3000788 Intro to Comp Molec Biol

Lecture 10: Transcriptomics

Fall 2025

Sira Sriswasdi, PhD

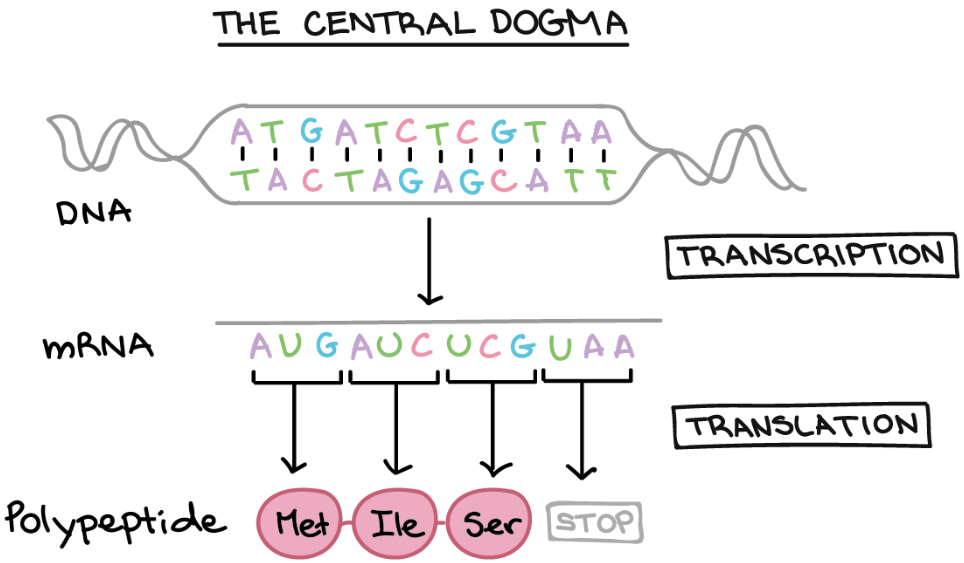
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today's agenda

- Technology for measuring gene expression
  - Oligonucleotide microarray
  - Nanostring
  - RNA sequencing

- Log-normal model for gene expression data

Why is transcriptomics so popular?

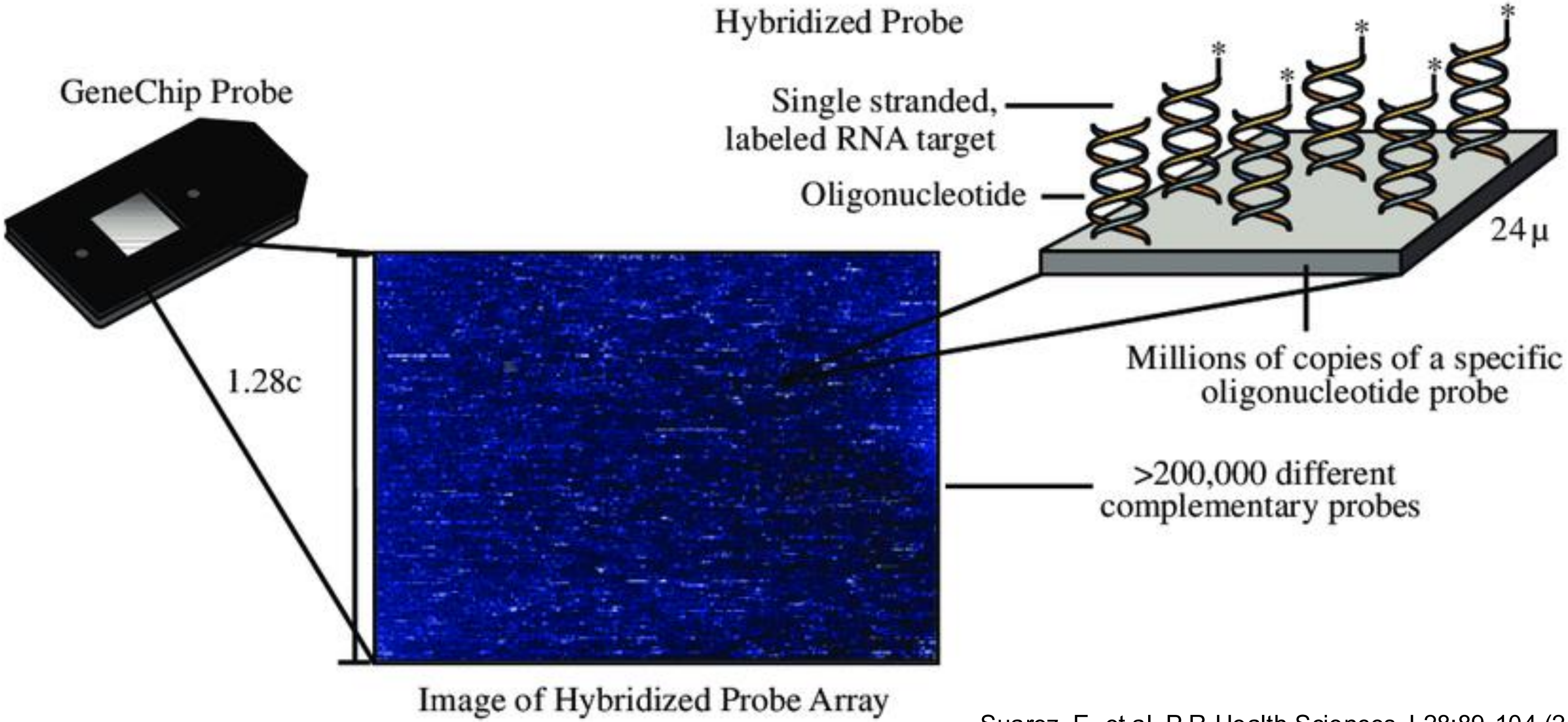


- Easy to quantify
- Explain broad cellular functions and phenotypes
- Proteins are difficult to study
  - DNA sequencing not applicable

https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/translation/a/intro-to-gene-expression-central-dogma

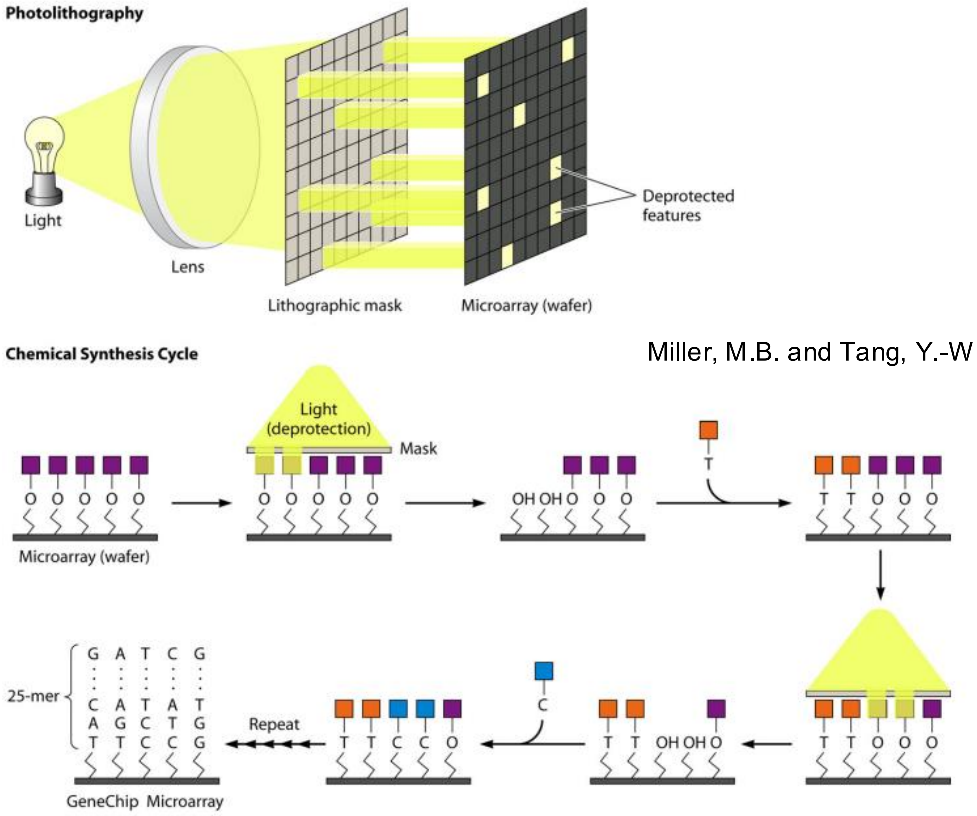
Oligonucleotide microarray

Microarray technology overview


Hybridized Probe
GeneChip Probe
Single stranded, labeled RNA target
Oligonucleotide
24 µ
1.28c
Millions of copies of a specific oligonucleotide probe
>200,000 different complementary probes
Image of Hybridized Probe Array

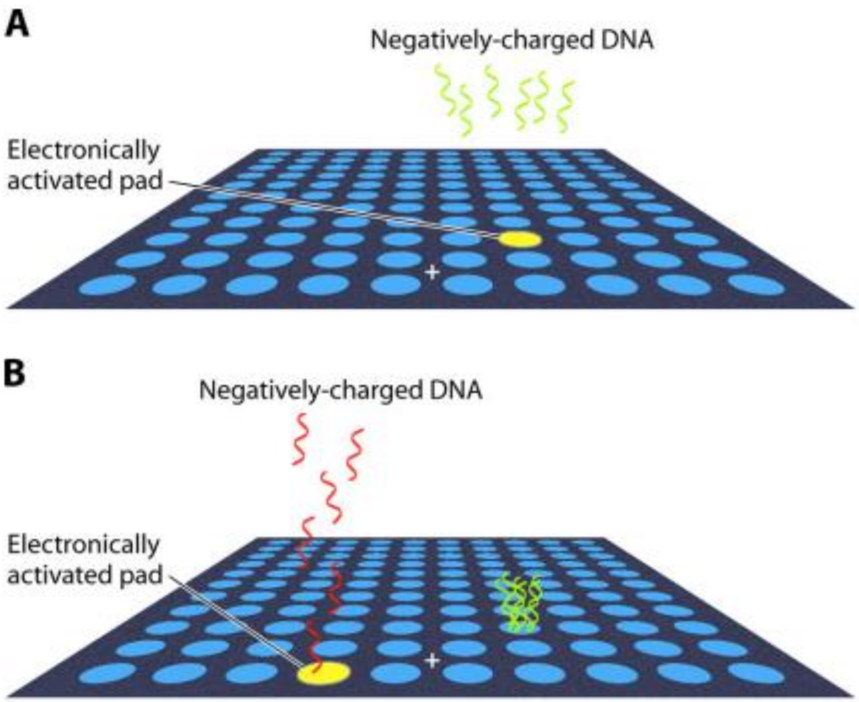
Suarez, E. et al. P R Health Sciences J 28:89-104 (2009)

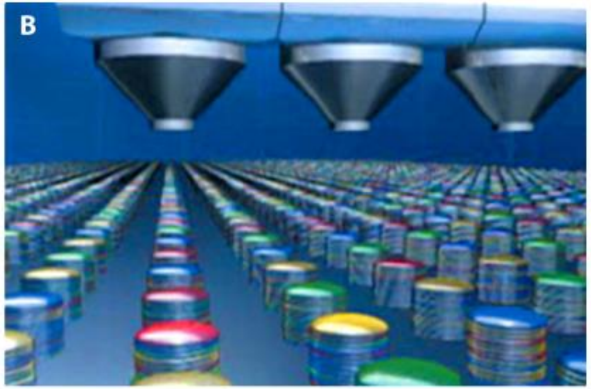
Microarray fabrication

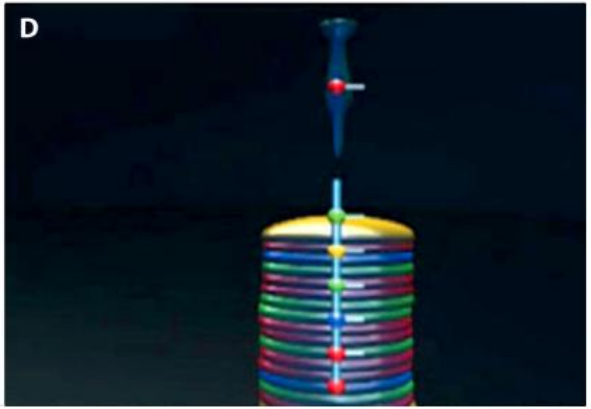


Miller, M.B. and Tang, Y.-W. Clin Microbiol Rev 611-633 (2009)

