RNA Isoforms	
2001-present	Genes have distal 5` TSSs and regulatory regions and share their genomic sequences with long and short noncoding RNAs encoded on both strands.	 Interleaved RNAs	 plus other more subtle phenotypes due to mutations or interactions with interleaved ncRNAs

Gingeras, Genome Res. 17(6):682-90 (2007)

Gene Expression

Transcriptome has mainly

- mRNA (messanger RNA)
- rRNA (ribosomal)
- tRNA (transfer)

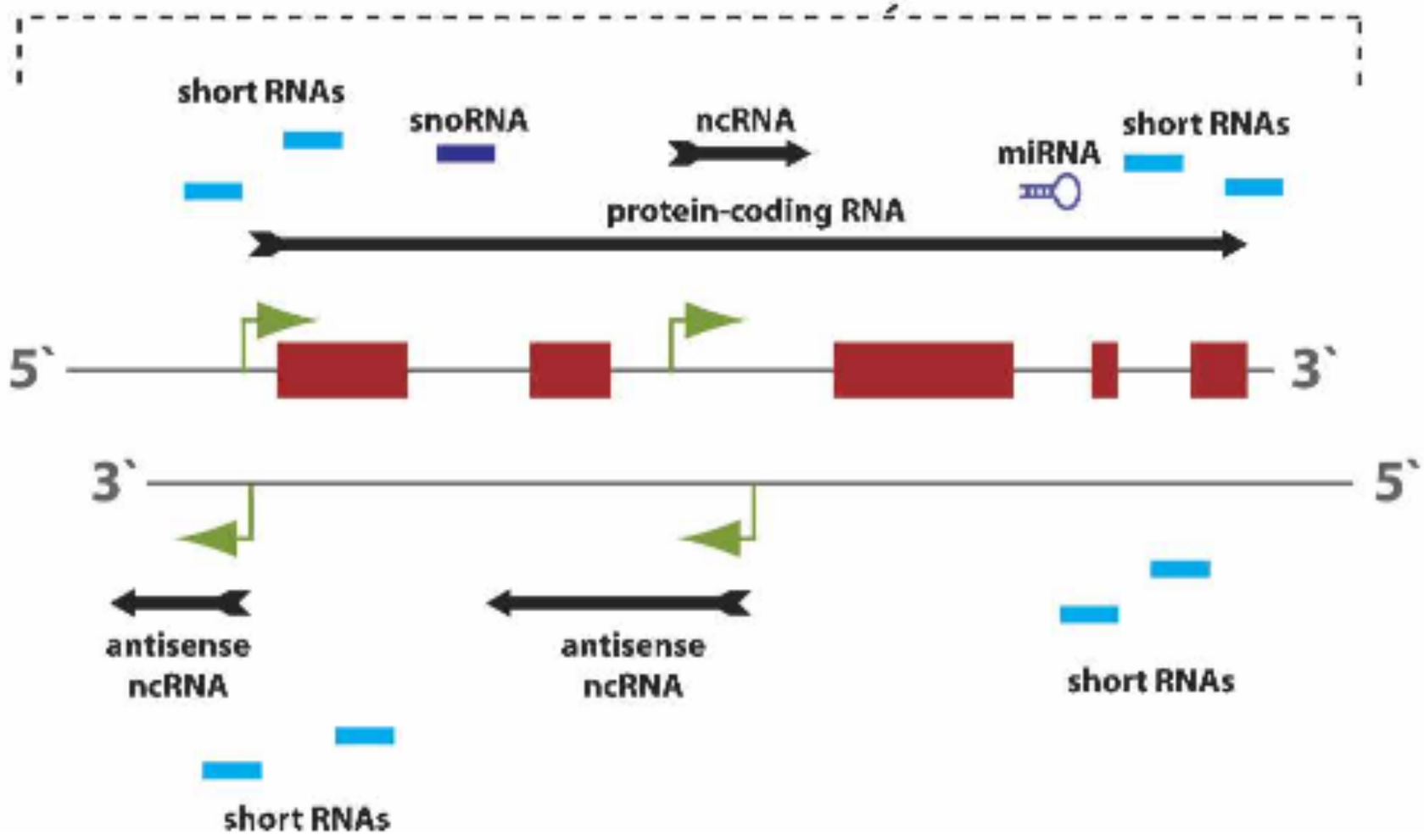
But also

- ncRNA (non-coding)
  - lincRNA (long intergenic ncRNA);
  - snRNA (small nuclear)
  - snoRNA (small nucleoral)
  - miRNA (micro)
  - siRNA (small interfering)
  - piRNAs (PIWI-interacting) RNAs
  - …

Gene Expression

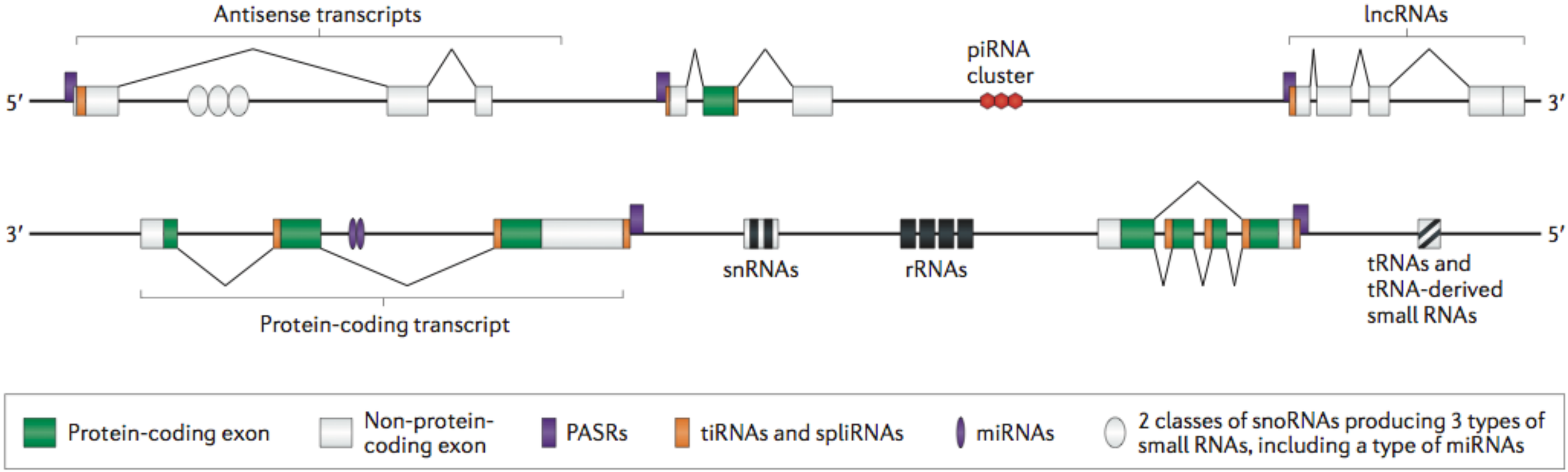
multiple start sites

coding
and
noncoding



Gingeras, Genome Res. 17(6):682-90 (2007)

Gene Expression



Transcriptional units not depicted to scale

PASRs, promoter-associated short RNAs

tiRNAs, transcription initiation RNAs

spliRNAs, splice site RNAs

Morris & Mattick, Nature Reviews Genetics 15, 423–437 (2014)

Gene Expression

ncRNA*	No. of known transcripts	Transcript lengths (nucleotides; nt)	Functions
Precursors to short RNAs			
miRNA	1,756	>1,000	Precursors to short (21–23 nt) regulatory RNAs
snoRNA	1,521	>100	Precursors to short (60–300 nt) RNAs that help to chemically modify other RNAs
snRNA	1,944	1,000	Precursors to short (150 nt) RNAs that assist in RNA splicing
piRNA	89	Unknown	Precursors to short (25–33 nt) RNAs that repress retrotransposition of repeat
tRNA	497	>100	Precursors to short (73–93 nt) transfer RNAs

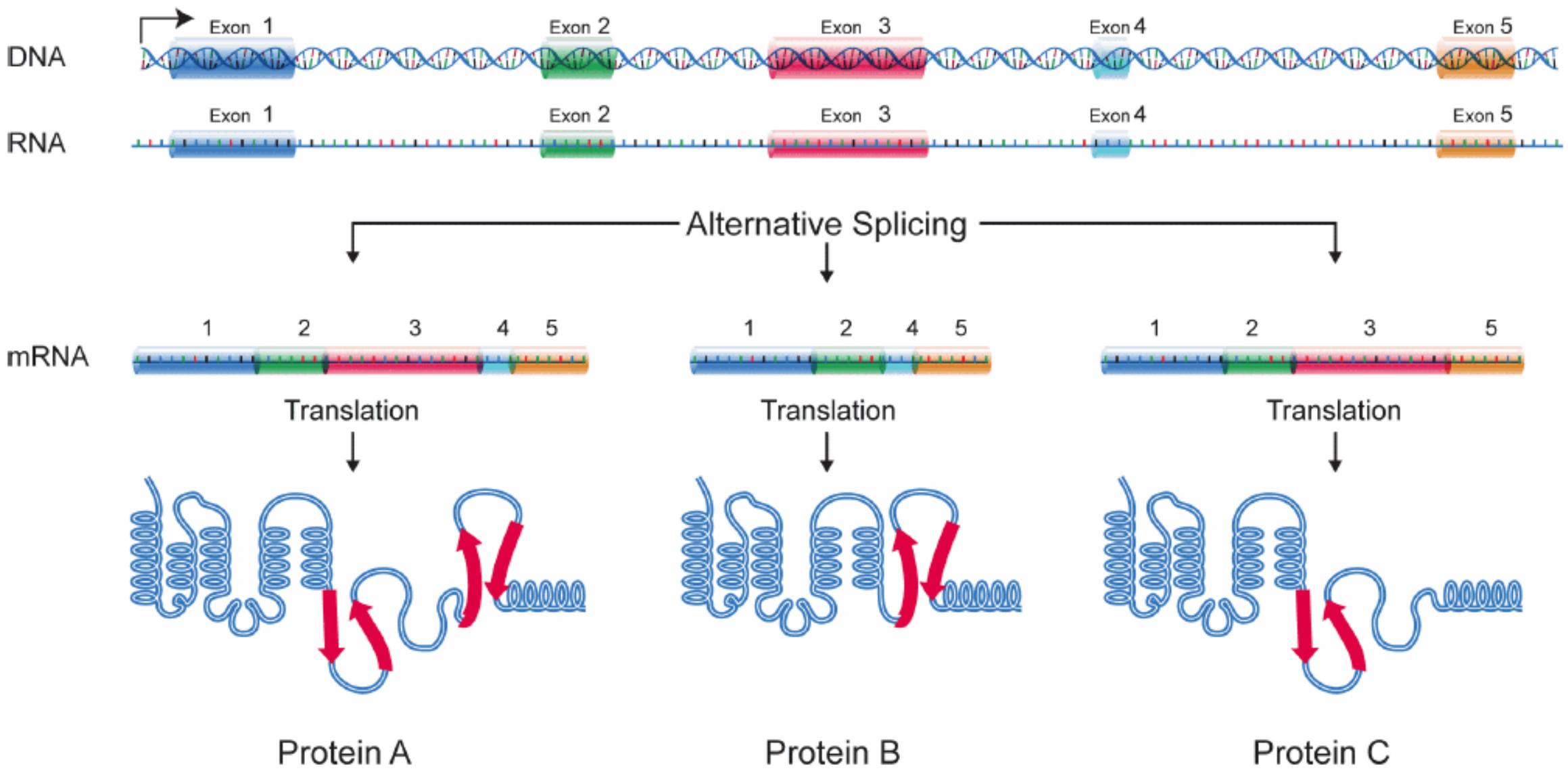
Kowalczyk & Higgs,Nature 482, 310–311 (2012)

Gene Expression

ncRNA*	No. of known transcripts	Transcript lengths (nucleotides; nt)	Functions
Long ncRNAs			
Antisense ncRNA	5,446	100–>1,000	Mostly unknown, but some are involved in gene regulation through RNA interference
Enhancer ncRNA (eRNA)	>2,000	>1,000	Unknown
Enhancer ncRNA (meRNA)	Not fully documented	As variable as the length of mRNAs	Unknown, but they resemble alternative gene transcripts
Intergenic ncRNA	6,742	10^2–10^5	Mostly unknown, but some are involved in gene regulation
Pseudogene ncRNA	680	10^2–10^4	Mostly unknown, but some are involved in regulation of miRNA
3' UTR ncRNA	12	>100	Unknown

Kowalczyk & Higgs,Nature 482, 310–311 (2012)

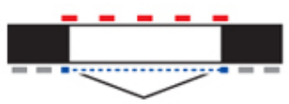
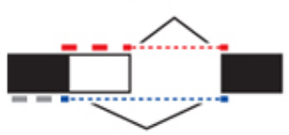
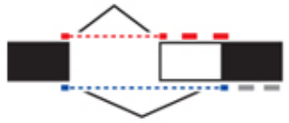
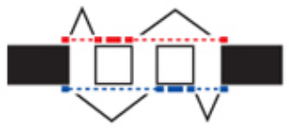
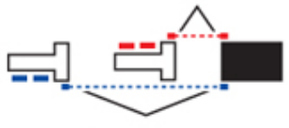
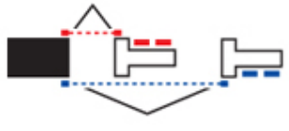
Gene Expression



NHGRI, https://www.genome.gov/Images/EdKit/bio2j_large.gif

Gene Expression

Pervasive tissue-specific regulation of alternative mRNA isoforms.