Microarray fabrication



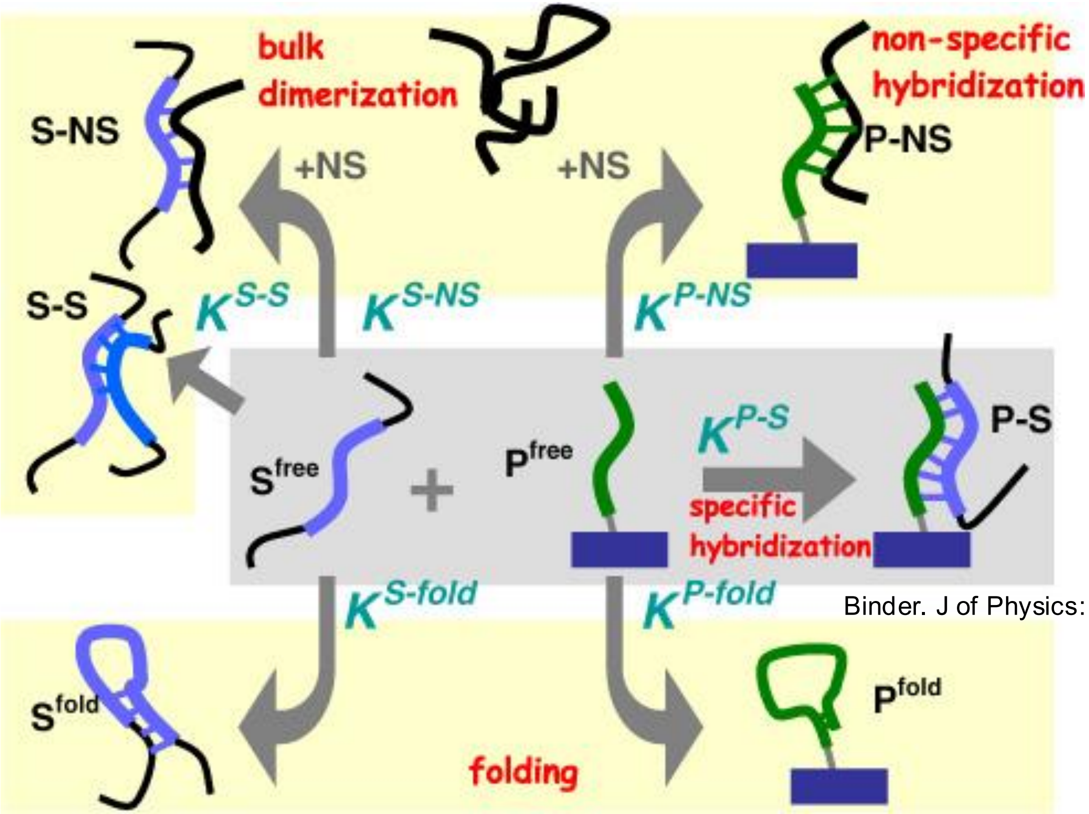




Miller, M.B. and Tang, Y.-W. Clin Microbiol Rev 611-633 (2009)

Probe design for microarray

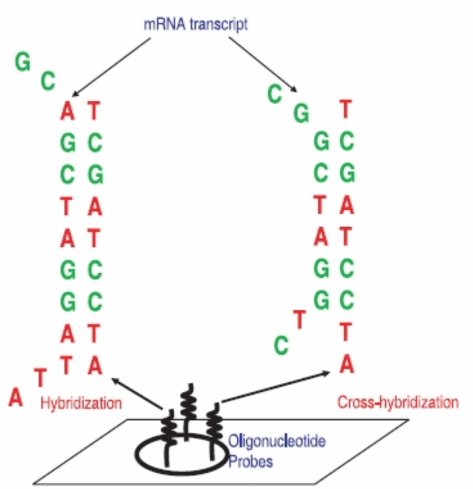
Unwanted probe interactions



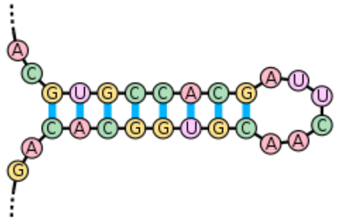
Binder. J of Physics: Condensed Matter 18. (2006)

Probe design principles

- **Sensitivity**
  - Complementary to each target genomic region
  - Multiple probes for each target
- **Specificity**
  - Reduced cross-hybridization
    - Check with BLAST
  - Negative control and mismatch probes
- **Technical issues**
  - Prevent secondary structure formation
  - Control hybridization energy
  - Redundant probes across array
    - Position-specific bias

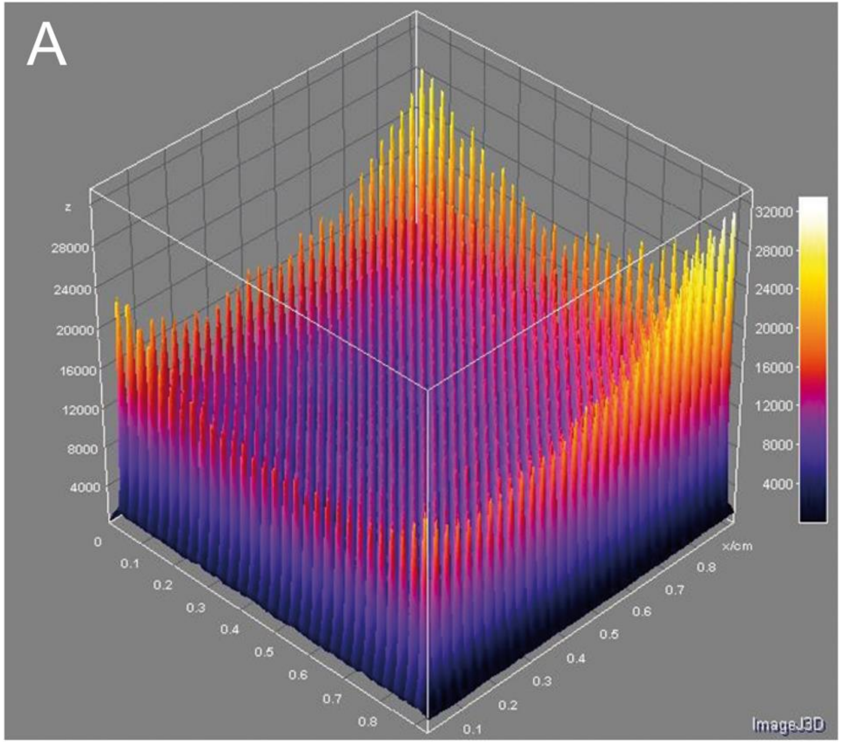


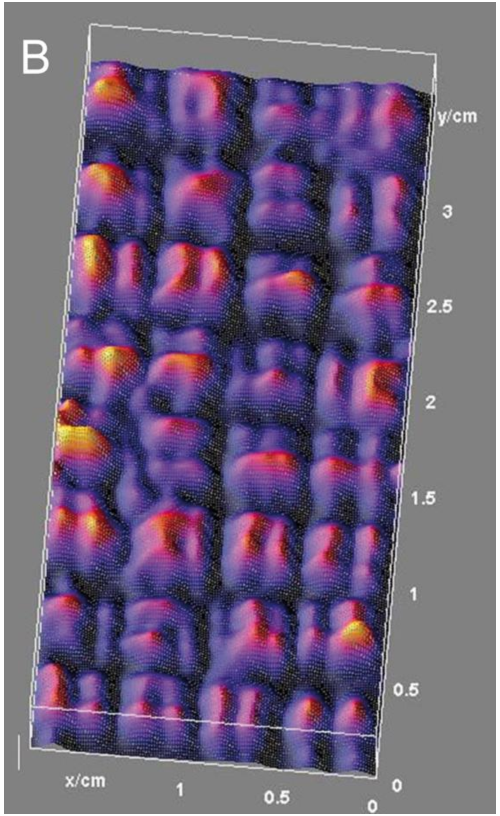
Malutan, R. and Vilda, P.G. Thermodynamics of Microarray Hybridization (2011)



https://en.wikipedia.org/wiki/Complementarity_%28molecular_biology%29

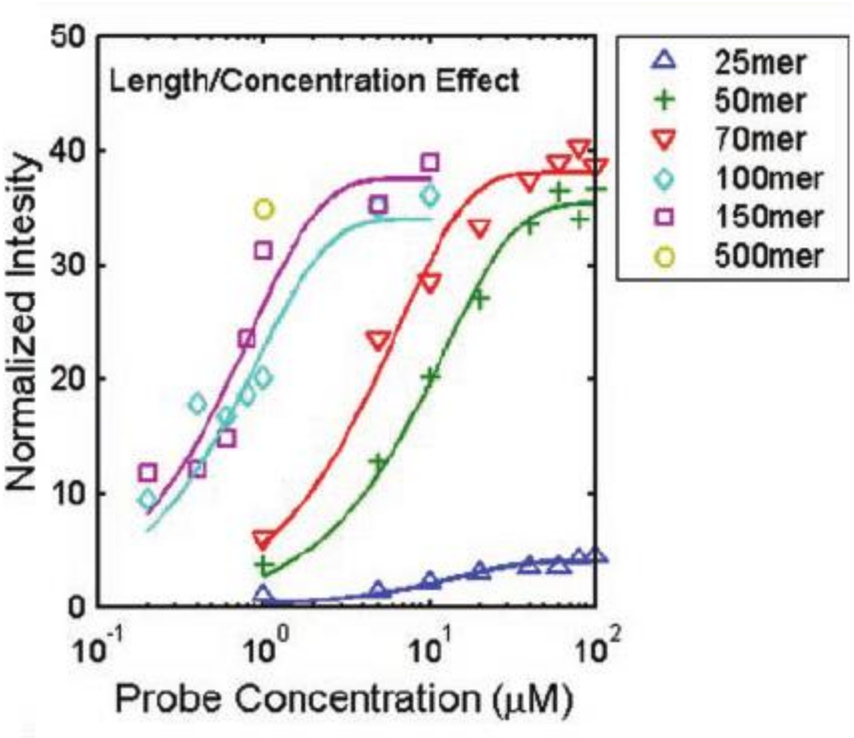
Position-specific bias in microarray

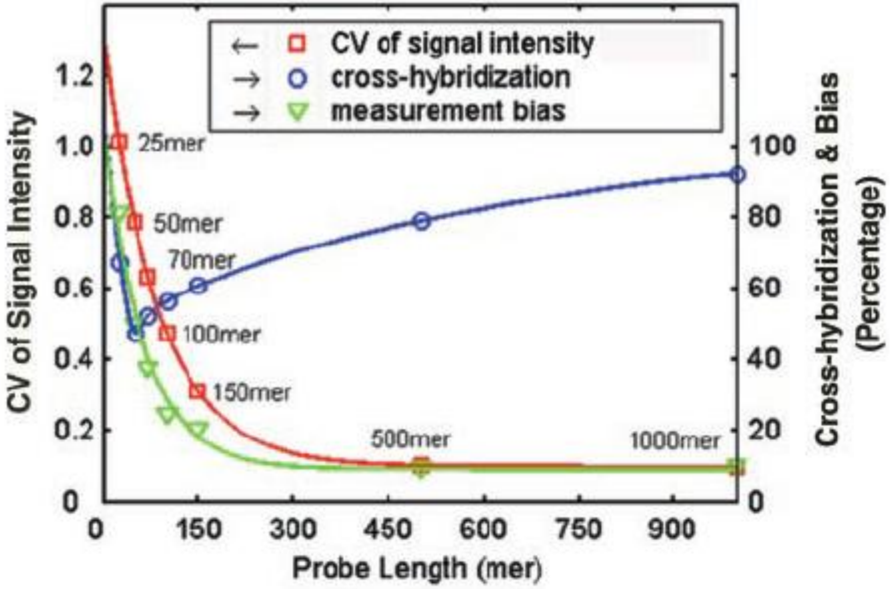




Steger, D. et al. PLoS ONE 6:e23727 (2011)

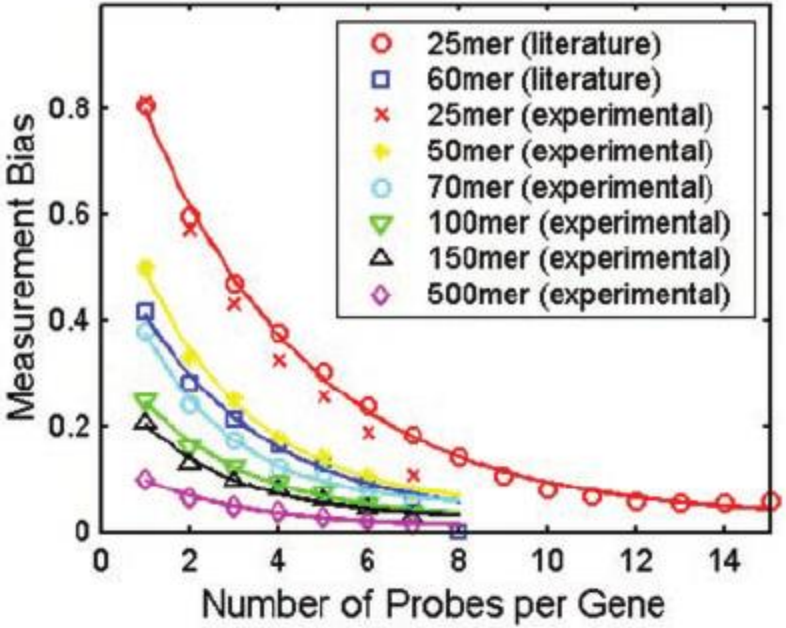
Impact of probe length



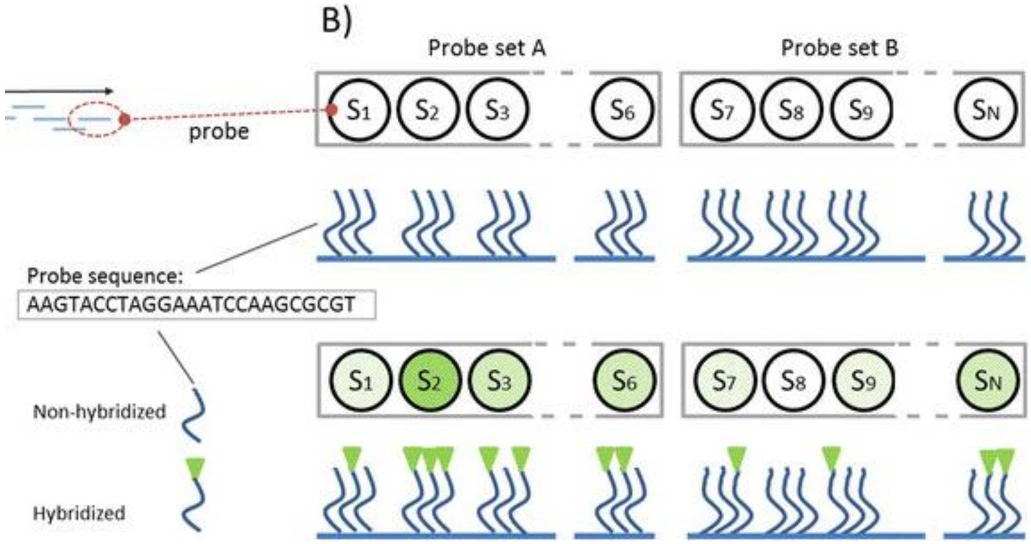


Chou, C.-C. NAR 32:e99 (2004)