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

Wang et al., Nature 456, 470-476 (27 November 2008)

Gene Expression Methods

- Northern Blots
  - Measure RNA levels by hybridization of a labeled probe to total RNA after gel electrophoresis and transfer to a membrane.
- Reporter Genes
  - Use of activity of a foreign protein to measure the amount of transcription from a promoter.
- Quantitative real-time RT-PCR.
  - Uses amplification by sequence-specific primers
- Microarrays
  - high throughput
- RNA-Sequencing

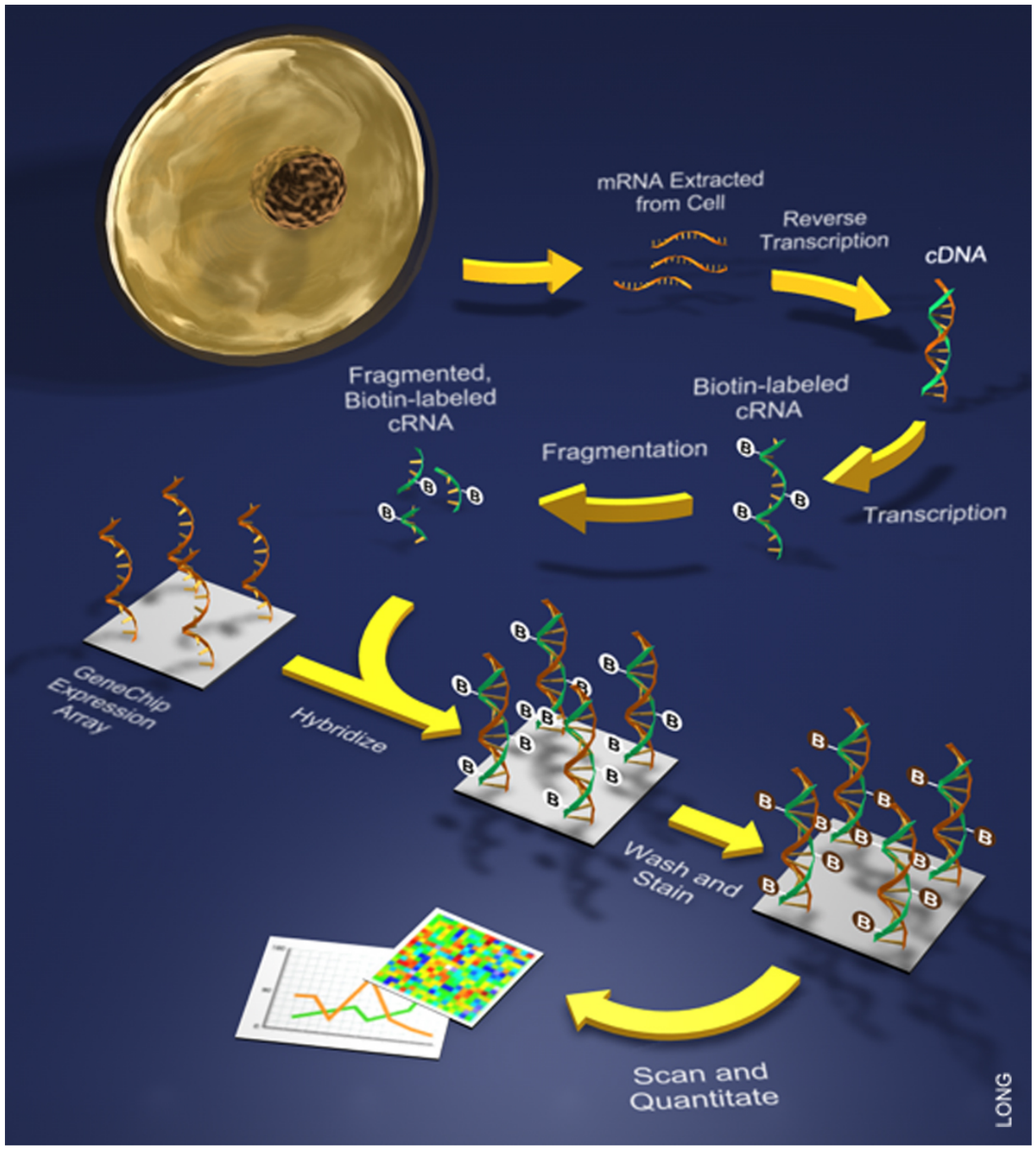
1-Color Expression Arrays

mRNA extracted

reverse transcribe to cDNA (complementary DNA),

Scan

Relative Expression



The Science Creative Quarterly, scq.ubc.ca (Jiang Long)

2-Color Expression Arrays

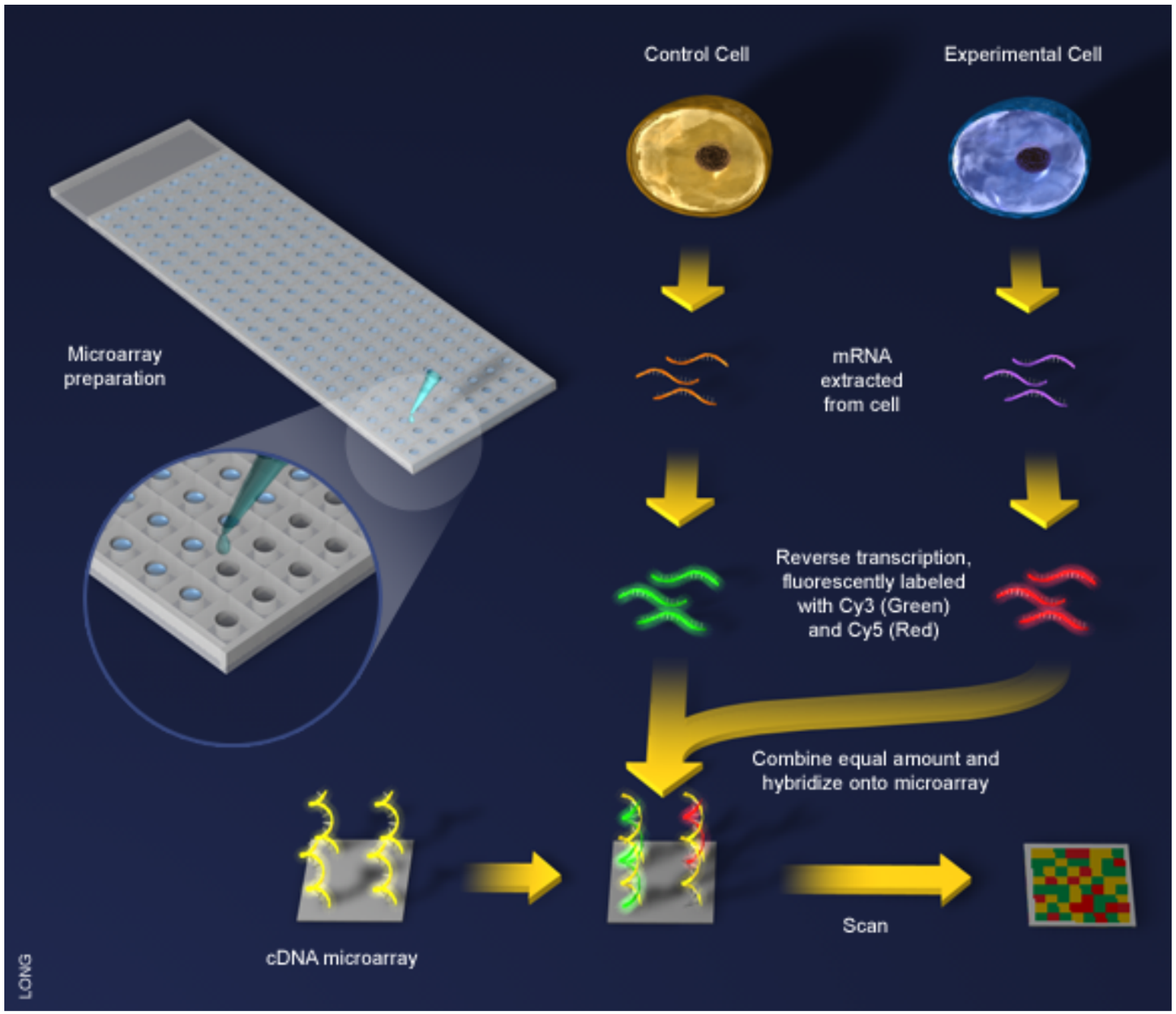
mRNA extracted

reverse transcribe to cDNA (complementary DNA),

Fluorescent labels (**Cy3**, **Cy5**)

Reference Relative ratios

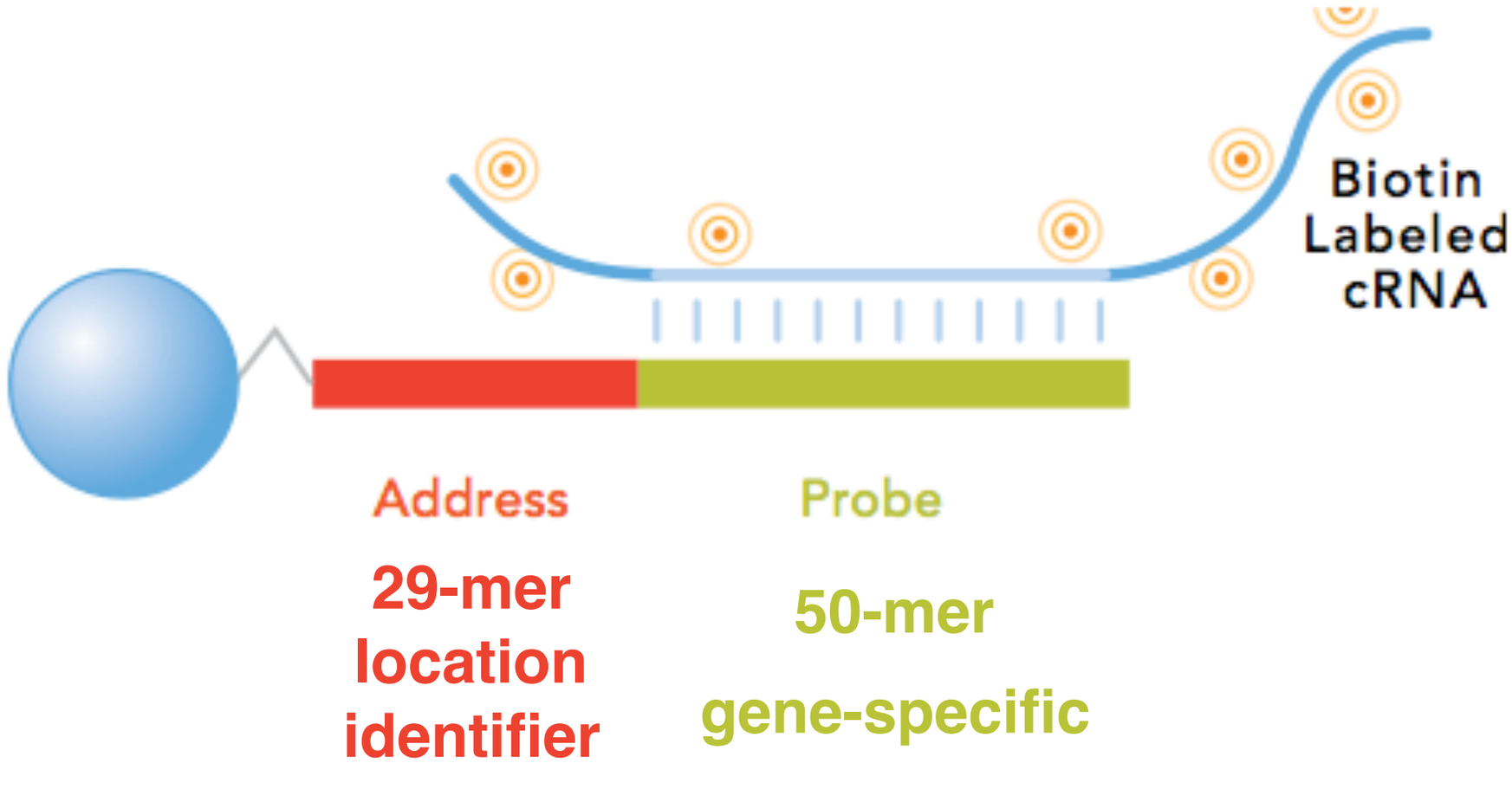
(control for amounts & conditions)



The Science Creative Quarterly, scq.ubc.ca (Jiang Long)