Probe set = multiple probes per gene

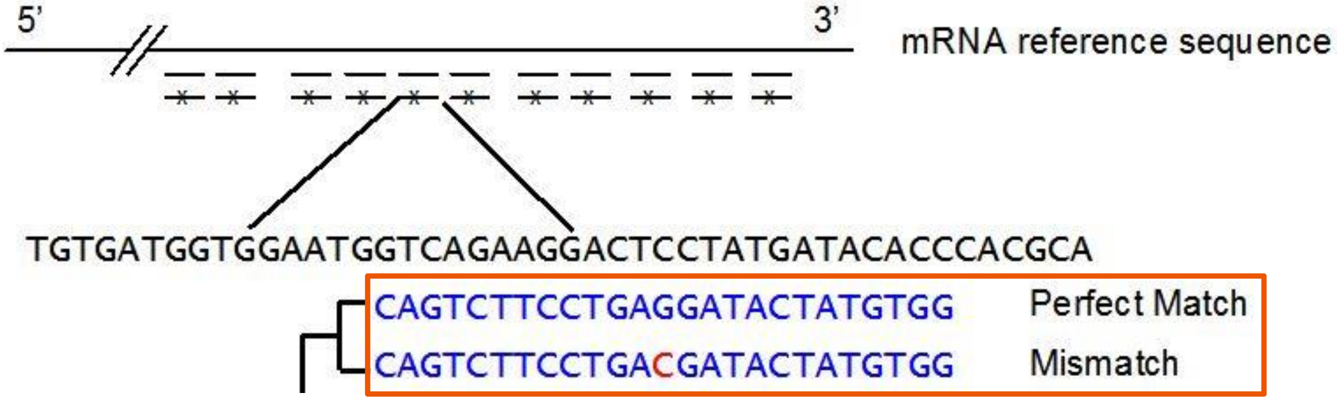


Chou, C.-C. NAR 32:e99 (2004)



Jaksik, R. et al. Biology Direct 10:46 (2015)

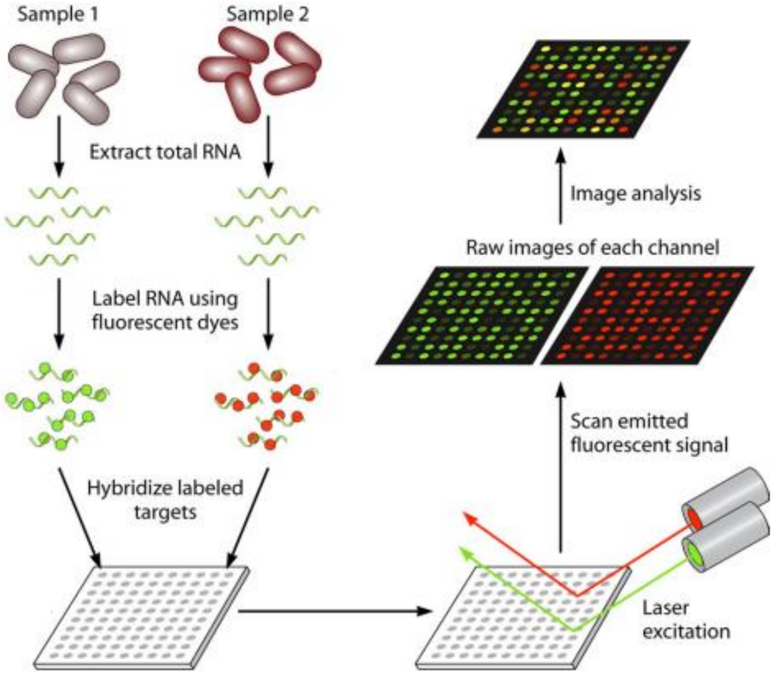
Perfect match (PM) and mismatch (MM)


5'
3'
mRNA reference sequence
TGTGATGGTGGAATGGTCAGAAGGACTCCTATGATACACCCACGCA
CAGTCTTCCTGAGGATACTATGTGG
Perfect Match
CAGTCTTCCTGACGATACTATGTGG
Mismatch

Malutan, R. and Vilda, P.G. Thermodynamics of Microarray Hybridization (2011)

- Compare signals between PM and MM probes
  - Expect more binding with PM probes
  - Equal signals = potentially non-specific match

Multi-channel microarray

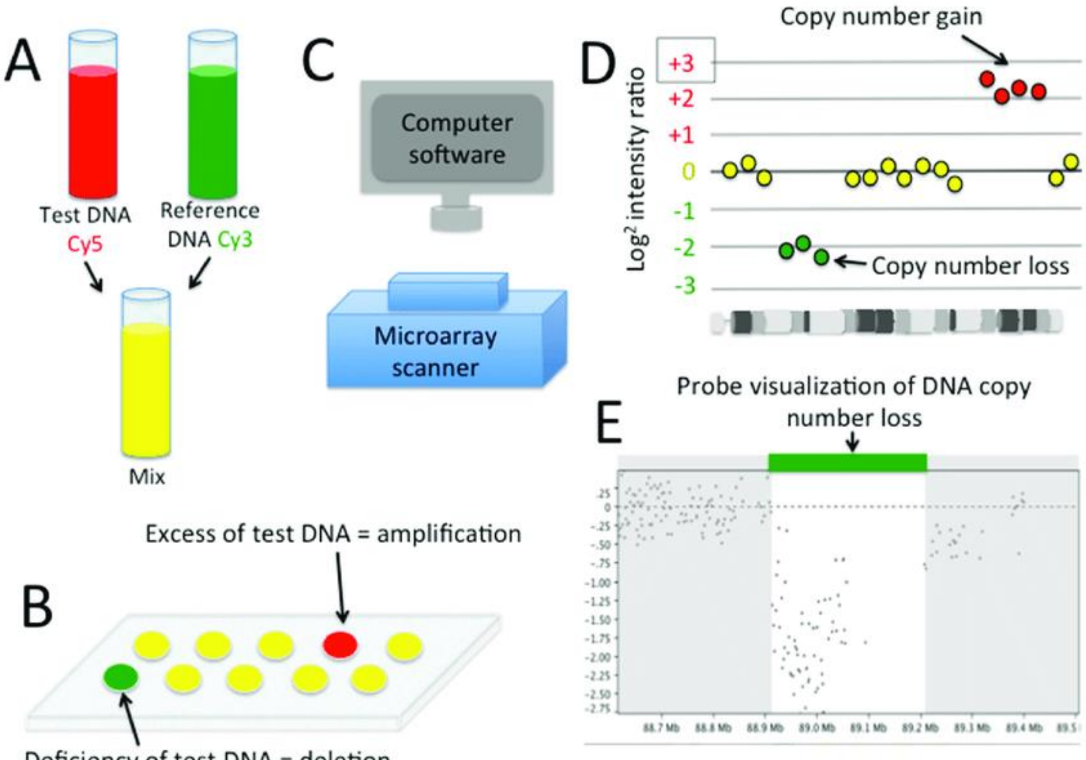

Sample 1
Sample 2
Extract total RNA
Label RNA using fluorescent dyes
Hybridize labeled targets
Laser excitation
Scan emitted fluorescent signal
Raw images of each channel
Image analysis

Miller, M.B. and Tang, Y.-W. Clin Microbiol Rev 611-633 (2009)

- Two samples are labeled with different dyes
- Mix and hybridize to microarray
- Relative fluorescence signal (ratio) directly indicates fold difference in gene expression
- **Minimize technical variance**

Other applications of microarrays

Comparative genome hybridization (CGH)



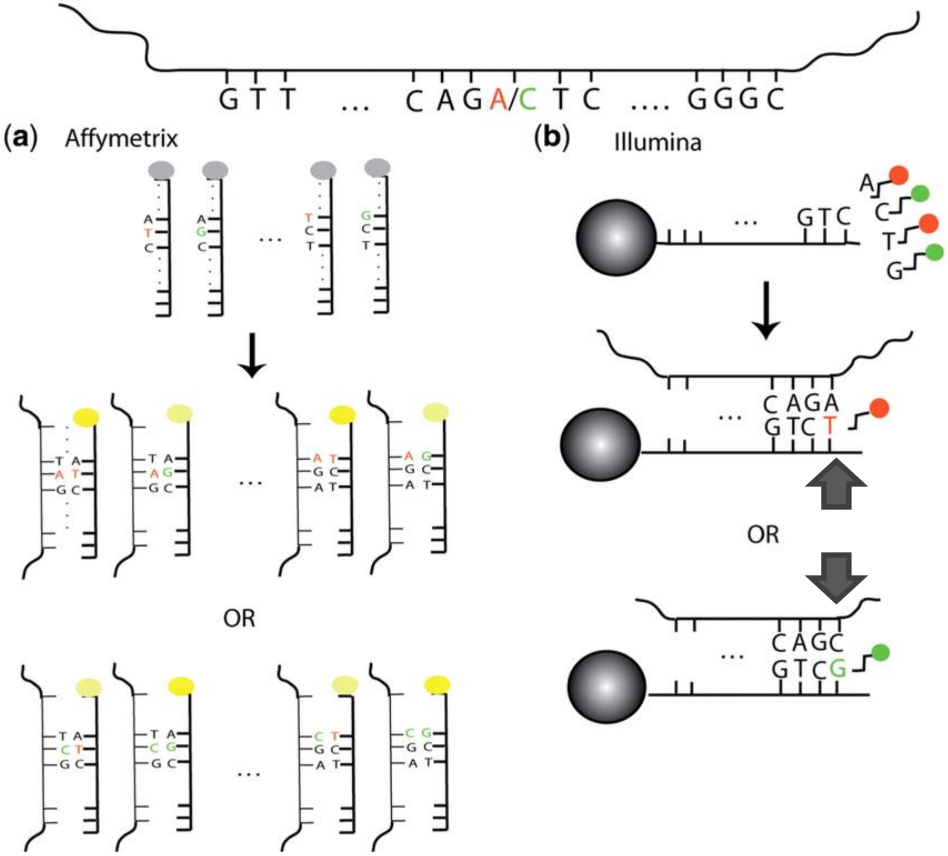
Hains, D.S. Pathogens 5:14 (2016)

- Design probes across genomic regions
- Compare to reference
- Loss of signal = deletion
- Gain of signal = DNA duplication

SNP genotyping array

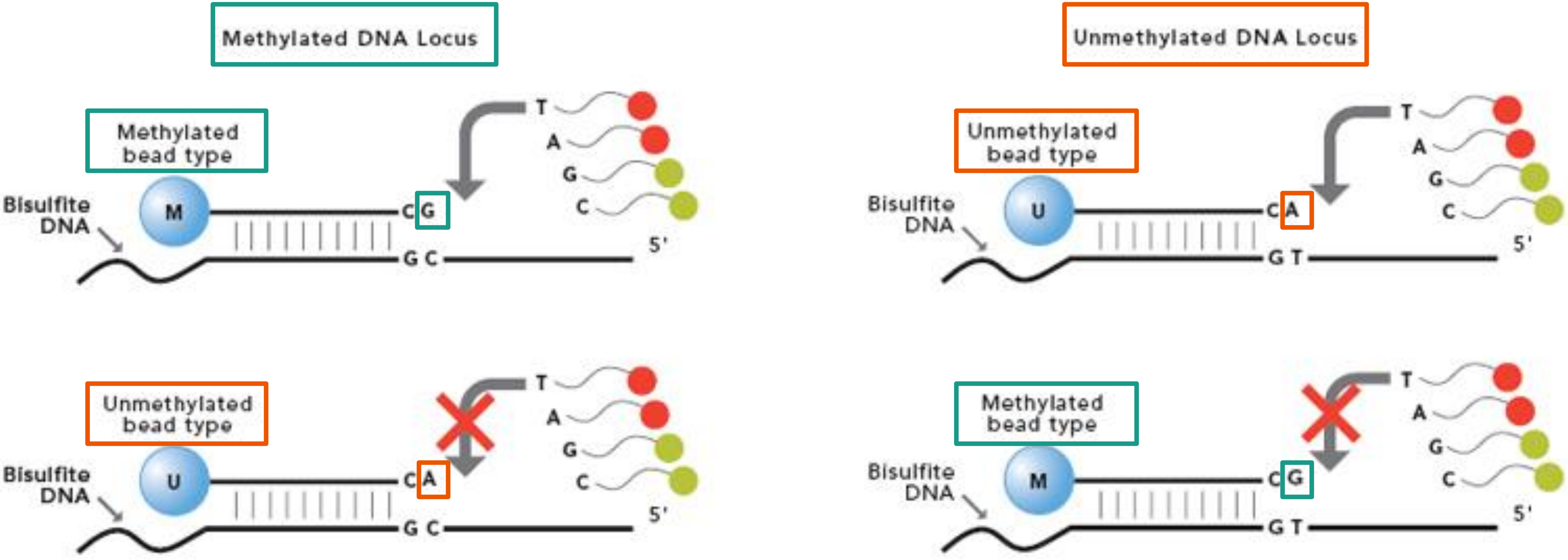
- Design probes for alternative SNPs at each position
  - Relative hybridization