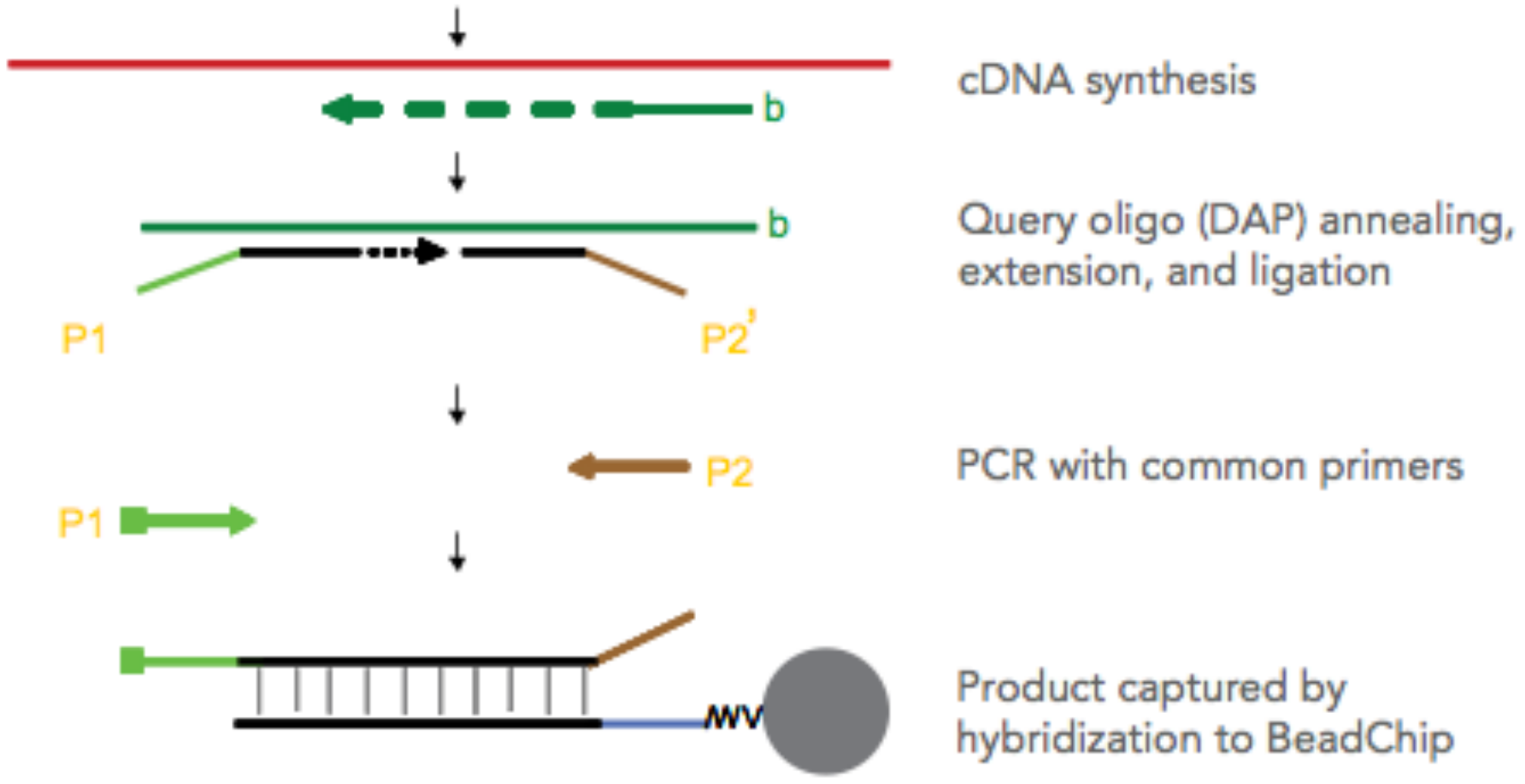
Expression BeadChips

Up to 30 beads per probe average



illumina.com

Expression BeadChips



illumina.com

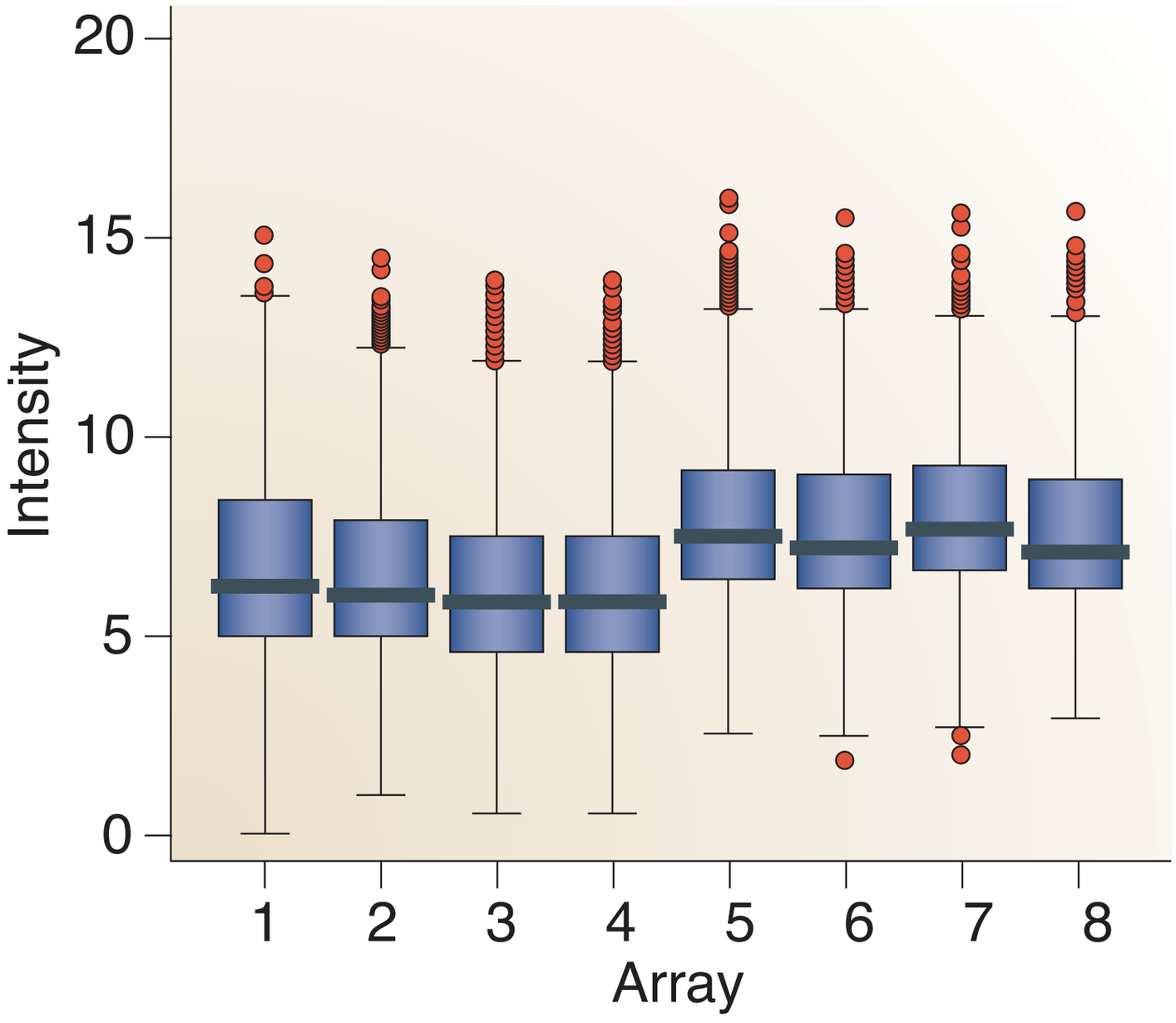
Expression BeadChips

Probes	Description	Human HT-12 v4.0* 12-sample	Mouse WG-6 v2.0 6-sample	Mouse Ref-8 v2.0 8-sample	Rat Ref-12 12-sample	Human WG DASL HT* 12-sample
RefSeq Content						
NM	Coding transcript, well-established annotation	28,688	26,766	24,854	6,277	27,253
XM	Coding transcript, provisional annotation	11,121	6,856	796	15,983	426
NR	Non-coding transcript, well-established annotation	1,752	56	47	1	1,580
XR	Non-coding transcript, provisional annotation	2,209			12	26
Source	RefSeq source release	Human RefSeq Rel 38	Mouse RefSeq Rel 22		Rat RefSeq Rel 16	Human RefSeq Rel 38
Supplementary Content						
UniGene	Experimentally confirmed mRNA sequences that align to EST clusters	3,461			250	
RIKEN FANTOM2	Exemplar protein-coding sequences from the RIKEN FANTOM2 database		5,659			
RefSeq Release 5	Transcripts with NM and XM annotation in RefSeq Release 5 (Build 33.1)		3,573			
MEEBO	Probes to transcripts that do not align with 100% accuracy to RefSeq, but are confirmed as valid mRNA mapping to clusters in Expressed Sequence Tag databases[6]		2,371			
Total		47,231	45,281	25,697	22,523	29,285

*> 99.98% of the bead types are present on any HumanHT-12 array

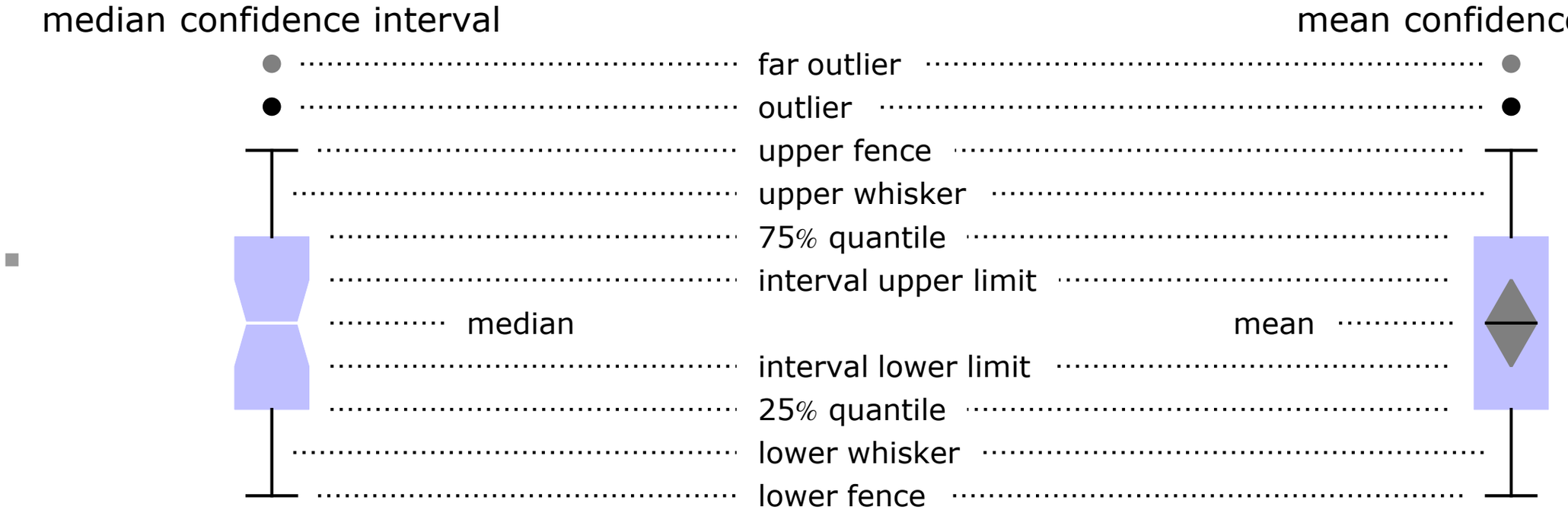
illumina.com

Box plots



Allison et al., Nature Reviews Genetics 7, 55-65 (January 2006)

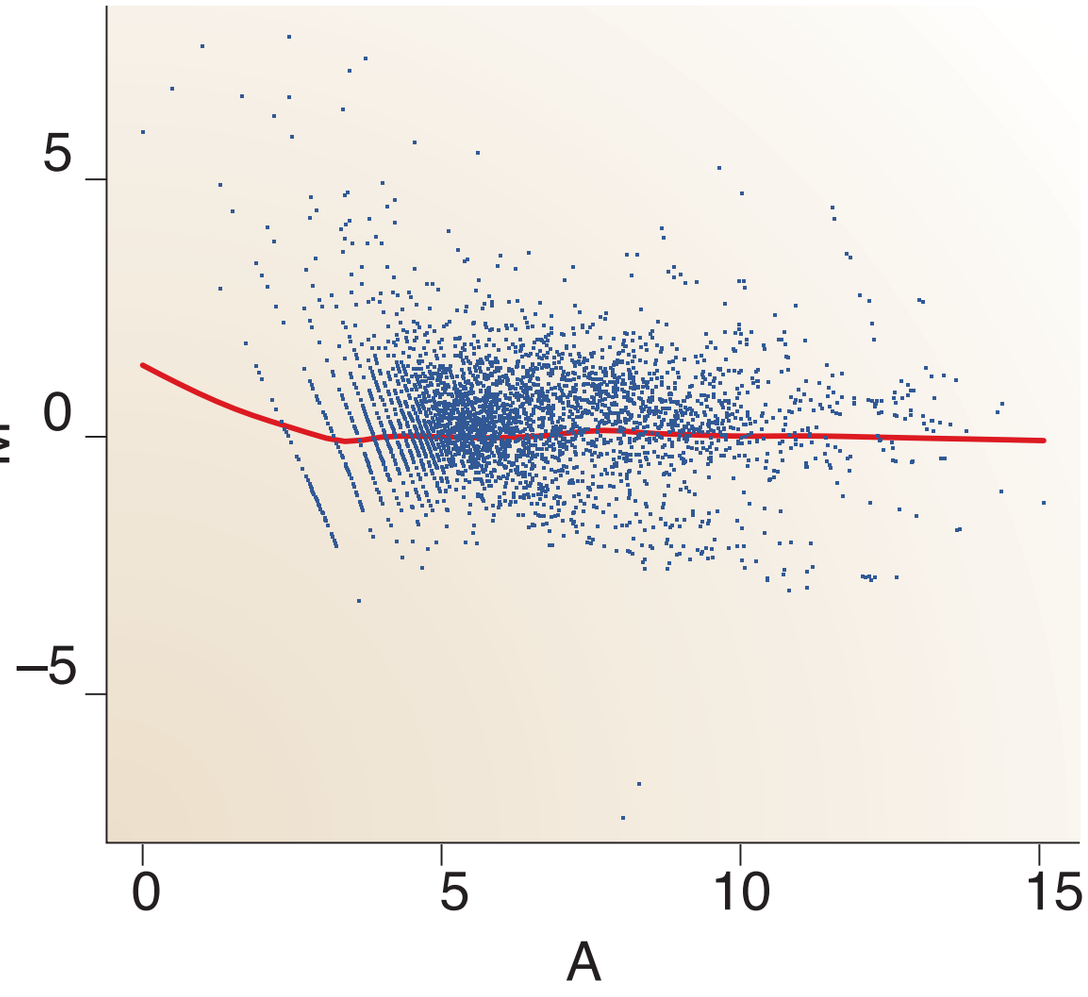
Box plots


median confidence interval
mean confidence interval
far outlier
outlier
upper fence
upper whisker
75% quantile
interval upper limit
median
mean
interval lower limit
25% quantile
lower whisker
lower fence

MA Plots

MA plots are used to detect intensity dependent artifacts

Used for normalizing two-colour cDNA microarrays, e.g. with **red (R)** and **green (G)** dyes.



Allison et al., Nature Reviews Genetics 7, 55-65 (January 2006)

MA Plots

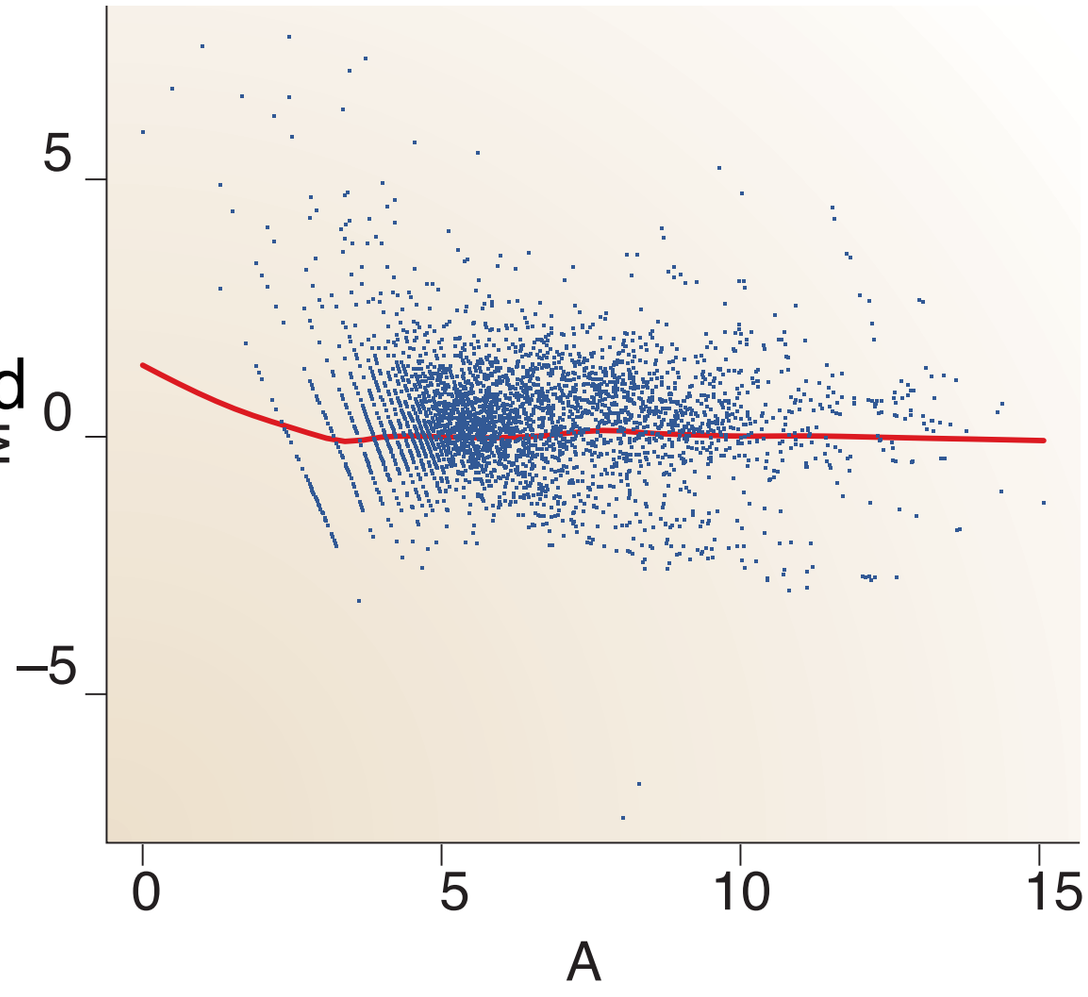
Define:

$M = \log_2(\mathbf{R}/\mathbf{G})$

$A = (1/2) \times \log_2(\mathbf{R} \times \mathbf{G})$.