- Single-nucleotide sequencing
  - Probe acts as primer
  - Match to the position right before the SNP
  - Sequence the SNP location



LaFramboise, T. NAR 37:4181-93 (2009)

Methylation array with bisulfite conversion



https://www.illumina.com/science/technology/microarray/infinium-methylation-assay.html

Microarray versus sequencing assays

- Microarray and DNA sequencing can be interchangeable
  - Genome tiling array
  - Fusion gene
  - ChIP-chip

- Microarray can be designed once for each task and **reused cheaply**

- **But microarray lack the ability to detect novel molecules**

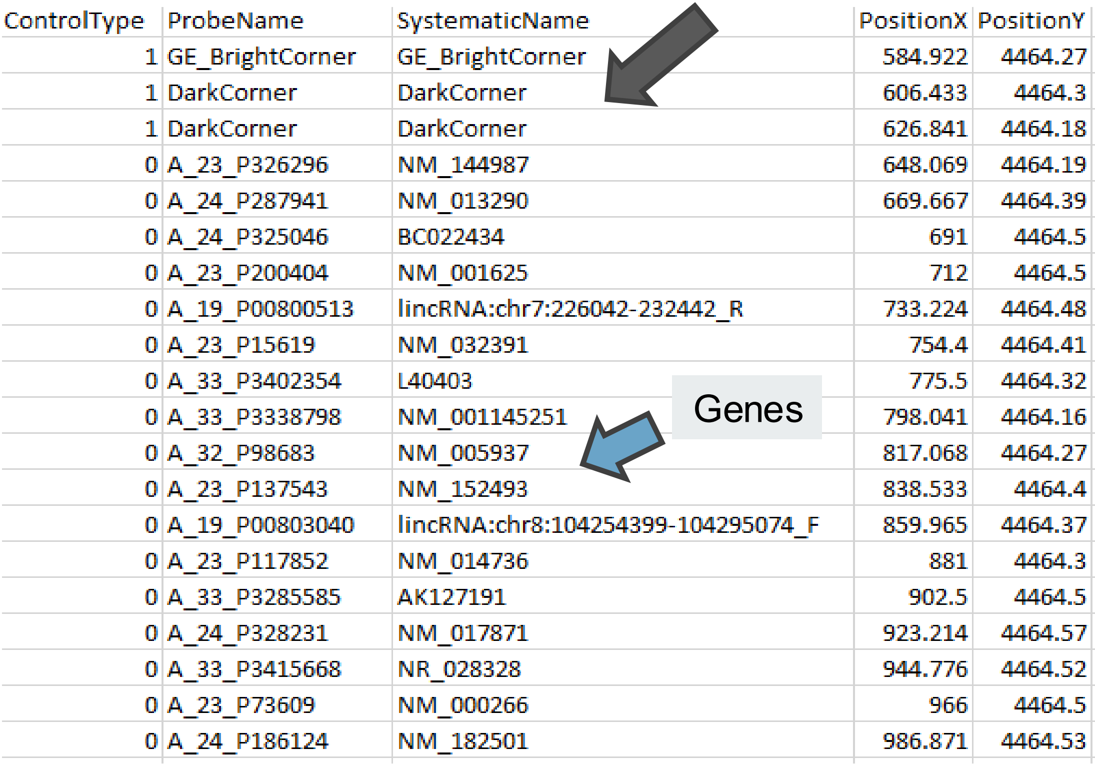
Microarray data processing

Microarray metadata

Positive and negative controls

Position on array

Genes



ControlType	ProbeName	SystematicName	PositionX	PositionY
1	GE_BrightCorner	GE_BrightCorner	584.922	4464.27
1	DarkCorner	DarkCorner	606.433	4464.3
1	DarkCorner	DarkCorner	626.841	4464.18
0	A_23_P326296	NM_144987	648.069	4464.19
0	A_24_P287941	NM_013290	669.667	4464.39
0	A_24_P325046	BC022434	691	4464.5
0	A_23_P200404	NM_001625	712	4464.5
0	A_19_P00800513	lincRNA:chr7:226042-232442_R	733.224	4464.48
0	A_23_P15619	NM_032391	754.4	4464.41
0	A_33_P3402354	L40403	775.5	4464.32
0	A_33_P3338798	NM_001145251	798.041	4464.16
0	A_32_P98683	NM_005937	817.068	4464.27
0	A_23_P137543	NM_152493	838.533	4464.4
0	A_19_P00803040	lincRNA:chr8:104254399-104295074_F	859.965	4464.37
0	A_23_P117852	NM_014736	881	4464.3
0	A_33_P3285585	AK127191	902.5	4464.5
0	A_24_P328231	NM_017871	923.214	4464.57
0	A_33_P3415668	NR_028328	944.776	4464.52
0	A_23_P73609	NM_000266	966	4464.5
0	A_24_P186124	NM_182501	986.871	4464.53

Key data processing steps

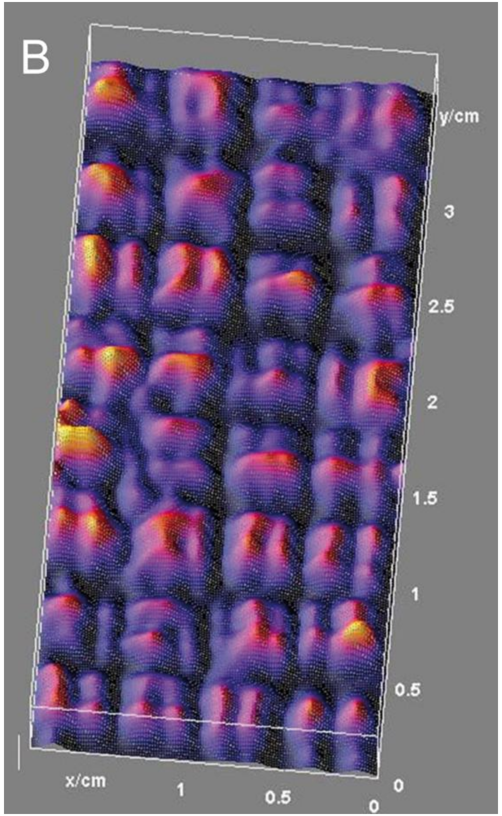
- Mapping probes to genes
  - BLAST to latest genome annotation
  - Already provided for commercial arrays

- Intensity correction
  - Position and sequence bias
  - Perfect match (PM) vs mismatch (MM)

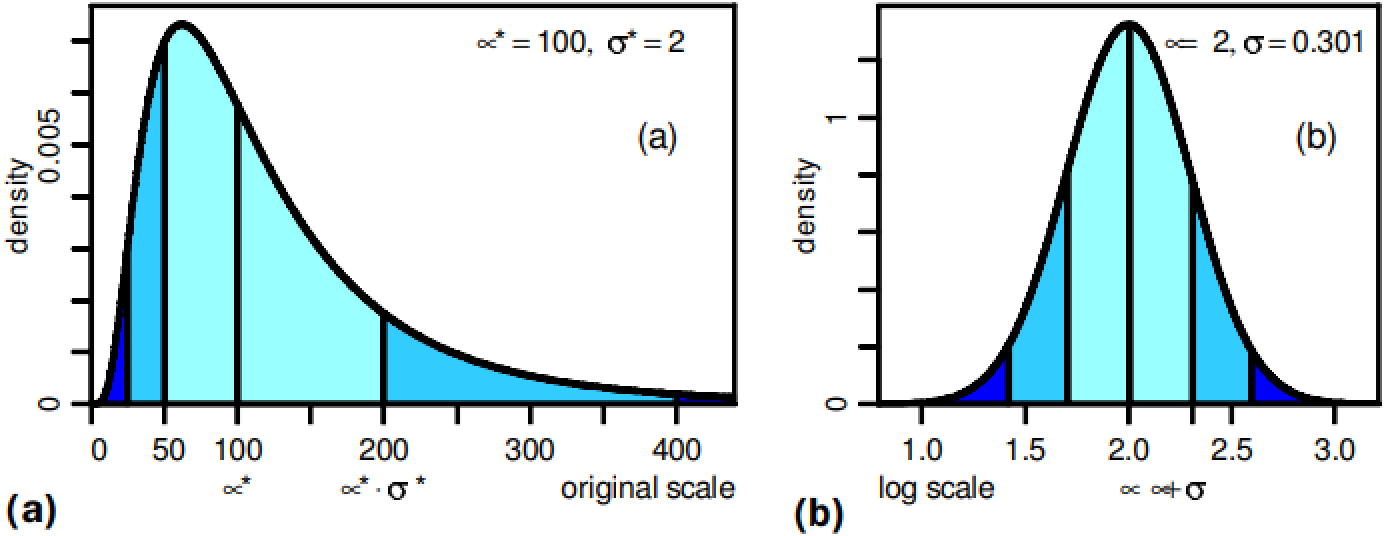
- Outlier removal

- Probe set aggregation for each gene / transcript

- Log-transformation



Log-normal distribution

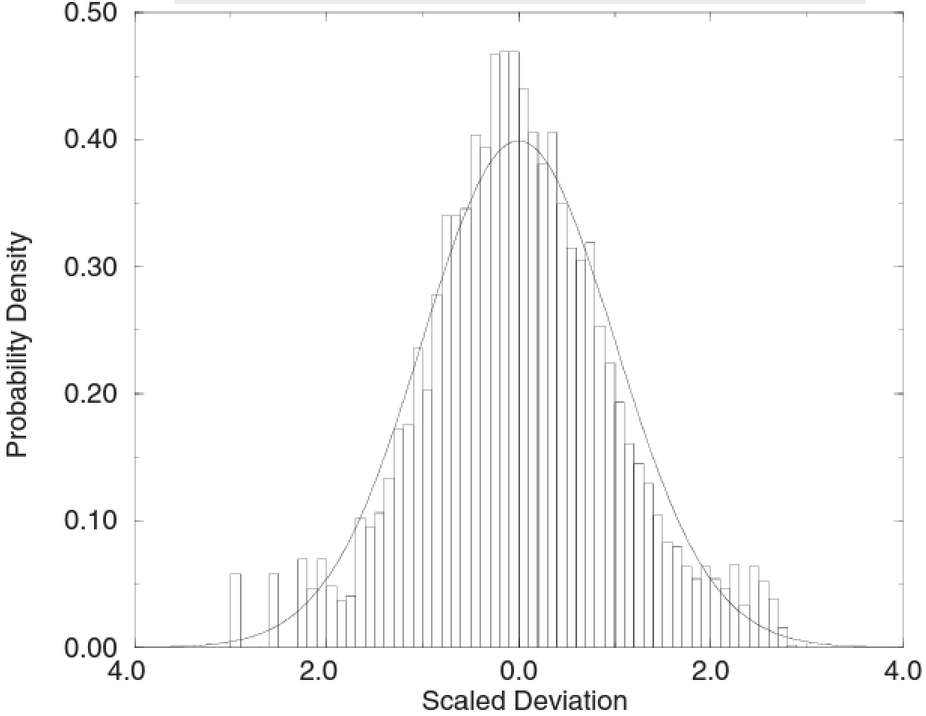


Limpert, Stahel, and Abbt. BioScience 2001.