Can fit with a lowess curve, which is used to normalize the gene-expression measurements.

Lowess, locally weighted scatterplot smoothing



Allison et al., Nature Reviews Genetics 7, 55-65 (January 2006)

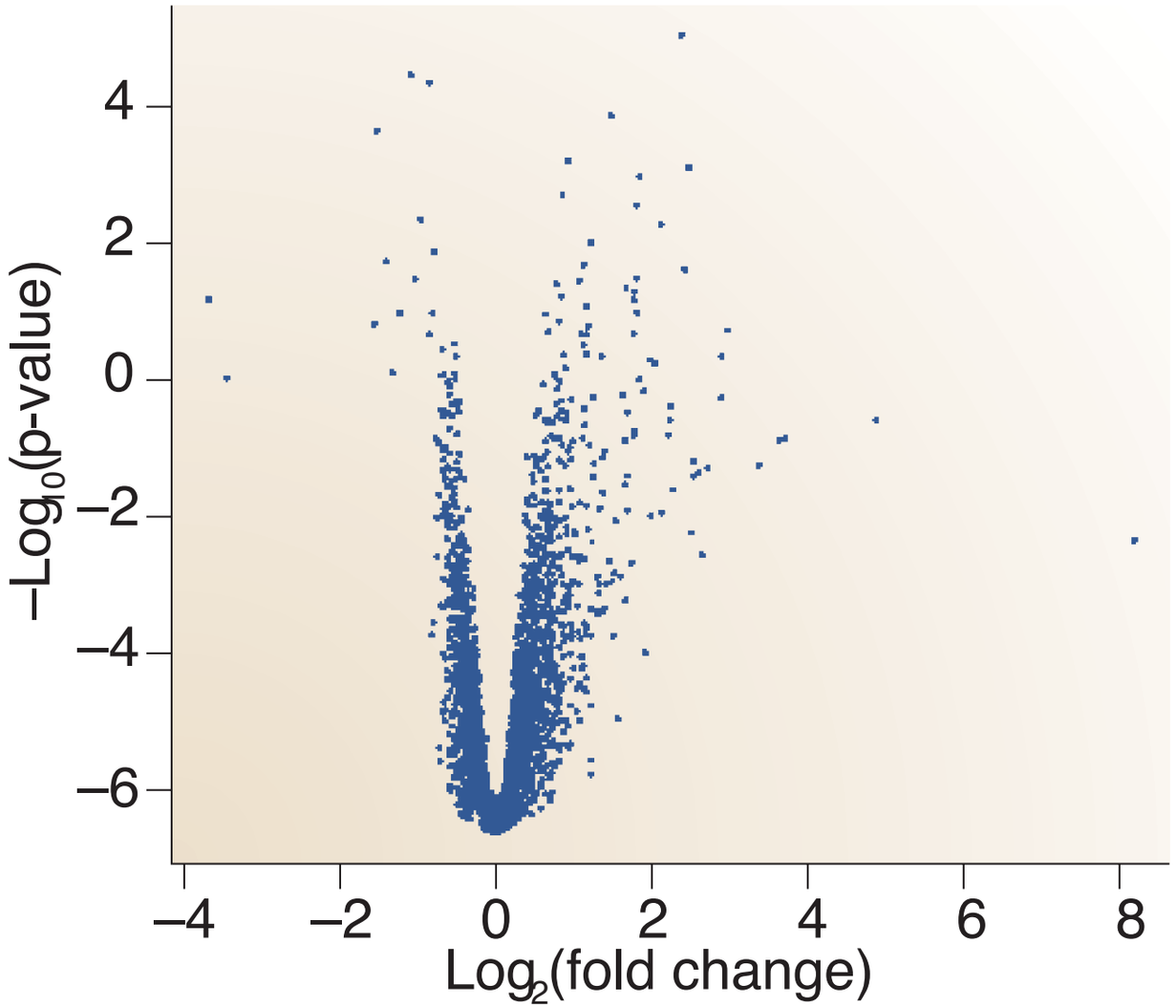
Volcano Plots

Statistical significance Vs. *fold changes* for all genes on the array.

Plotted as

– $\log_{10}$(*p*-values) Vs.

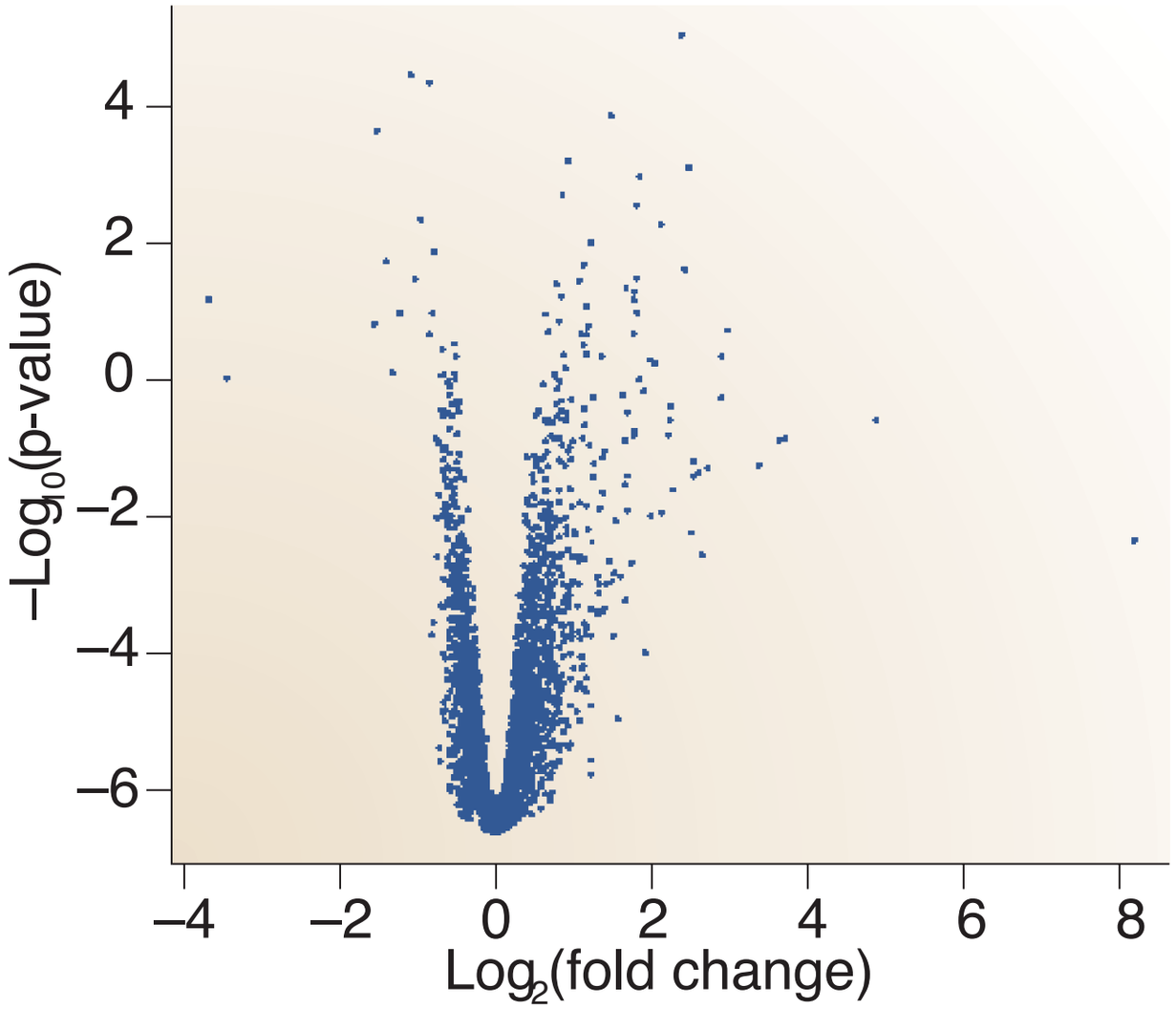
$\log_2$ (fold change).



Allison et al., Nature Reviews Genetics 7, 55-65 (January 2006)

Volcano Plots

Upper corners represent genes that show both statistical significance and large fold changes.

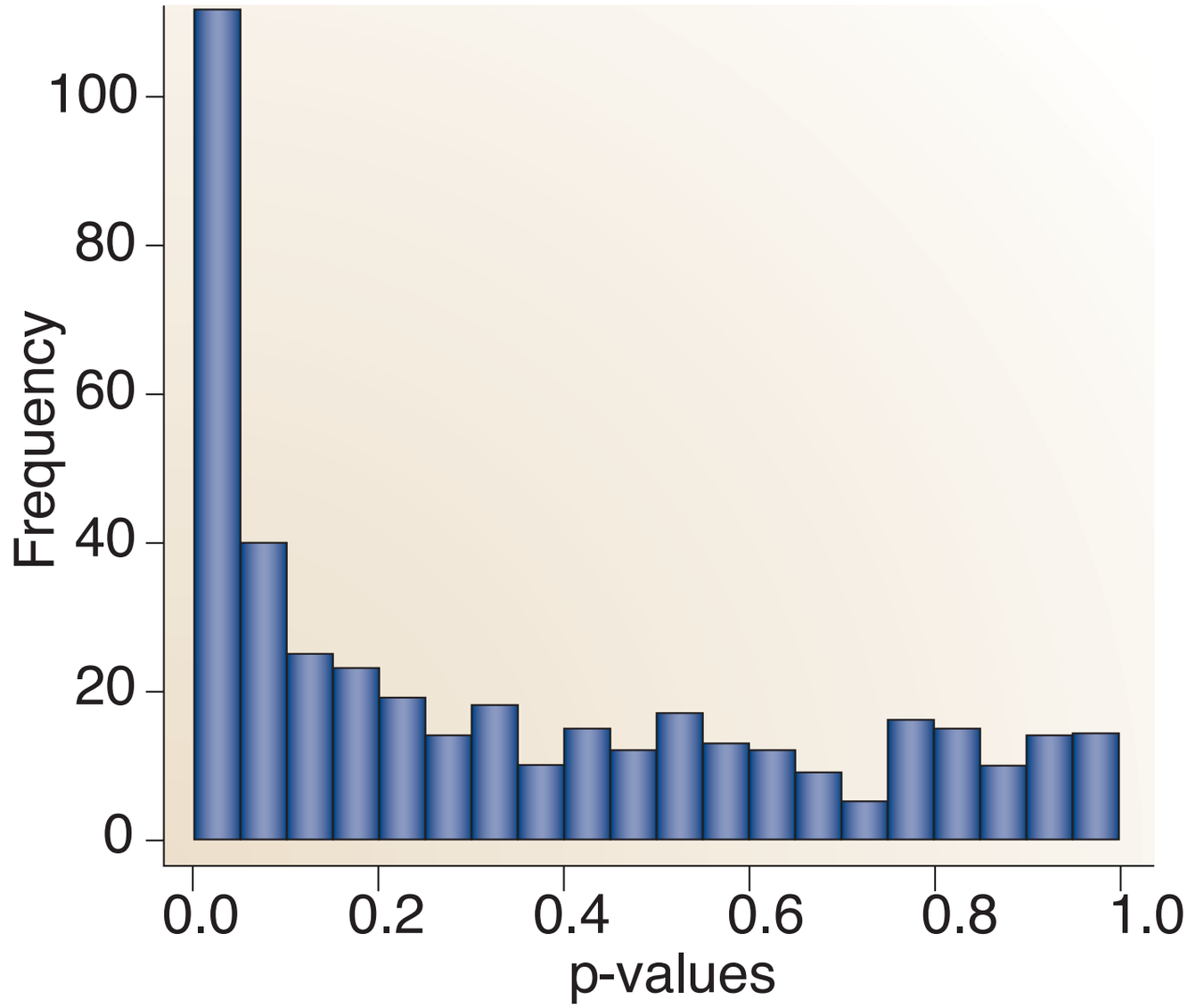


Allison et al., Nature Reviews Genetics 7, 55-65 (January 2006)

p-value Histograms

p-values for a chosen test of differential expression for each gene.

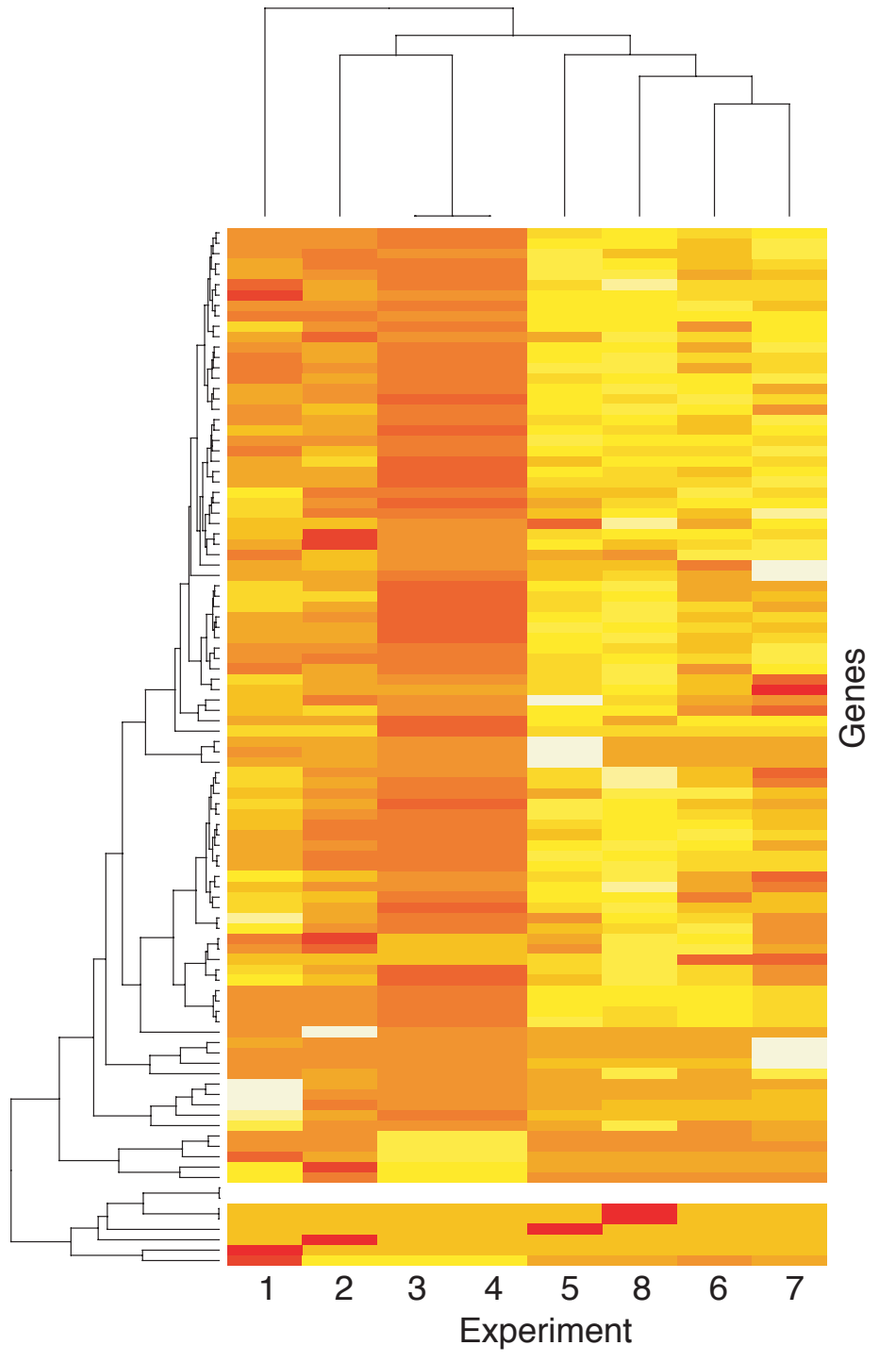
Can be used for false-discovery rates estimations.



Allison et al., Nature Reviews Genetics 7, 55-65 (January 2006)

Heatmaps & Dendrograms

- Each cell represents a value (scale important)
- Often generated following hierarchical clustering analyses.
- Dendrograms shown to the sides indicating *clustering* results.
- Clustering can be both vertical and horizontal.

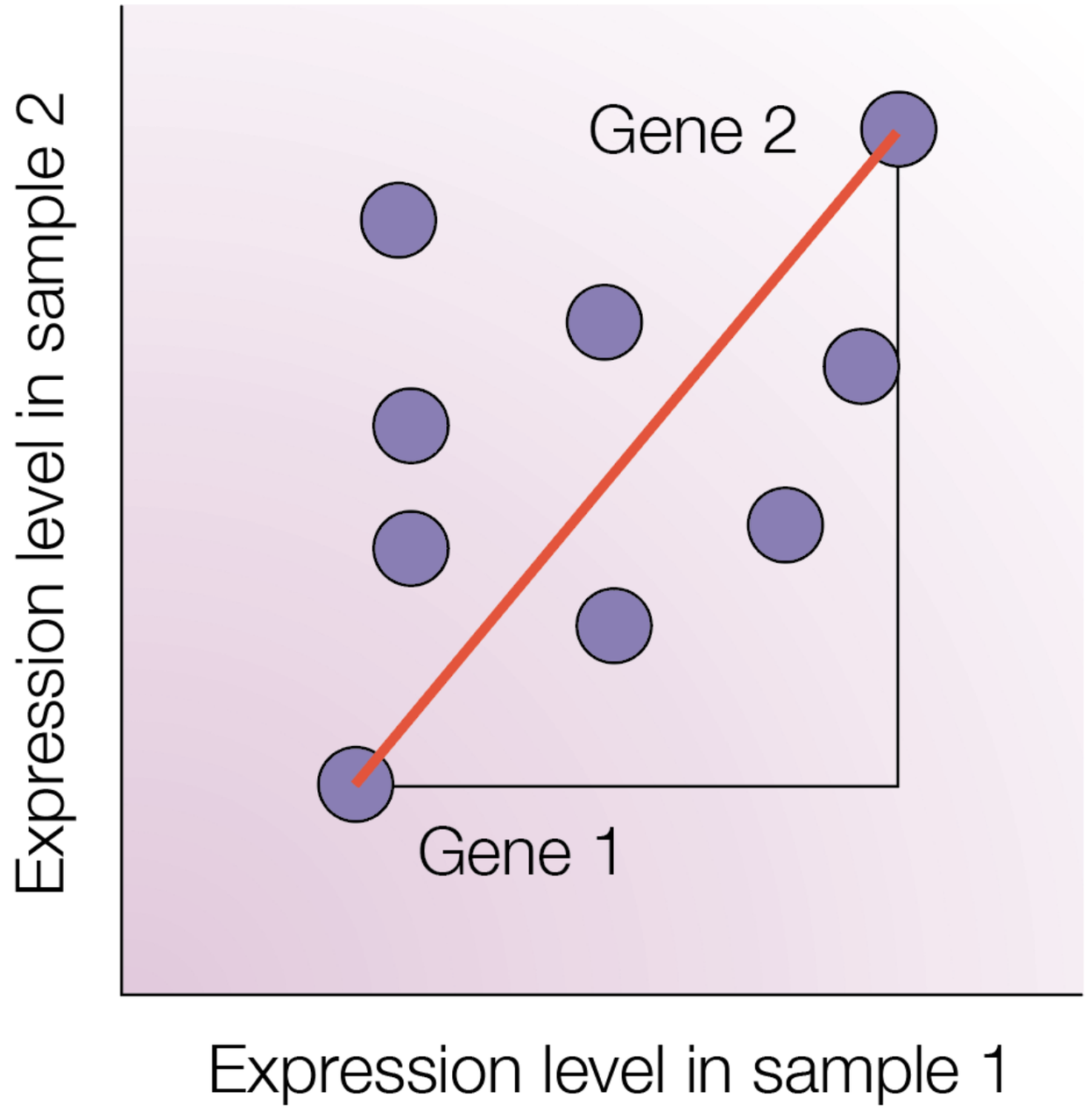


Allison et al., Nature Reviews Genetics 7, 55-65 (2006)

Dissimilarity Measures

Clustering techniques compare genes pairwise to compute *similarity*.

Euclidean distance (compare to usual vectors and using Pythagorean theorem)

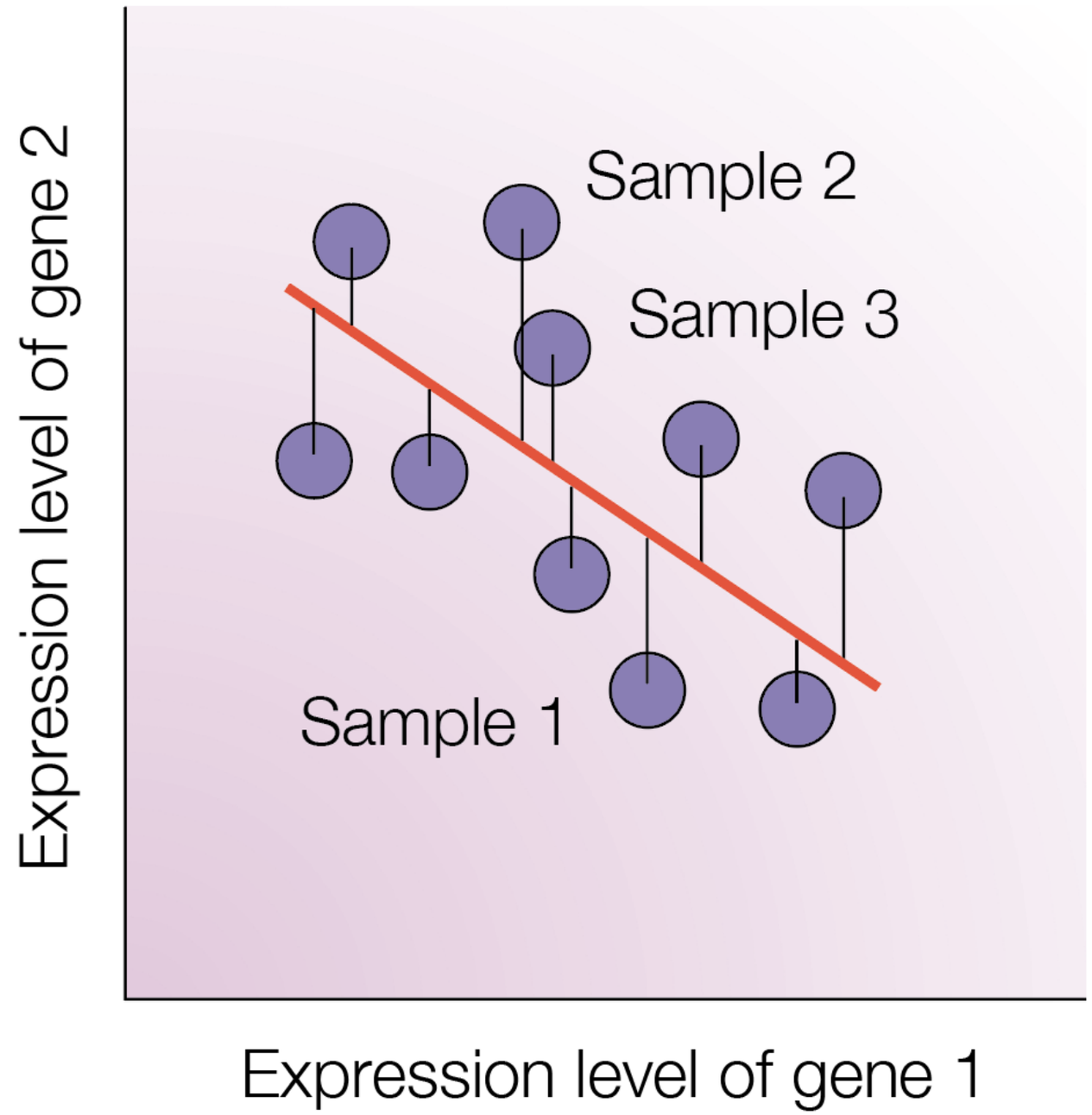


Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Dissimilarity Measures

Pearson correlation coefficient as a distance

Calculated from the distances of each point from the linear-regression line (known as residuals).



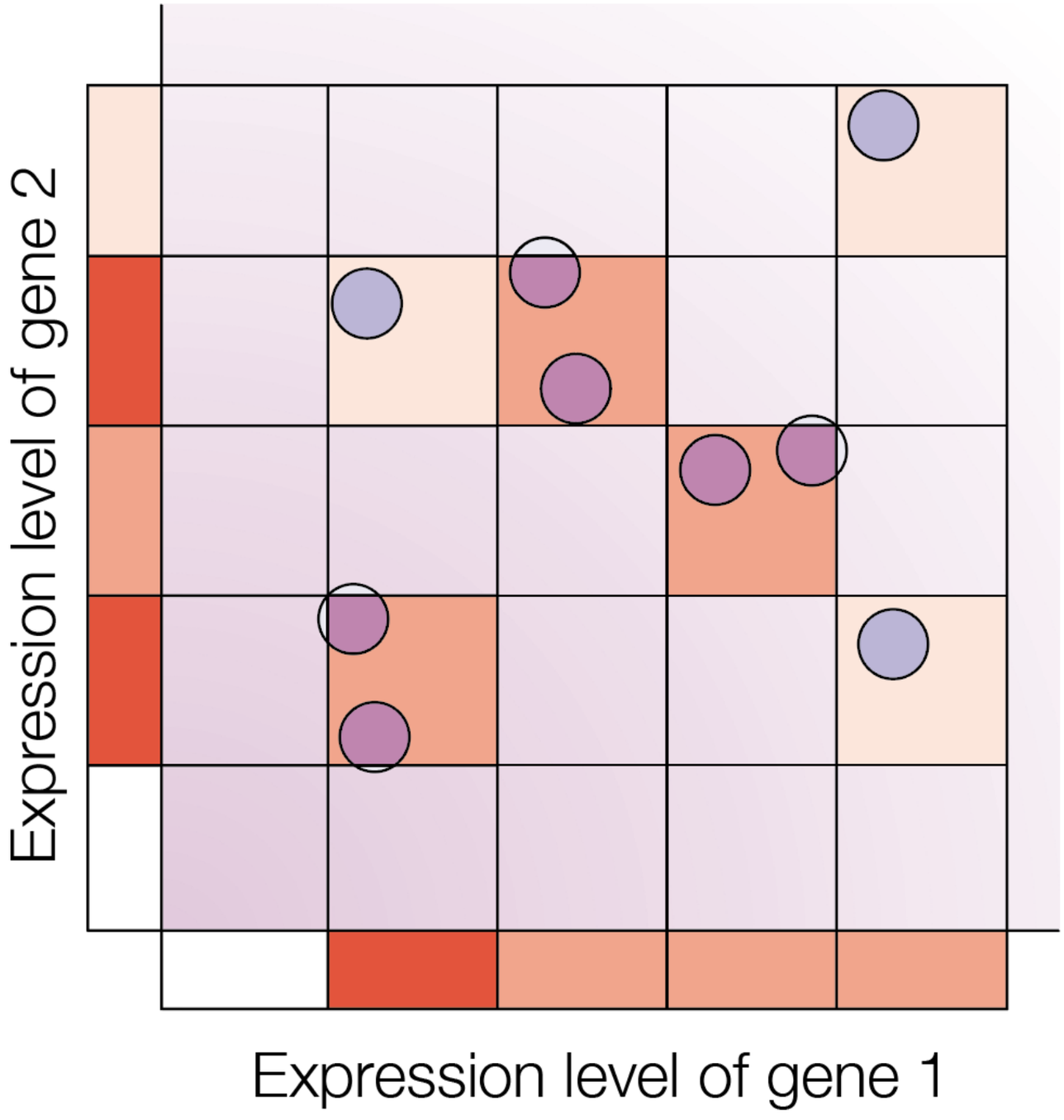
Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Dissimilarity Measures

Mutual information

Model-free measurement of the degree of information content in one gene known from another gene.

Highest when genes are randomly distributed separately, but show a non-random joint distribution.



Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Dissimilarity Measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \left\| e_{fc} - e_{gc} \right\|$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (\mathbf{e}_f - \mathbf{e}_g)' \Sigma^{-1} (\mathbf{e}_f - \mathbf{e}_g)$, where $\boldsymbol{\Sigma}$ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace $\boldsymbol{e_{gc}}$ with the rank of $\boldsymbol{e_{gc}}$ within the expression values of gene g across all conditions $\boldsymbol{c = 1 \ldots C}$
Absolute or squared correlation	$d_{fg} = 1 - \left\| r_{fg} \right\|$ or $d_{fg} = 1 - r_{fg}^2$

d_{fg}, distance between expression patterns for genes f and g. e_{gc}, expression level of gene g under condition c.