- Data whose log-transformed values are normally distributed
  - Light and fluorescence intensity, ion count

Microarray data are log-normally distributed

Log-transformed microarray intensity

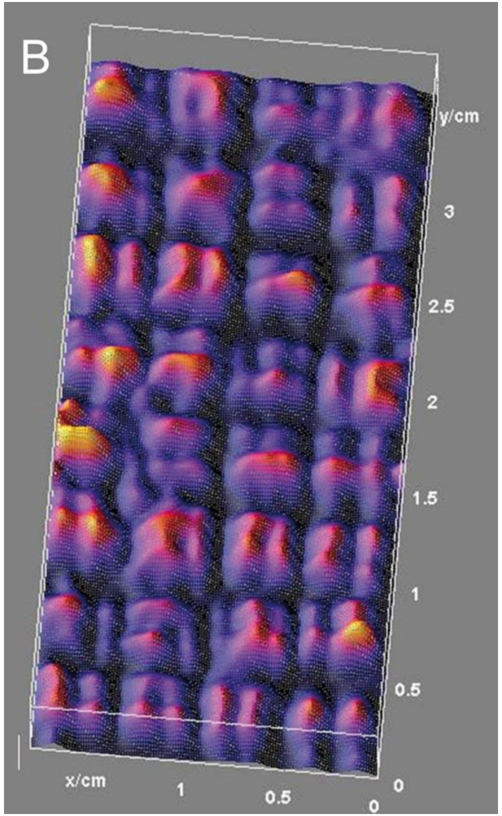


Hoyle, D. et al. Bioinformatics 18:576-584 (2001)

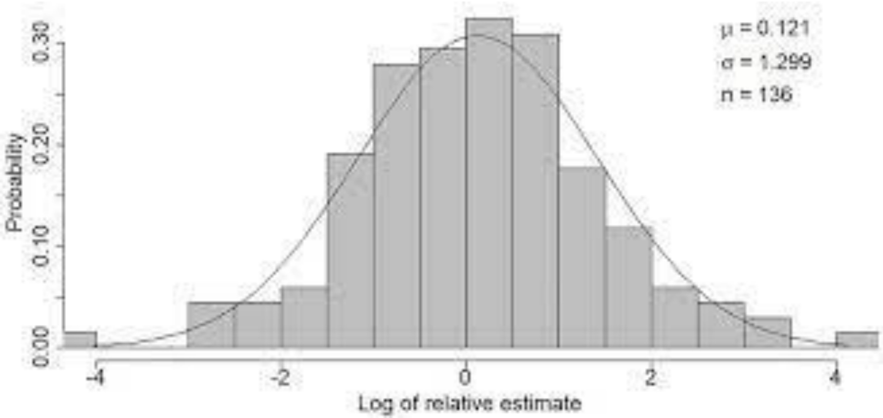
Position-specific noise correction sketch

- **Null Hypothesis**:
  - Noise is normally distributed, and **its mean and variance are independent of the position on array**
    - Mean noise = μ (same across positions)
    - Variance = σ^2 (same across positions)

- **Alternative Hypothesis**:
  - Noise is normally distributed, **its mean depend on the position on array, with a common variance**
    - Mean noise at (x, y) = $ax + by + \mu$
    - Variance = σ^2 (same across positions)



Fitting data to a normal distribution (finding μ and σ^2)



- Consider negative control probe intensities: $n_1, n_2, \ldots, n_k$
  - Assume Normal distribution (μ, σ^2)
  - Likelihood P(data | μ, σ^2) = $\prod_i P(n_i | \mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^k e^{-\frac{1}{2}\sum_i\left(\frac{n_i-\mu}{\sigma}\right)^2}$
  - **Maximum likelihood**: Take the partial derivatives w.r.t. μ and σ^2 and set to 0

Linear effect model

- Negative probe *i* with intensity n_i is located at position (x_i, y_i)

- Hypothesis: $\mu(x_i, y_i) = ax_i + by_i + \mu$
  - Solve for a, b, c that minimize squared difference $\sum_i \left(n_i - (ax_i + by_i + \mu)\right)^2$

- With calculus:
  - $0 = \frac{\delta MSE}{\delta a} = \sum_i 2\left(n_i - (ax_i + by_i + \mu)\right)(-x_i)$
  - $0 = \frac{\delta MSE}{\delta b} = \sum_i 2\left(n_i - (ax_i + by_i + \mu)\right)(-y_i)$
  - $0 = \frac{\delta MSE}{\delta \mu} = \sum_i 2\left(n_i - (ax_i + by_i + \mu)\right)(-1)$

The algebra is not as bad as it looks

- Three linear equations with three variables!
  - $0 = \sum_i 2\left(n_i - (ax_i + by_i + \mu)\right)(-x_i)$
  - $0 = \sum_i 2\left(n_i - (ax_i + by_i + \mu)\right)(-y_i)$
  - $0 = \sum_i 2\left(n_i - (ax_i + by_i + \mu)\right)(-1)$

- Or equivalently
  - $a \sum_i x_i^2 + b \sum_i x_i y_i + \mu \sum_i x_i = \sum_i n_i x_i$
  - $a \sum_i x_i y_i + b \sum_i y_i^2 + \mu \sum_i y_i = \sum_i n_i y_i$
  - $a \sum_i x_i + b \sum_i y_i + k\mu = \sum_i n_i$

- Most of the terms are numbers from your data

Hypothesis testing for likelihoods

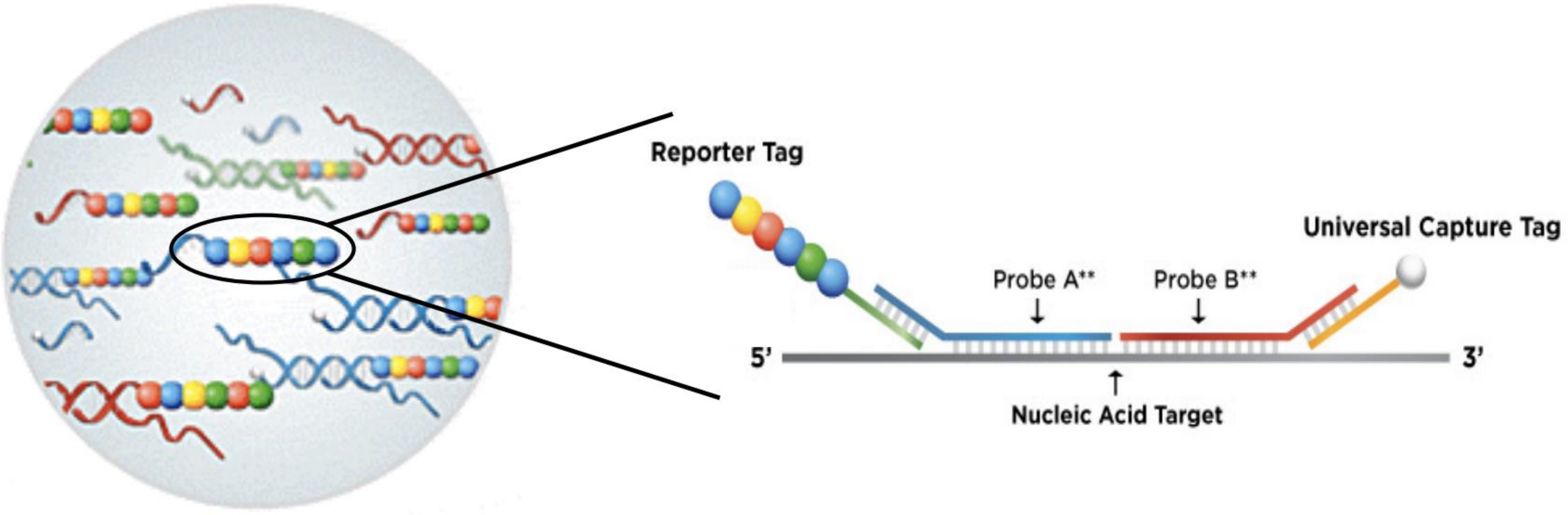
- Likelihood for Null Hypothesis (2 parameters, μ, σ^2)
- Likelihood for Alternate Hypothesis (4 parameters, a, b, μ, σ^2)

- **Information criterion**
  - **Akaike**: AIC = 2*k* – 2 log(likelihood)
  - **Bayesian**: BIC = log(*n*) *k* – 2 log(likelihood)
  - Favor model with **low number of parameters (*k*) and high likelihood**

- **Nested model / likelihood ratio test**
  - Score = – 2 x delta log(likelihood)
  - Chi-square difference test of score with delta degree of freedom = 4 – 2

Nanostring

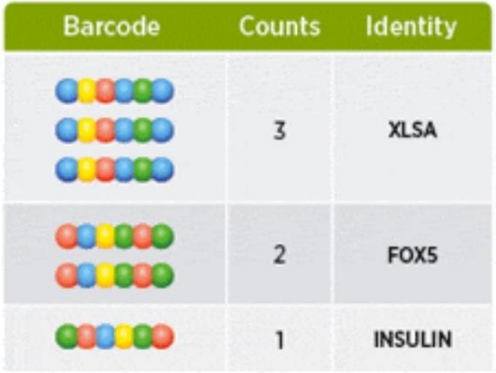
Transcript-specific probes & fluorescence barcodes



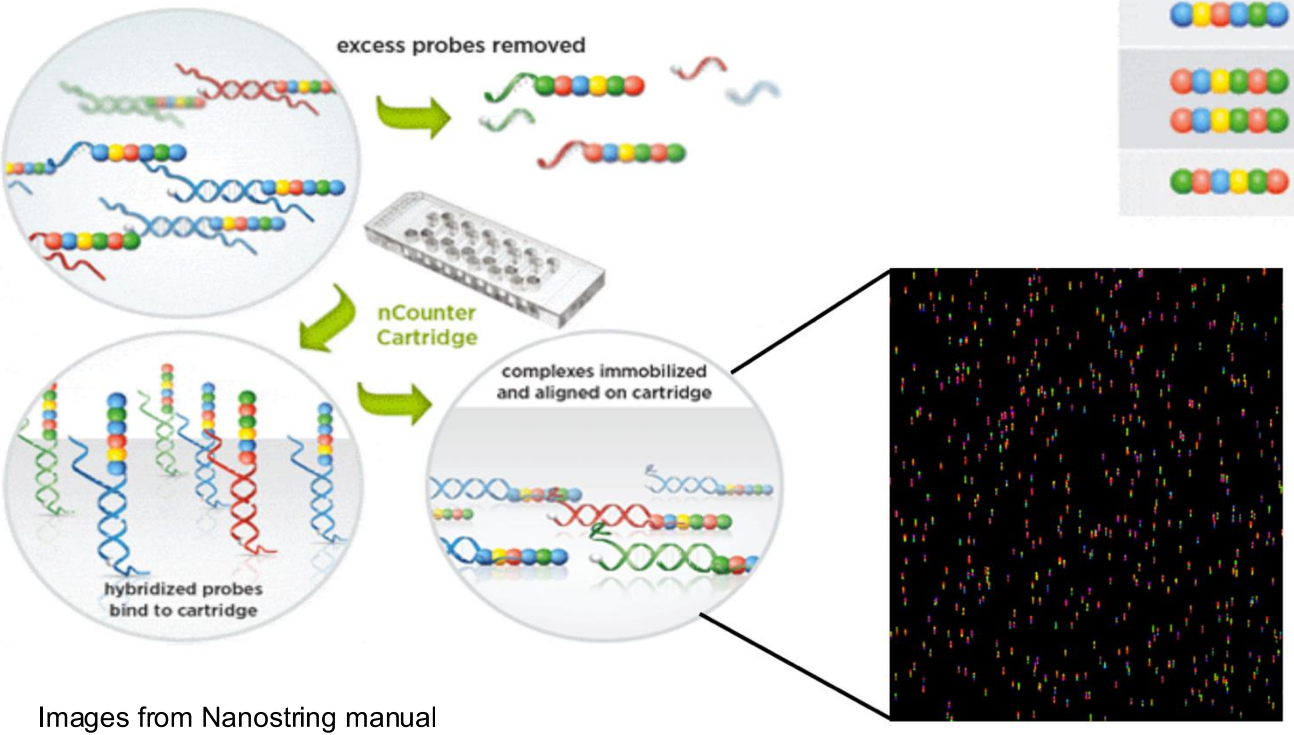
solution phase hybridization

Images from Nanostring manual

Counting number of molecules

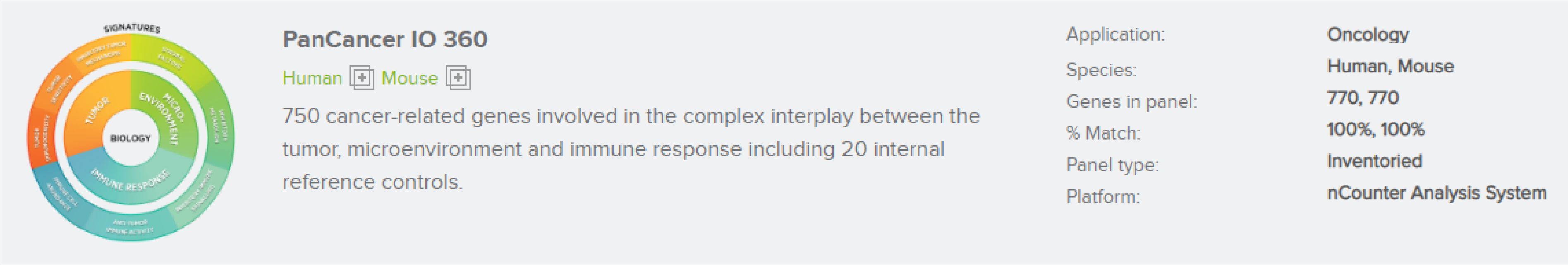


Barcode	Counts	Identity
🔵🟡🔴🔵🟢🔵 🔵🟡🔴🔵🟢🔵 🔵🟡🔴🔵🟢🔵	3	XLSA
🔴🔵🟡🟢🔴🟢 🔴🔵🟡🟢🔴🟢	2	FOXS
🟢🔴🔵🟡🟢🔴	1	INSULIN



Images from Nanostring manual

Prebuilt gene set (up to ~800 targets)

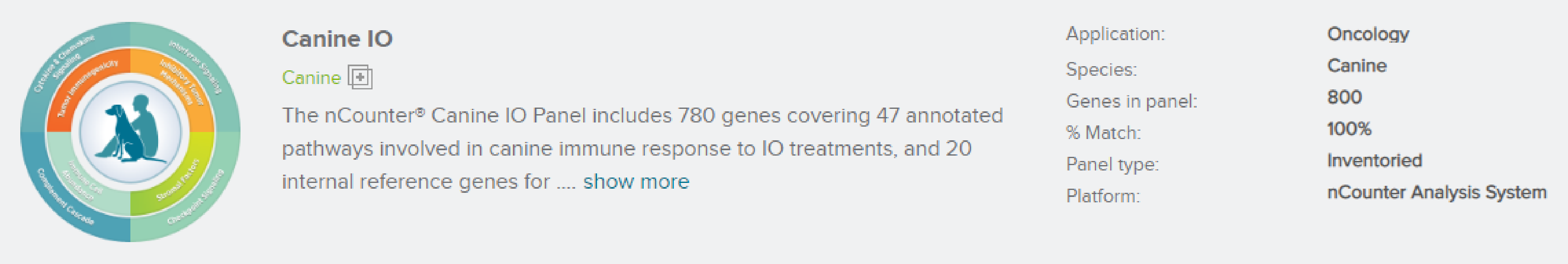


PanCancer IO 360

Human ⊞ Mouse ⊞

750 cancer-related genes involved in the complex interplay between the tumor, microenvironment and immune response including 20 internal reference controls.