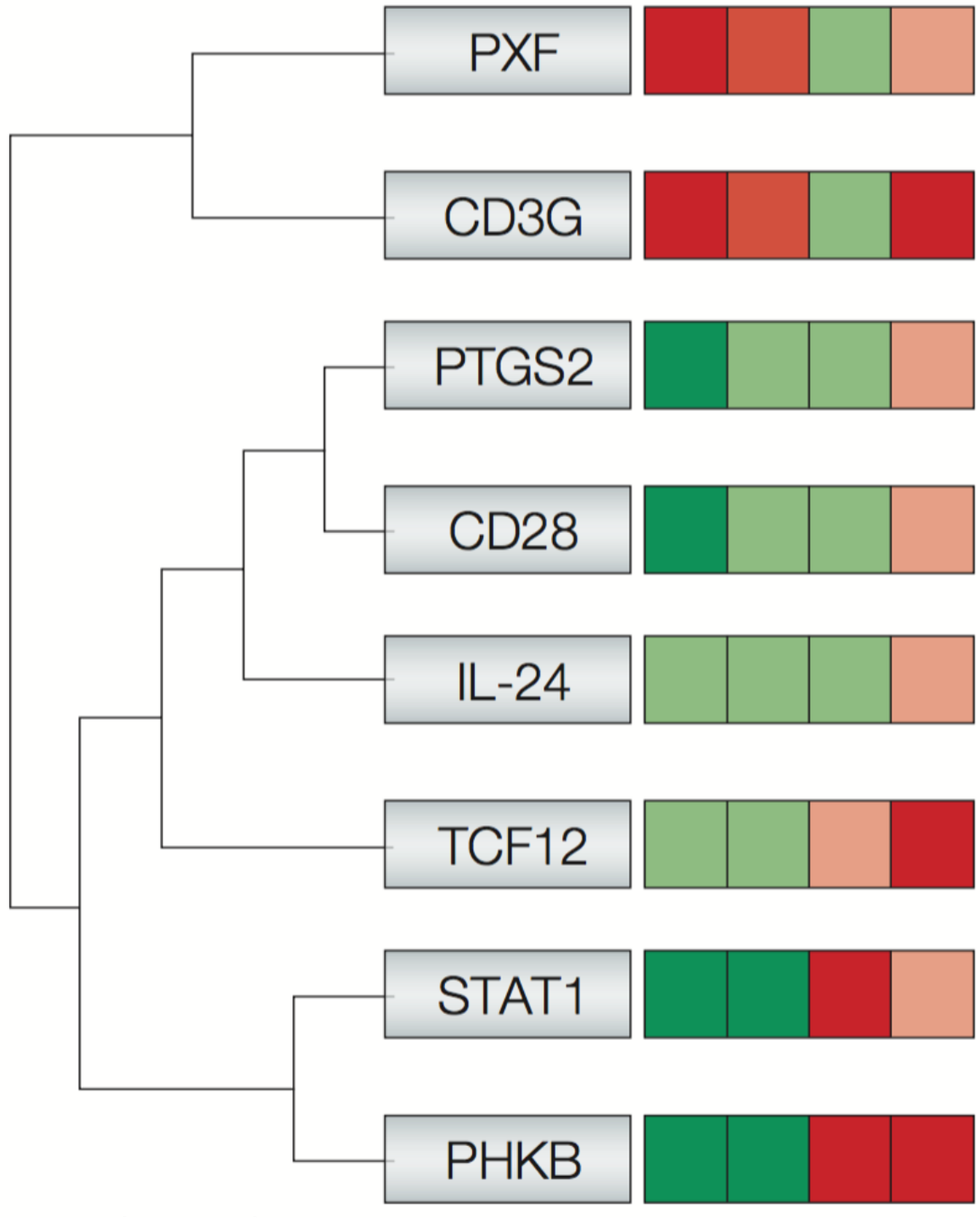
D’haeseleer, Nature Biotechnology 23, 1499 - 1501 (2005)

Clustering Methods

Hierarchical Clustering

Sorting so similar genes appear near each other.

The length of the branch is inversely proportional to the degree of similarity.



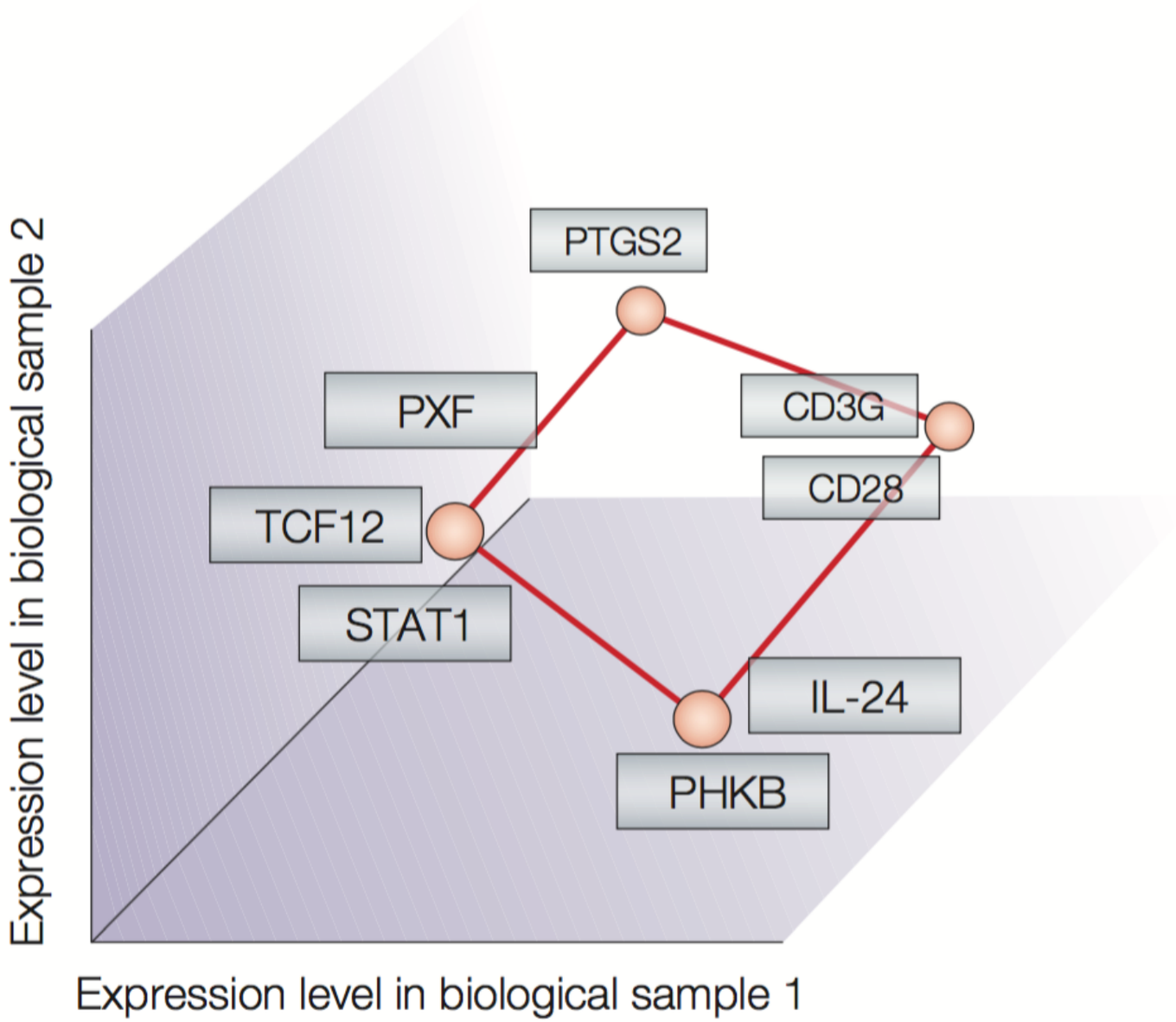
Shades of red indicate increased relative expression; shades of green indicate decreased relative expression.

Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Clustering Methods

Self Organizing Maps (SOMs)

Self-organizing maps find variable-sized clusters of genes that are similar to each other, given the input number of clusters to find.



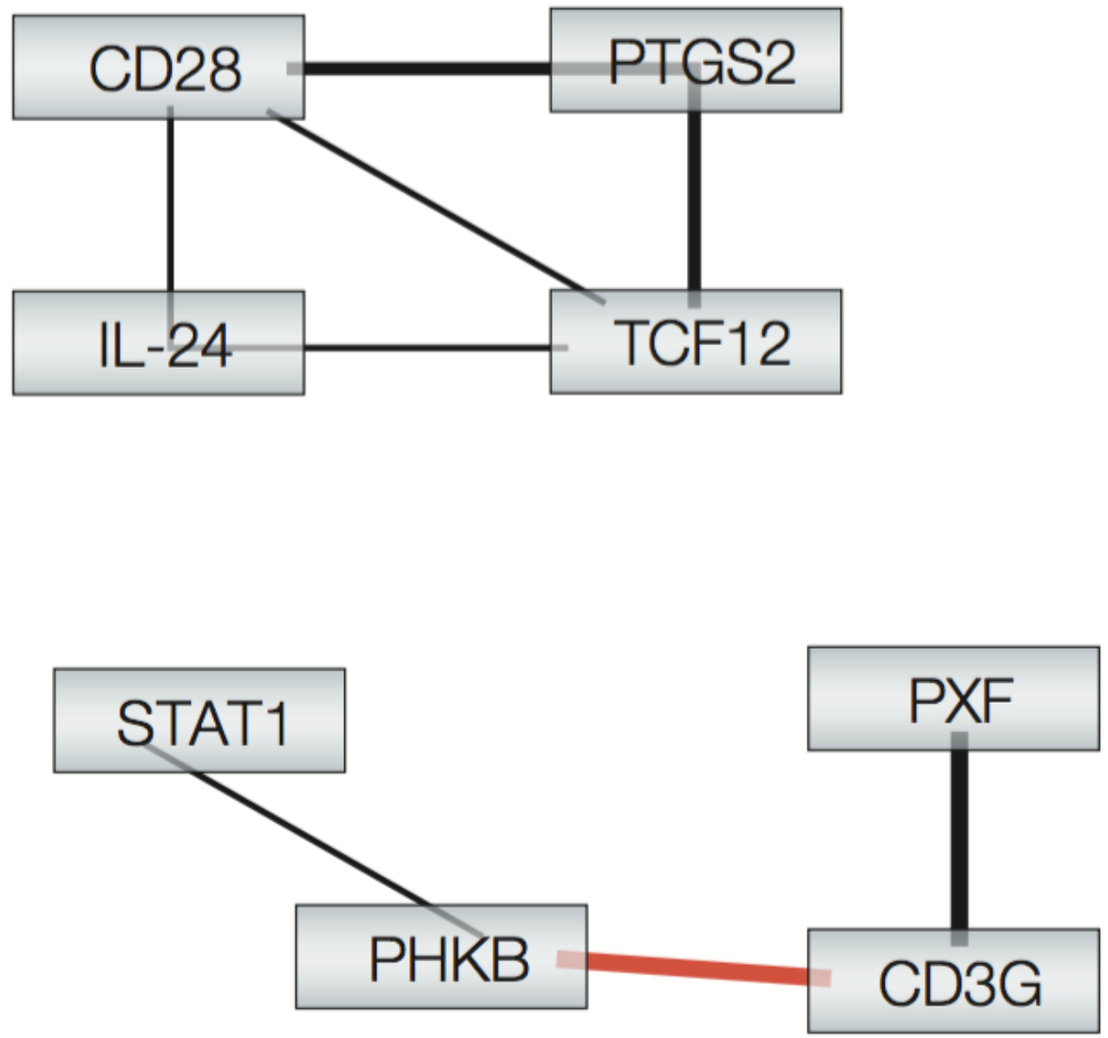
Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Clustering Methods

Relevance Networks

Find and display pairs of genes with strong positive and negative correlations, then construct networks from these gene pairs.

Strength of correlation is proportional to the thickness of the edges, and red indicates a negative correlation.



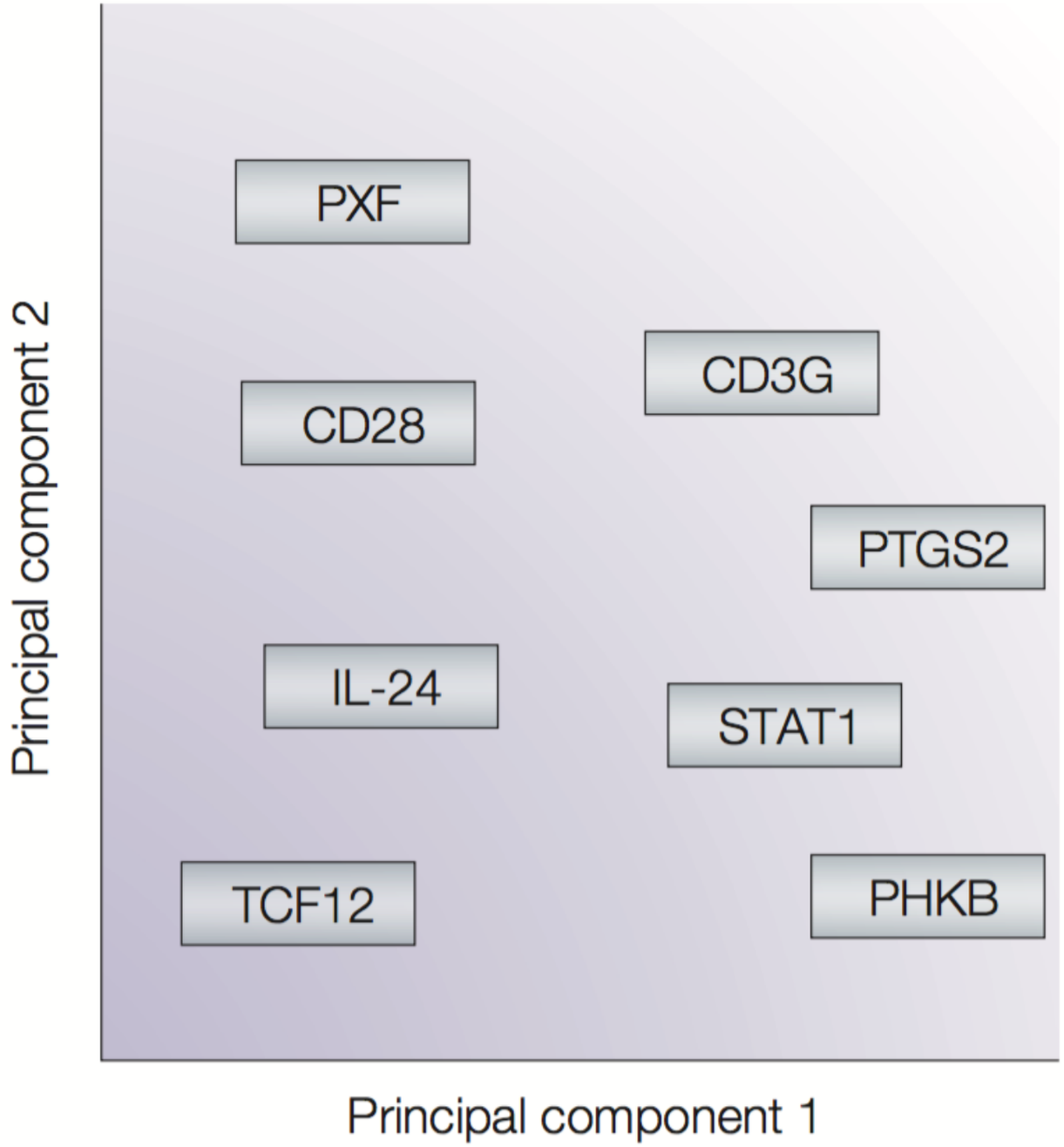
Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Clustering Methods

Principal-components analysis

Shows clustering or scatter of genes (or samples) when viewed along two or three principal components

Principal component can be thought of as a ‘meta-biological sample’, which combines all the biological samples so as to capture the most variation in gene expression.

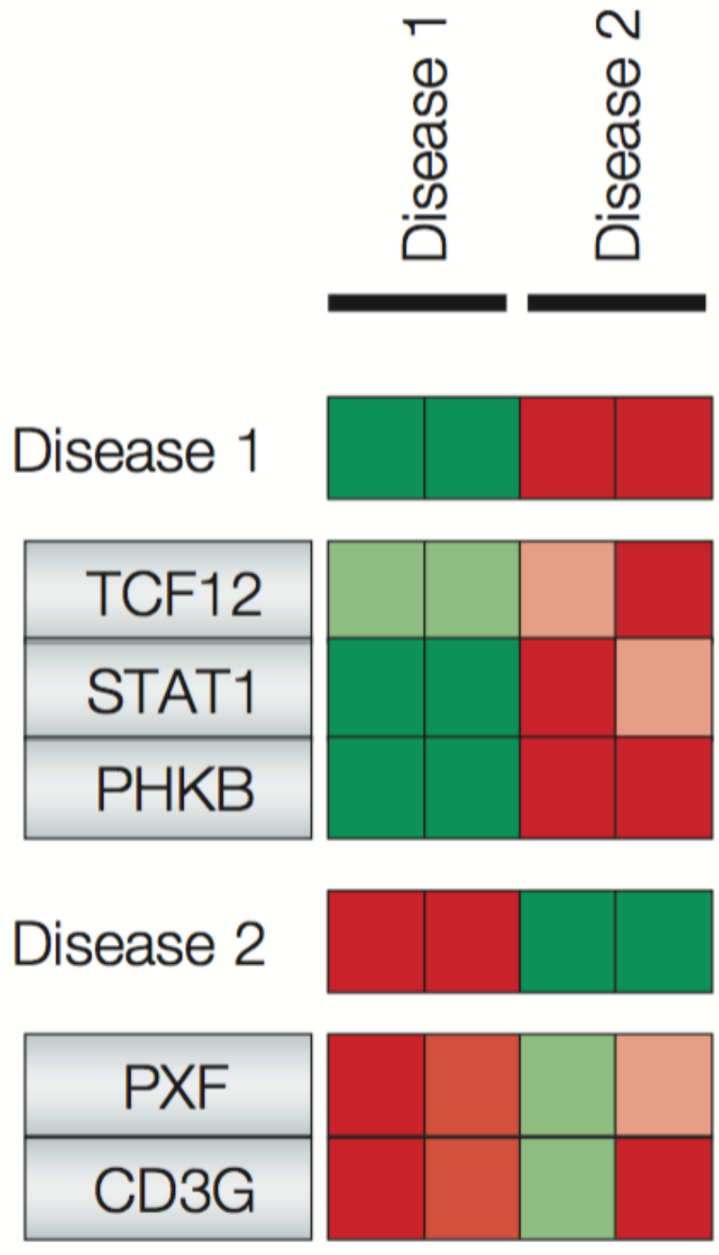


Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Clustering Methods

The nearest-neighbour supervised method

1. First construct hypothetical genes that best fit the desired patterns (e.g.a gene with high expression in disease 1 and low expression in disease 2, or vice versa).
2. Find individual genes that are most similar to the hypothetical genes.

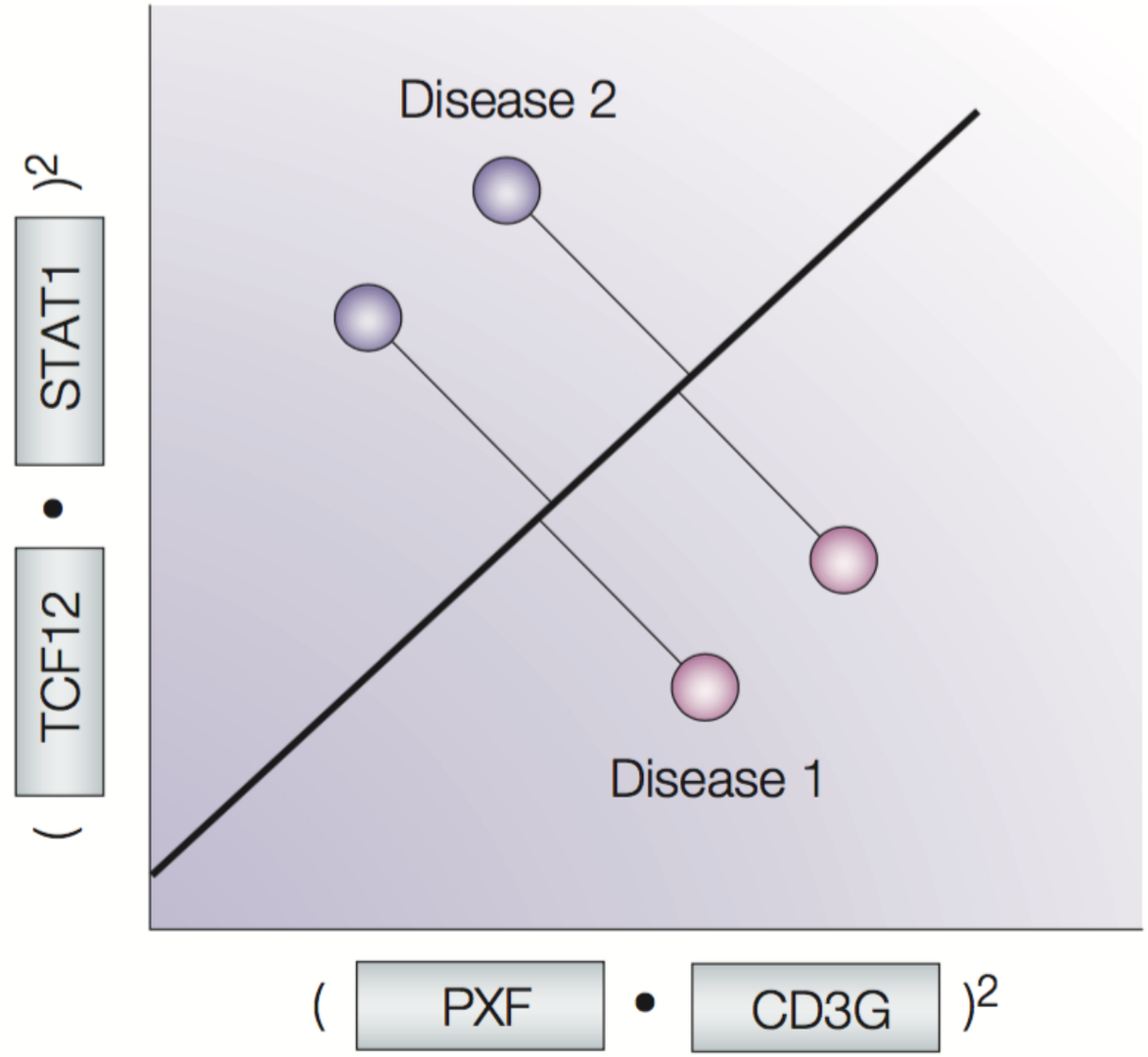


Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Clustering Methods

Support Vector Machines

Instead of restricting to individual genes, find the line (or plane) that best separates groups of biological samples.

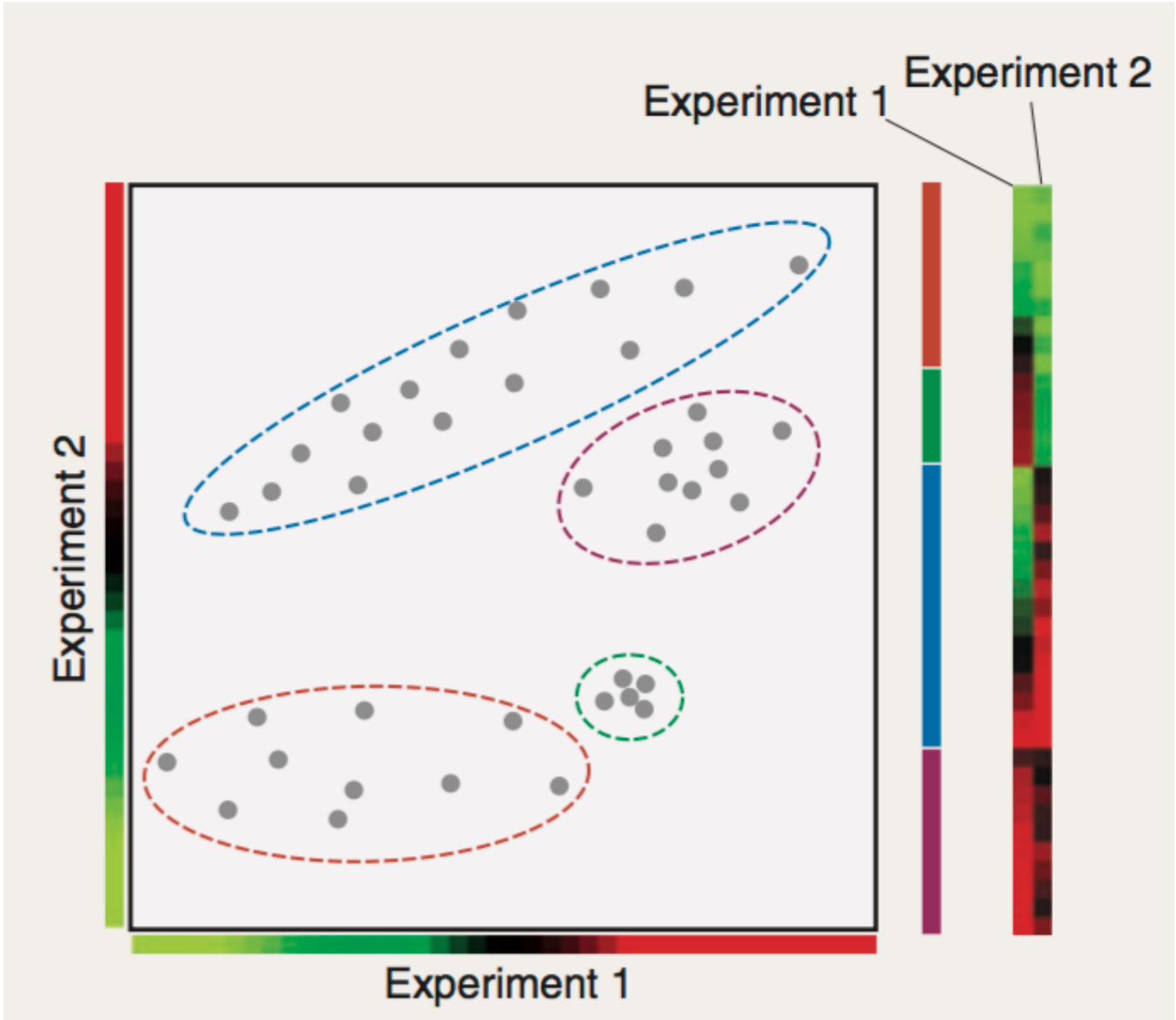


Atul Butte, Nature Reviews Drug Discovery 1, 951-960 (2002)

Clustering Examples

40 genes measured
under two different conditions.

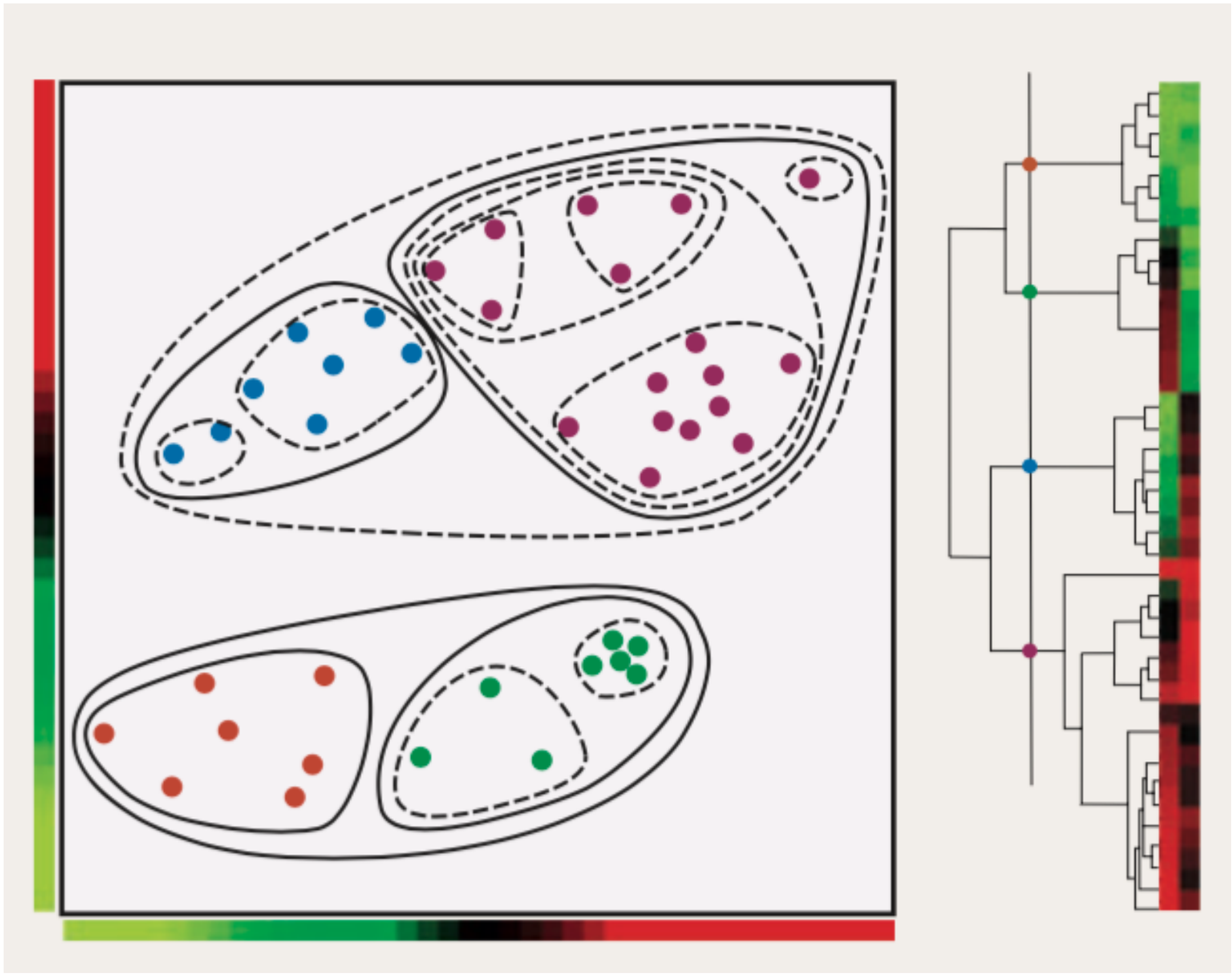
4 clusters of different sizes, shapes and numbers of genes..
Euclidean distance, which corresponds to the straight-line distance between points in this graph, was used for clustering.
Right: the standard red-green representation of the data and corresponding cluster identities



D'haeseleer, Nature Biotechnology 23, 1499 - 1501 (2005)