Application:	Oncology
Species:	Human, Mouse
Genes in panel:	770, 770
% Match:	100%, 100%
Panel type:	Inventoried
Platform:	nCounter Analysis System



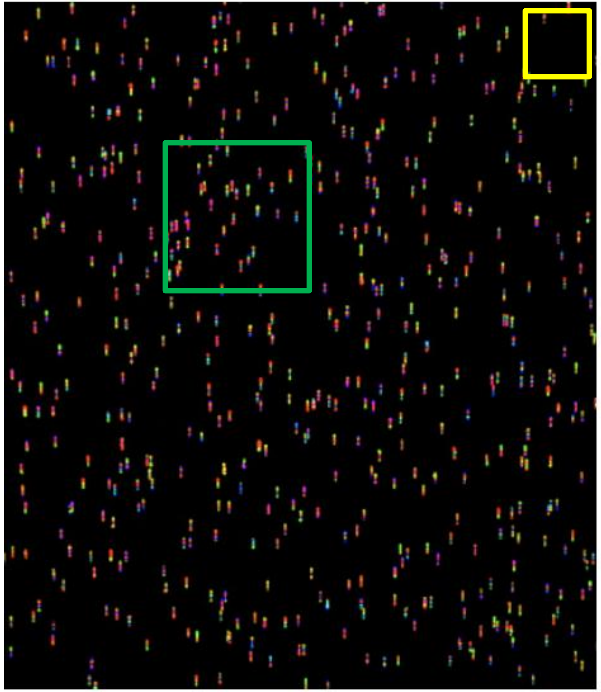
Canine IO

Canine ⊞

The nCounter® Canine IO Panel includes 780 genes covering 47 annotated pathways involved in canine immune response to IO treatments, and 20 internal reference genes for .... show more

Application:	Oncology
Species:	Canine
Genes in panel:	800
% Match:	100%
Panel type:	Inventoried
Platform:	nCounter Analysis System

Nanostring's built-in quality control

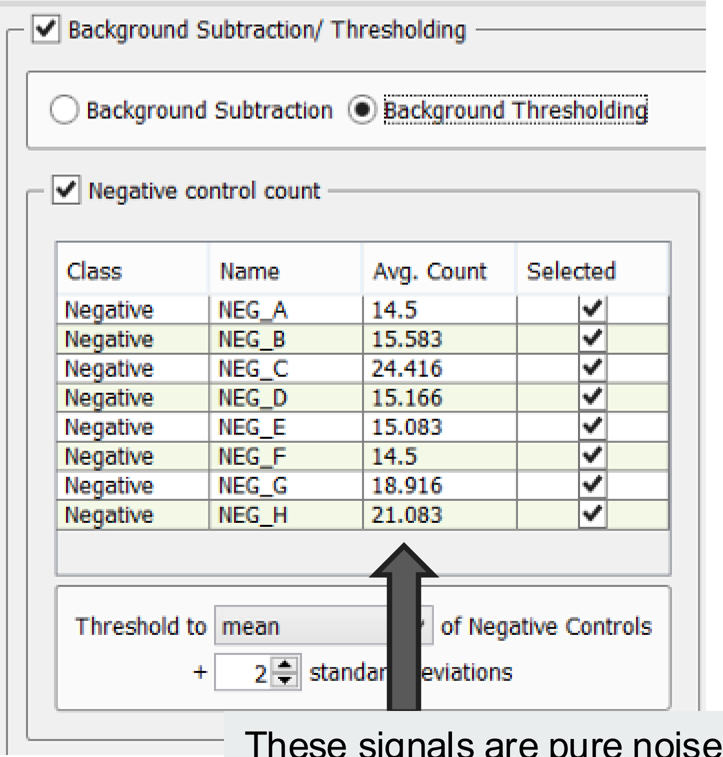


- Imaging QC
  - % of successful imaging field of view > 75%
- Binding QC
  - 0.1-2 molecules per square micron
- Positive control
  - Six synthetic DNA ranging from 0.125-128 fM
- Negative control
  - Eight synthetic DNA that do not bind to probe

Nanostring data processing

Through nCounter / nSolver software

Using negative and positive controls to normalize



☑ Background Subtraction/ Thresholding

○ Background Subtraction ● Background Thresholding

☑ Negative control count

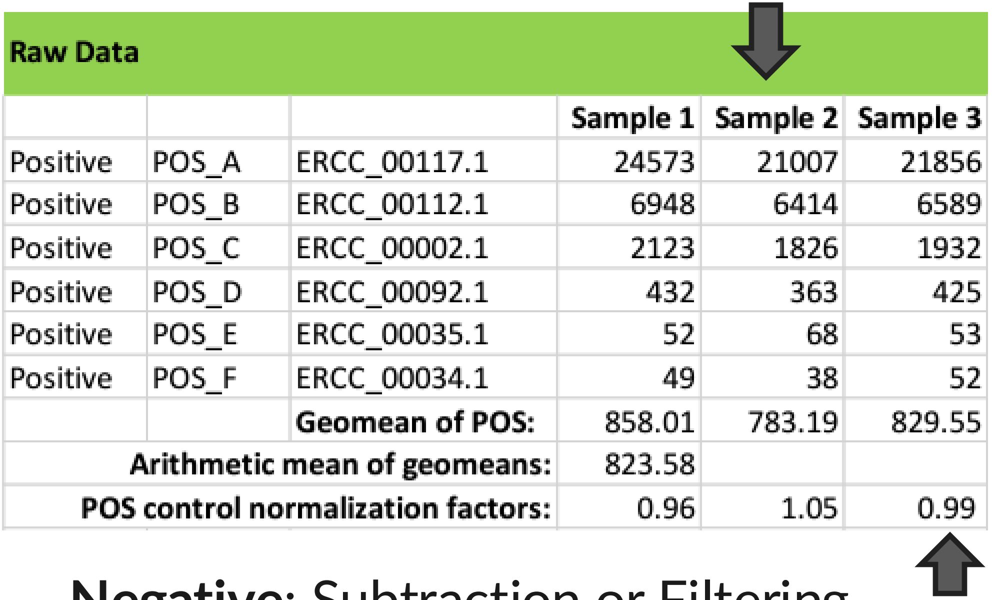
Class	Name	Avg. Count	Selected
Negative	NEG_A	14.5	✓
Negative	NEG_B	15.583	✓
Negative	NEG_C	24.416	✓
Negative	NEG_D	15.166	✓
Negative	NEG_E	15.083	✓
Negative	NEG_F	14.5	✓
Negative	NEG_G	18.916	✓
Negative	NEG_H	21.083	✓

Threshold to mean of Negative Controls

+ 2 standar eviations

These signals are pure noises

Positive control probes added with known concentrations

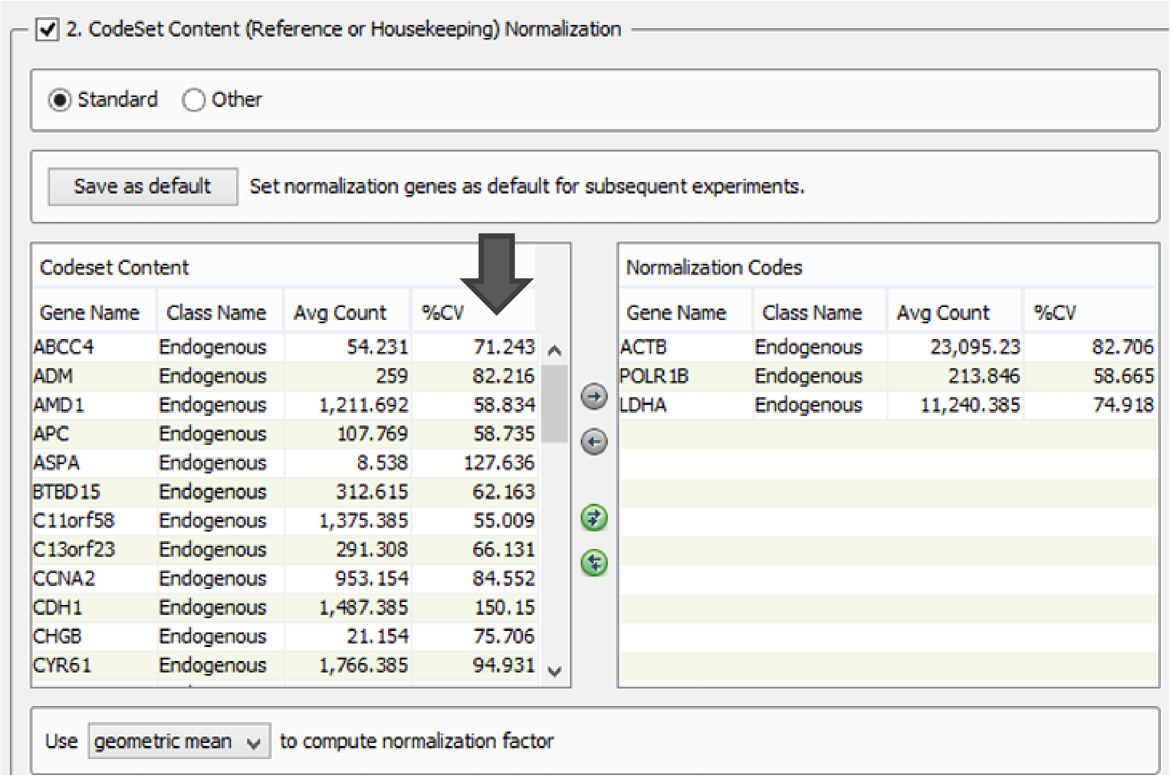


Raw Data

			Sample 1	Sample 2	Sample 3
Positive	POS_A	ERCC_00117.1	24573	21007	21856
Positive	POS_B	ERCC_00112.1	6948	6414	6589
Positive	POS_C	ERCC_00002.1	2123	1826	1932
Positive	POS_D	ERCC_00092.1	432	363	425
Positive	POS_E	ERCC_00035.1	52	68	53
Positive	POS_F	ERCC_00034.1	49	38	52
		Geomean of POS:	858.01	783.19	829.55
Arithmetic mean of geomeans:			823.58		
POS control normalization factors:			0.96	1.05	0.99

- **Negative**: Subtraction or Filtering
- **Positive**: Scale data by geometric mean of ratios between observed / expected

Housekeeping genes as control



☑ 2. CodeSet Content (Reference or Housekeeping) Normalization

● Standard ○ Other

Save as default — Set normalization genes as default for subsequent experiments.

Codeset Content

Gene Name	Class Name	Avg Count	%CV
ABCC4	Endogenous	54.231	71.243
ADM	Endogenous	259	82.216
AMD1	Endogenous	1,211.692	58.834
APC	Endogenous	107.769	58.735
ASPA	Endogenous	8.538	127.636
BTBD15	Endogenous	312.615	62.163
C11orf58	Endogenous	1,375.385	55.009
C13orf23	Endogenous	291.308	66.131
CCNA2	Endogenous	953.154	84.552
CDH1	Endogenous	1,487.385	150.15
CHGB	Endogenous	21.154	75.706
CYR61	Endogenous	1,766.385	94.931

Normalization Codes

Gene Name	Class Name	Avg Count	%CV
ACTB	Endogenous	23,095.23	82.706
POLR1B	Endogenous	213.846	58.665
LDHA	Endogenous	11,240.385	74.918

Use geometric mean to compute normalization factor

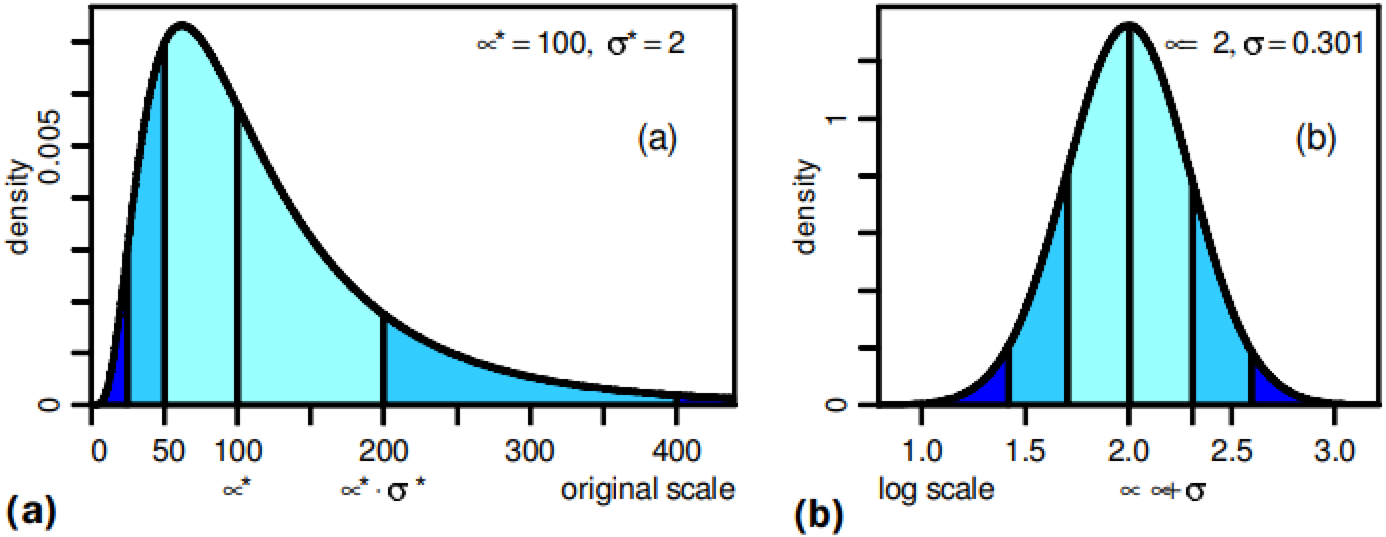
- **Housekeeping genes** are essential genes with basic cellular functions that should be stable across conditions

- Assume similar abundance across all samples
  - **Coefficient of Variation:** CV = S.D. / mean

Arithmetic versus geometric mean and logarithm

- AM = $\frac{x_1 + \cdots + x_n}{n}$
- GM = $\sqrt[n]{x_1 \ldots x_n}$
- Two data points: 1 and 10000
  - AM = 5000.5
  - GM = 100 (leaning toward smaller values)
- Three data points: 1, 10, and 100
  - AM = 37
  - GM = 10 (leaning toward smaller values)

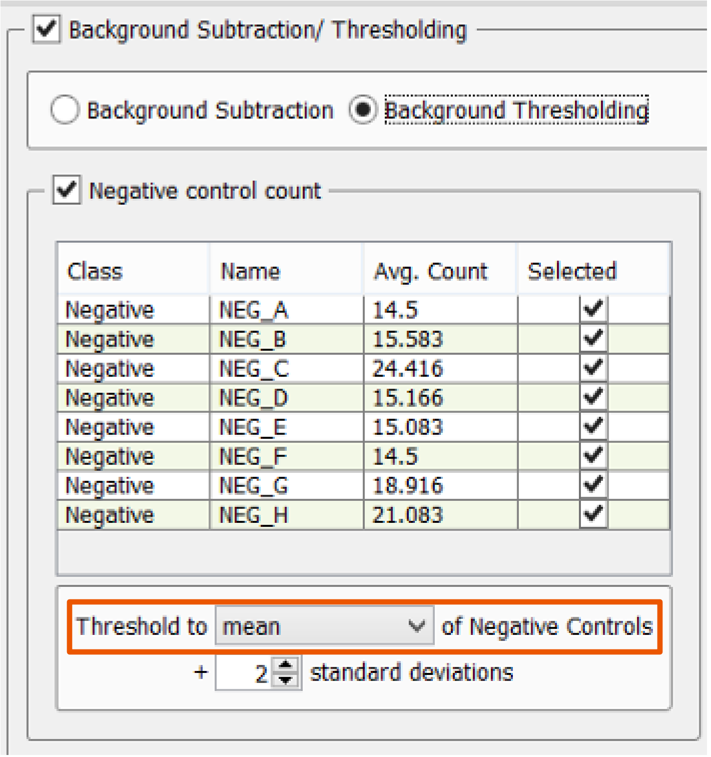
Arithmetic versus geometric mean and logarithm



Limpert, Stahel, and Abbt. BioScience 2001.

- $\frac{\log(x_1)+\cdots+\log(x_n)}{n} = \log\left(\sqrt[n]{x_1 \ldots x_n}\right)$ → AM of log data = GM of original data

Arithmetic mean for background noises



☑ Background Subtraction/ Thresholding

○ Background Subtraction ● Background Thresholding

☑ Negative control count

Class	Name	Avg. Count	Selected
Negative	NEG_A	14.5	☑
Negative	NEG_B	15.583	☑
Negative	NEG_C	24.416	☑
Negative	NEG_D	15.166	☑
Negative	NEG_E	15.083	☑
Negative	NEG_F	14.5	☑
Negative	NEG_G	18.916	☑
Negative	NEG_H	21.083	☑

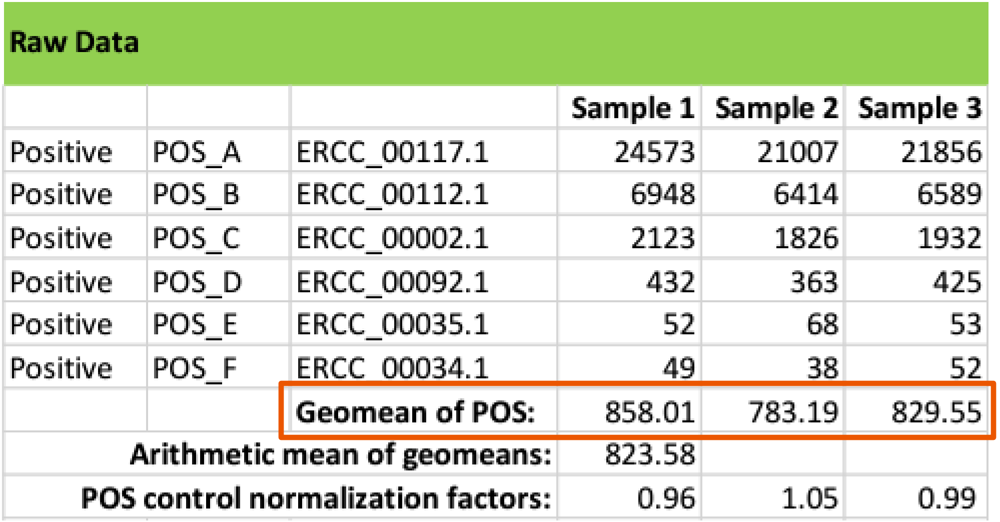
Threshold to mean of Negative Controls

+ 2 standard deviations

- Background noises are assumed to be **Normally distributed**

- **Arithmetic mean** is used

Geometric mean for molecular counts



Raw Data

			Sample 1	**Sample 2**	**Sample 3**
Positive	POS_A	ERCC_00117.1	24573	21007	21856
Positive	POS_B	ERCC_00112.1	6948	6414	6589
Positive	POS_C	ERCC_00002.1	2123	1826	1932
Positive	POS_D	ERCC_00092.1	432	363	425
Positive	POS_E	ERCC_00035.1	52	68	53
Positive	POS_F	ERCC_00034.1	49	38	52
		Geomean of POS:	858.01	783.19	829.55
Arithmetic mean of geomeans:			823.58		
POS control normalization factors:			0.96	1.05	0.99

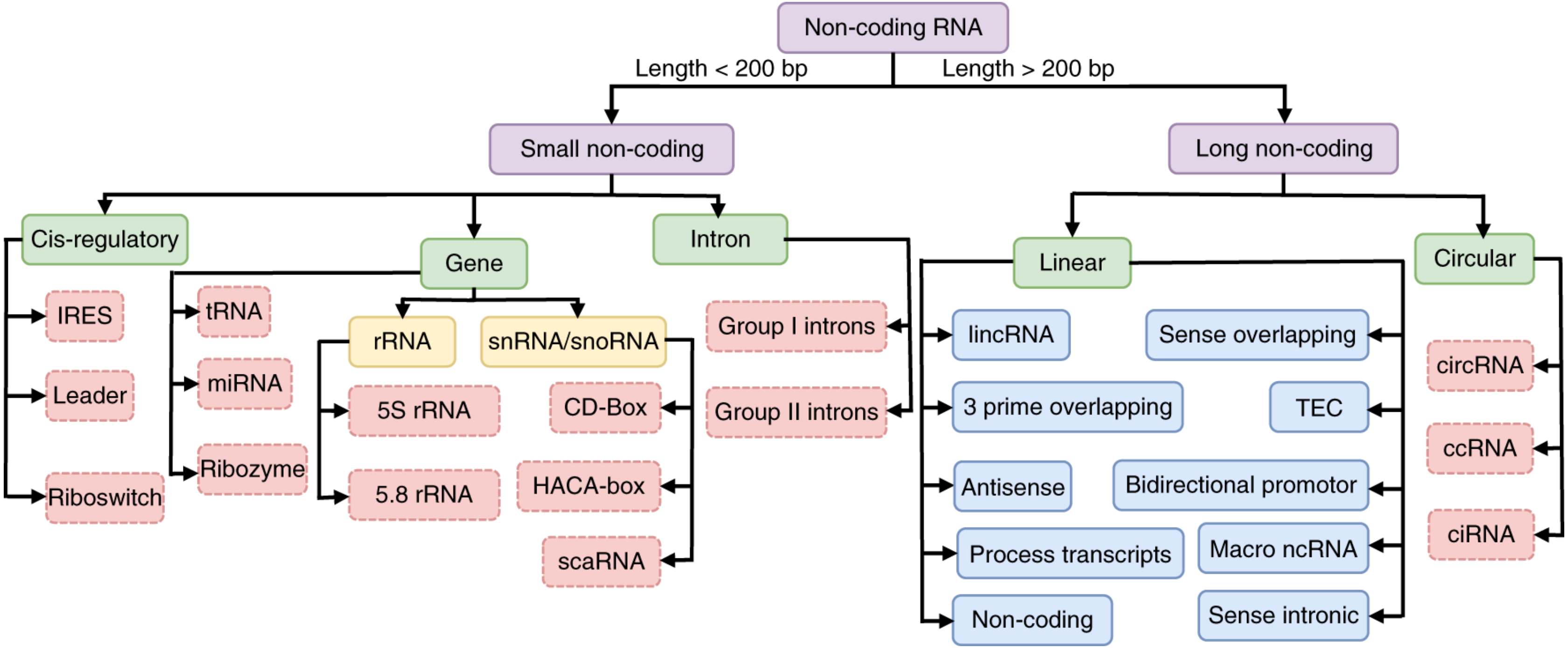
- Real expression data are assumed to be **log-normally distributed**

- **Geometric mean** is used

- Equivalent to log-transforming the data first and then use arithmetic mean

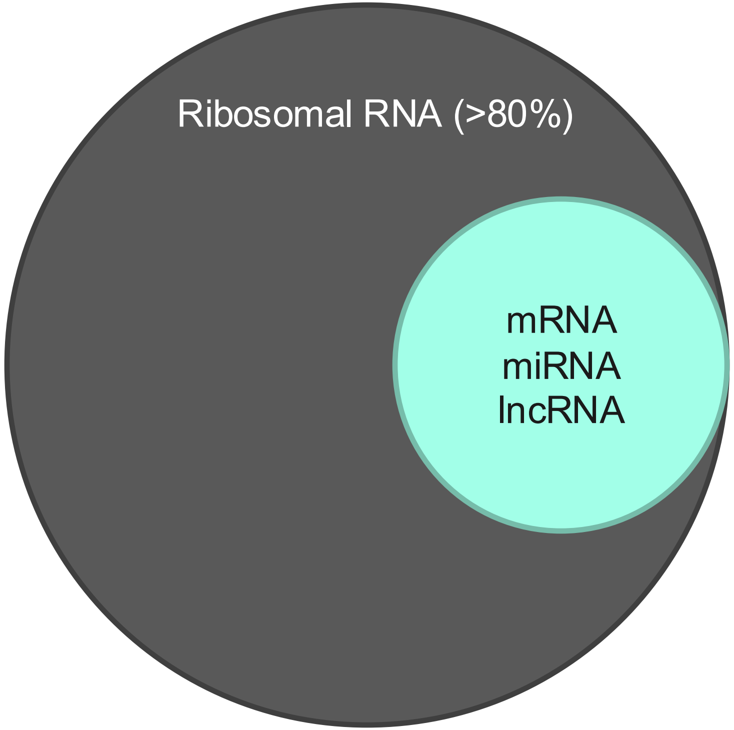
RNA sequencing

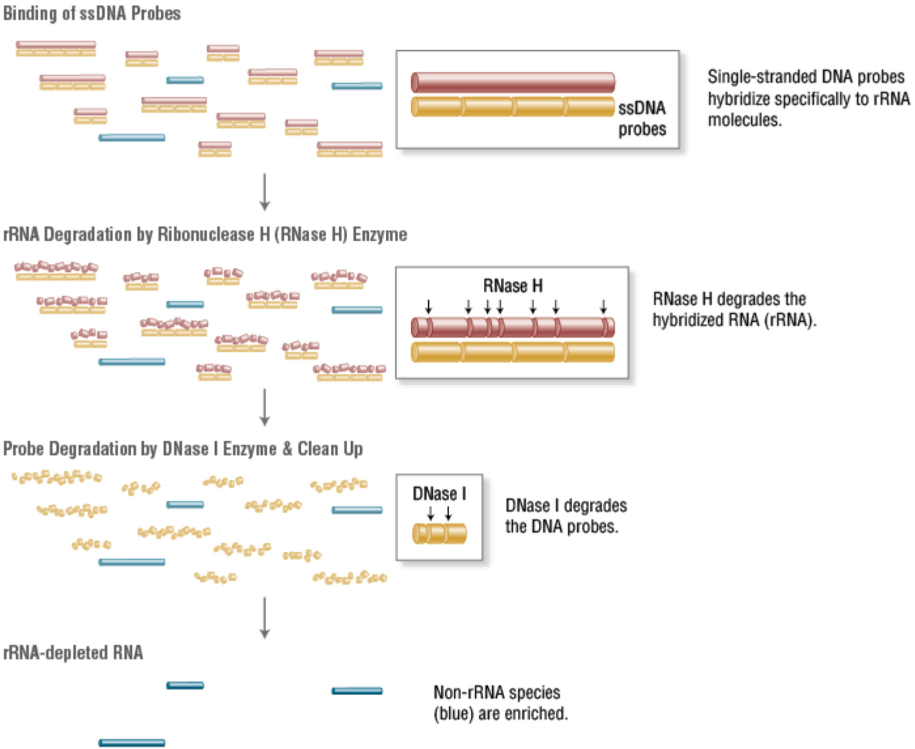
Reminder: There are non-coding RNAs



Amin, N. et al. Nature Machine Intelligence 1:246-256 (2019)