Clustering Examples

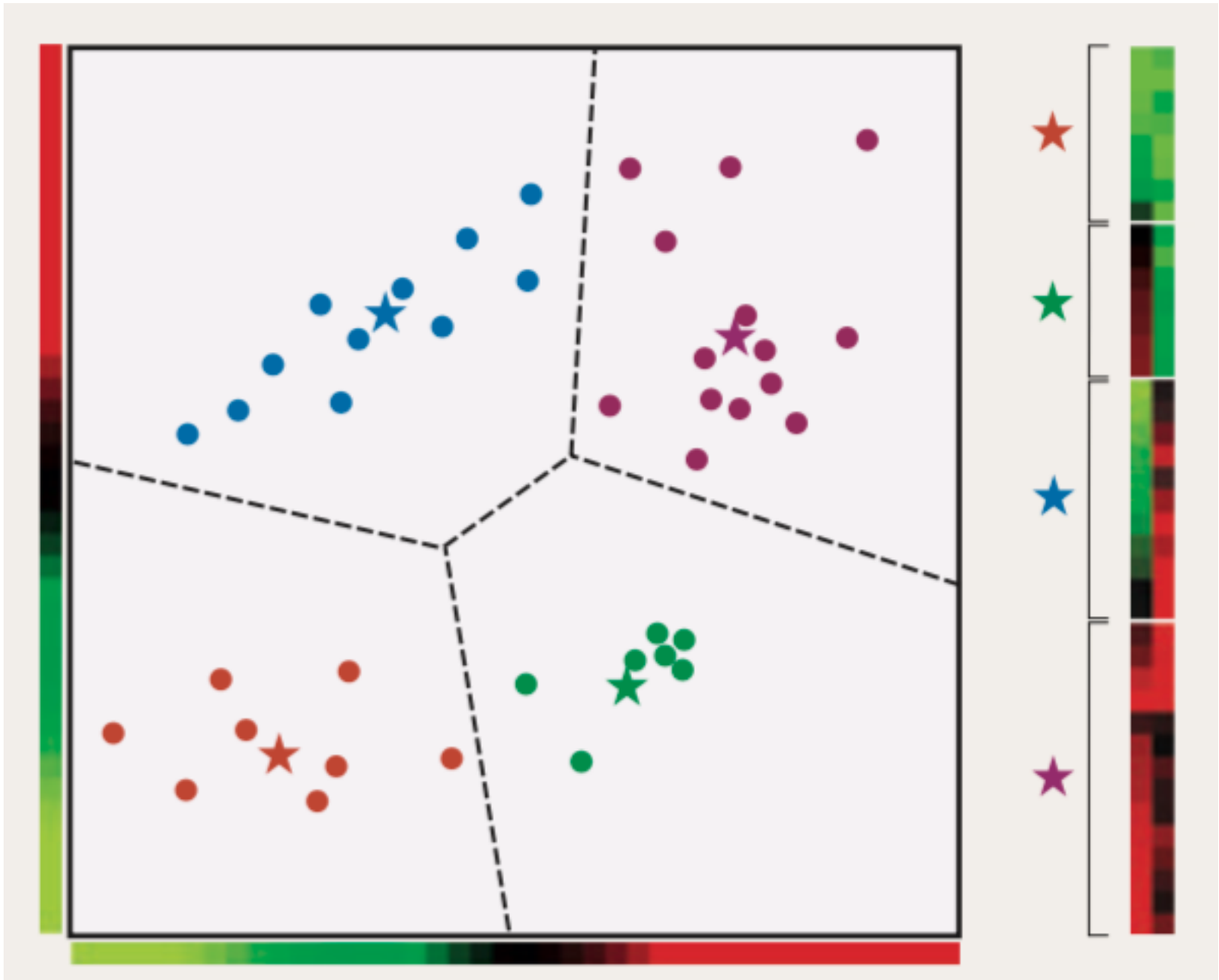
Hierarchical clustering finds an entire hierarchy of clusters. The tree was cut at the level indicated to yield four clusters.



D'haeseleer, Nature Biotechnology 23, 1499 - 1501 (2005)

Clustering Examples

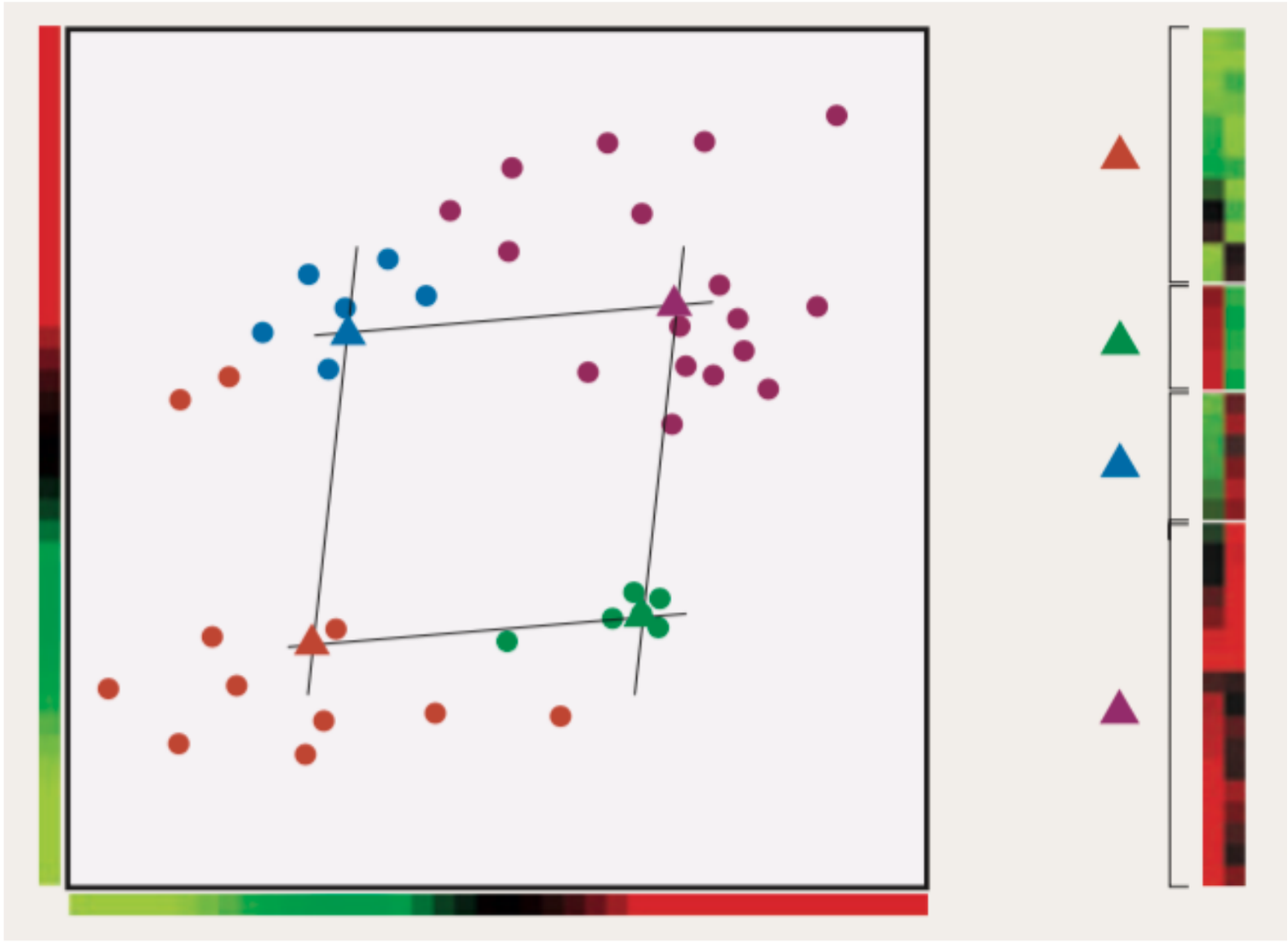
k-means (with k = 4) partitions the space into four subspaces, depending on which of the four cluster centroids (stars) is closest.



D'haeseleer, Nature Biotechnology 23, 1499 - 1501 (2005)

Clustering Examples

SOM finds clusters, which are organized into a grid structure
(in this case a 2 × 2 grid).

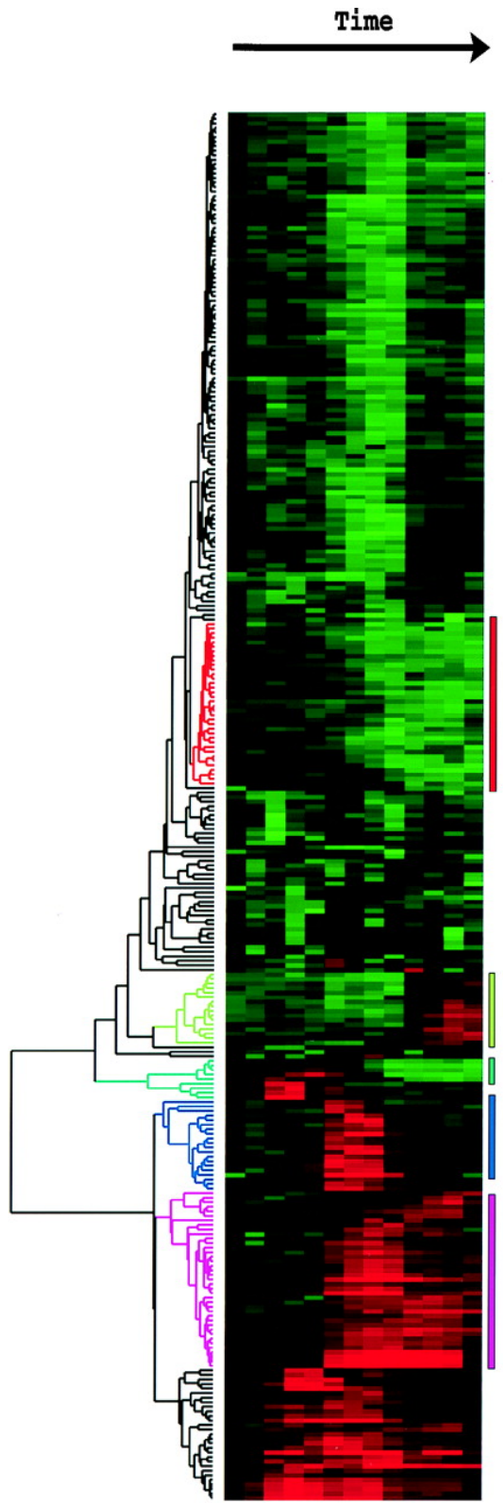


D'haeseleer, Nature Biotechnology 23, 1499 - 1501 (2005)

Patterns in Data

- Foreskin fibroblasts grown in culture and were deprived of serum for 48 hr.
- Serum was added back and samples taken at time 0, 15 min, 30 min, 1 hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr - final datapoint was from unsynchronized sample.
- cDNA microarray with 8,600 distinct human genes.
- Measurements are relative to time 0.
- Genes were selected if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. the corresponding region of the dendrogram.

Eisen et al. PNAS 95:14863-14868 (1998)
Iyer et al., Science 283(5398), pp. 83-87 (1999)



Clusters with genes in:

(A) cholesterol biosynthesis

(B) the cell cycle

(C) the immediate–early response

(D) signaling and angiogenesis,

(E) wound healing and tissue remodeling

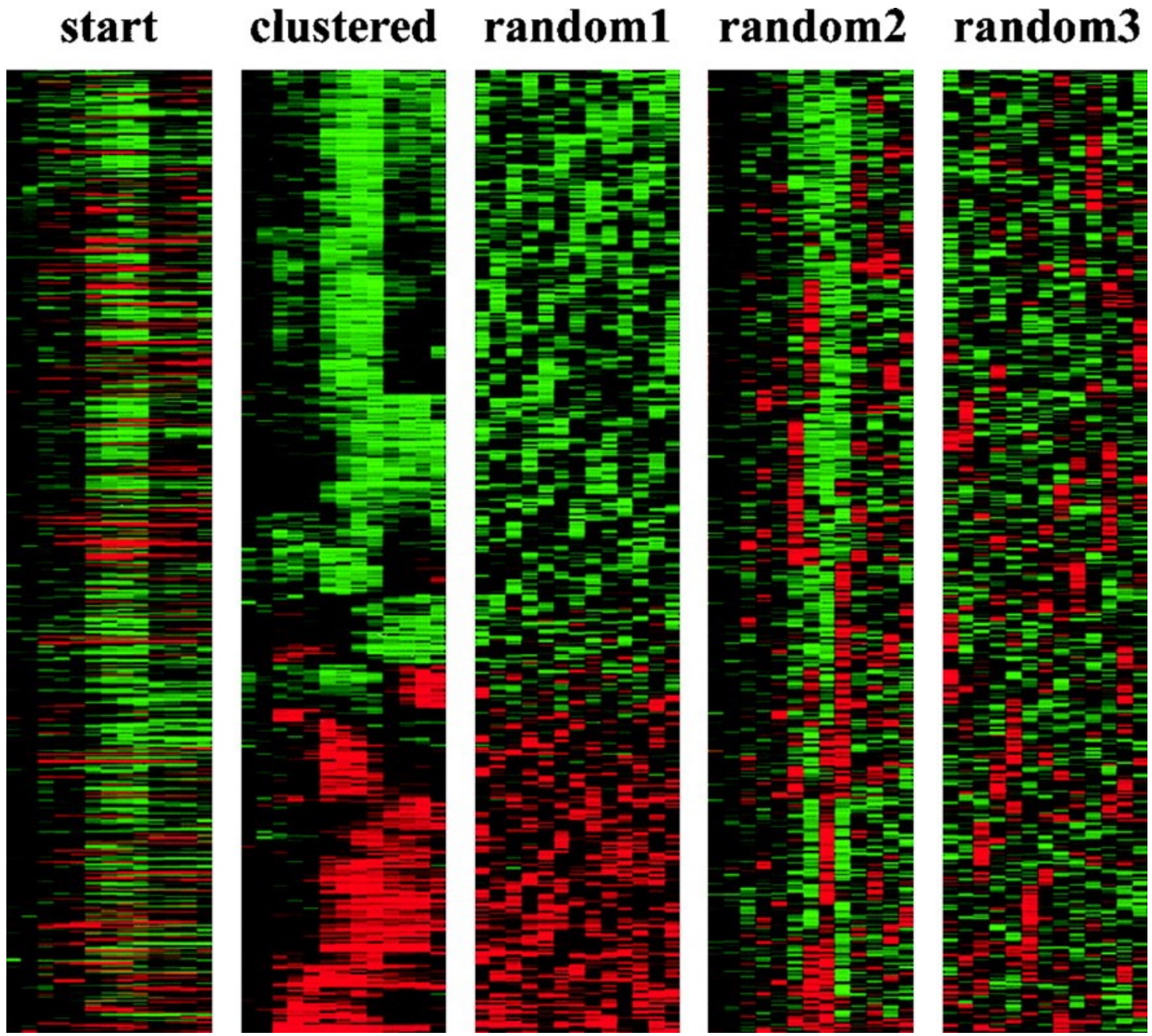
Patterns in Data

Permutations:

random 1, within rows

random 2, within columns

random 3, both directions



Eisen et al. PNAS 95:14863-14868 (1998)