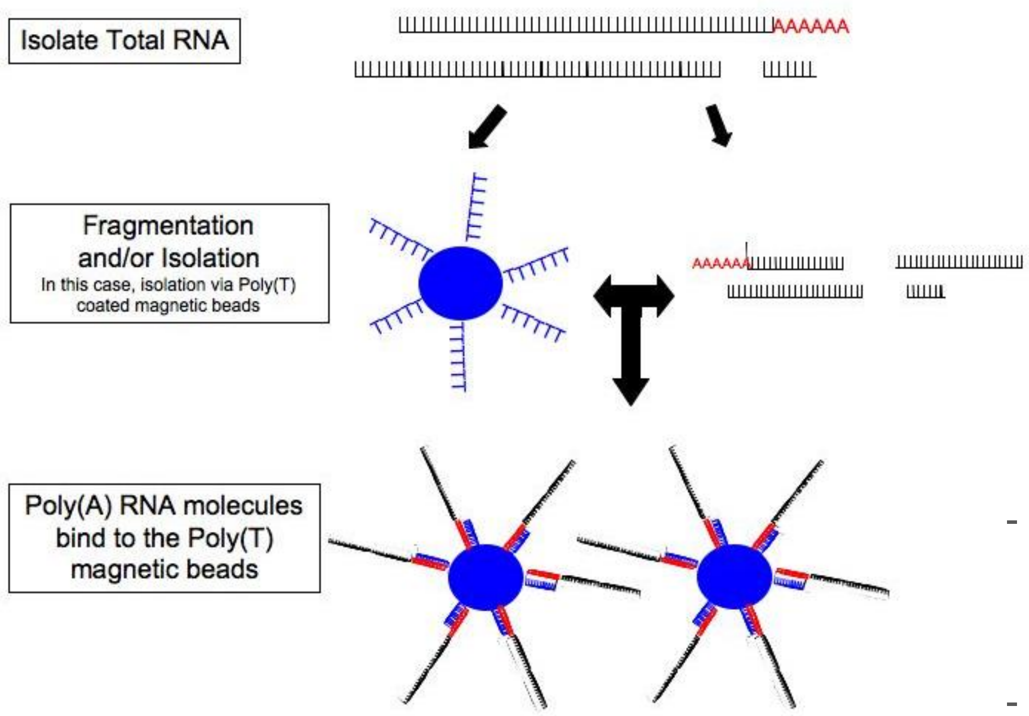
Total RNA sequencing = removal of rRNA




Binding of ssDNA Probes
ssDNA probes
Single-stranded DNA probes hybridize specifically to rRNA molecules.
rRNA Degradation by Ribonuclease H (RNase H) Enzyme
RNase H
RNase H degrades the hybridized RNA (rRNA).
Probe Degradation by DNase I Enzyme & Clean Up
DNase I
DNase I degrades the DNA probes.
rRNA-depleted RNA
Non-rRNA species (blue) are enriched.

Source: New England BioLabs

Enrichments of mRNA and miRNA

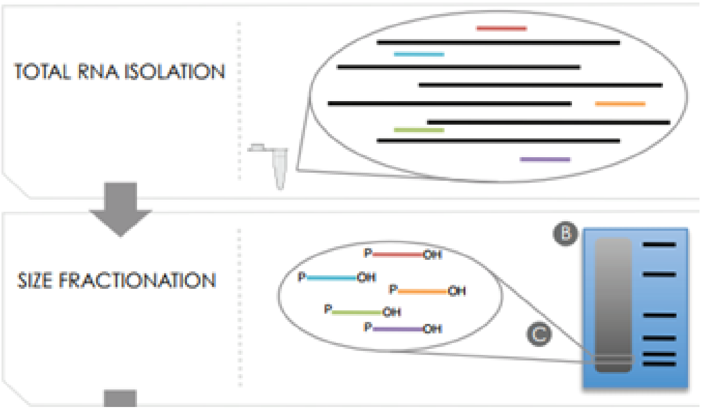


Isolate Total RNA

Fragmentation and/or Isolation

In this case, isolation via Poly(T) coated magnetic beads

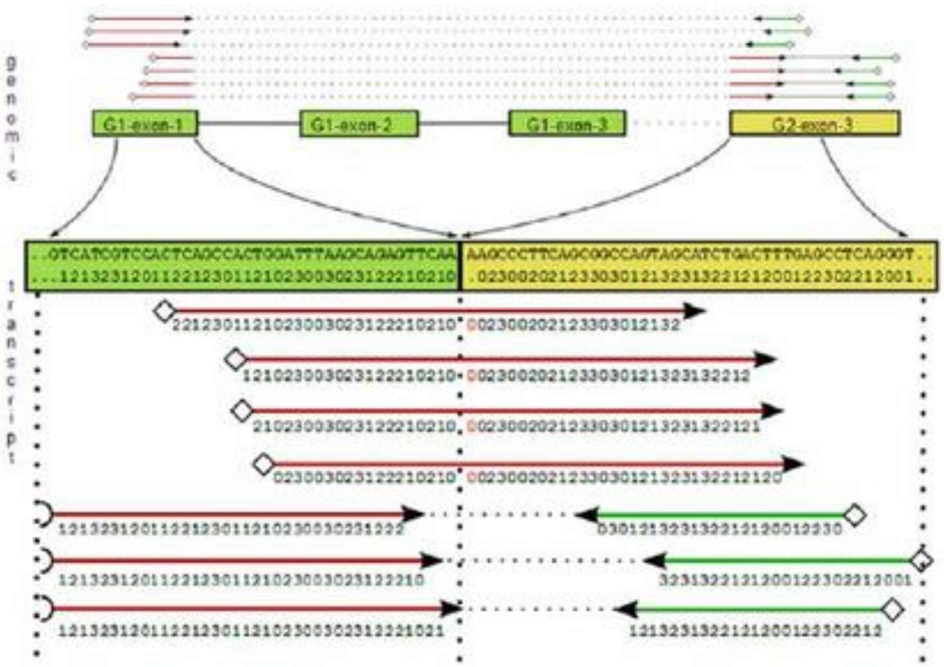
Poly(A) RNA molecules bind to the Poly(T) magnetic beads



- Selection by polyT probe or size fractionation

- Some lncRNA also have polyA

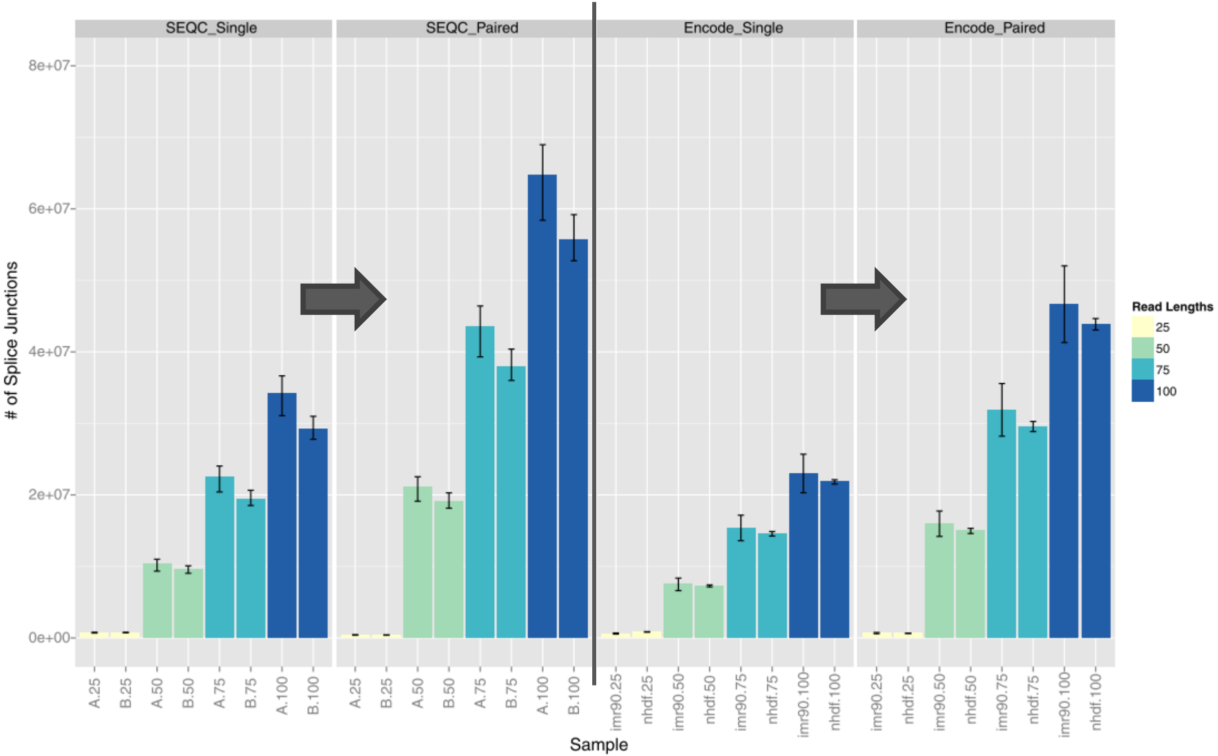
Transcript isoform and splice junction detection



Sakarya, O. et al. PLoS Comp Biol 8:e1002464 (2012)

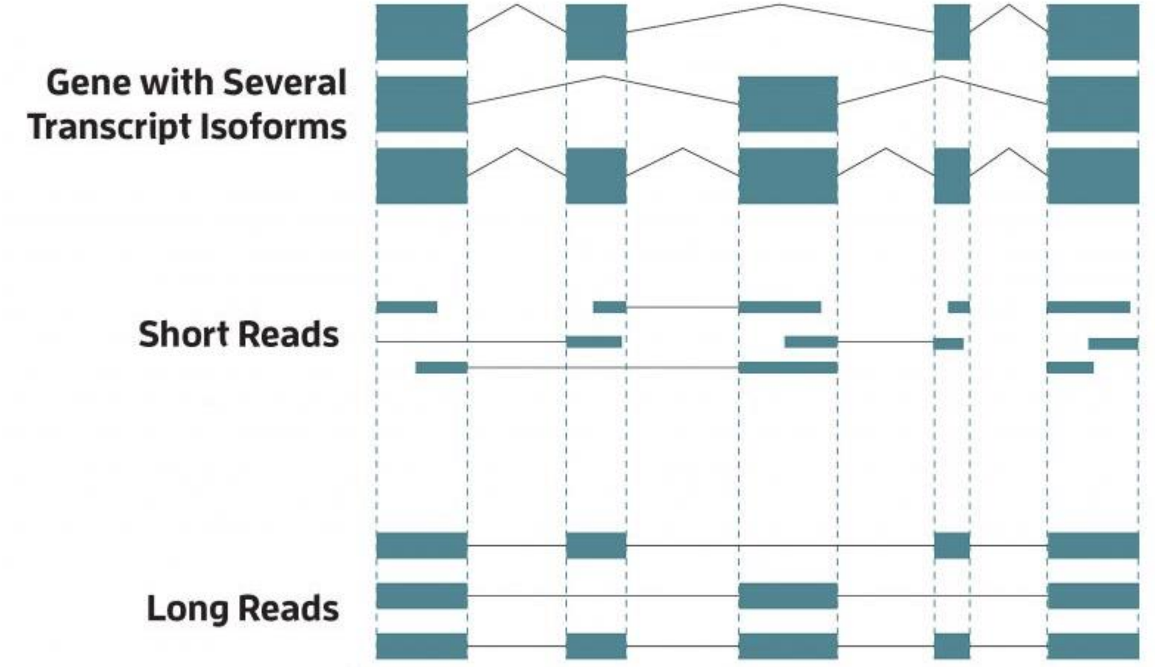
- One gene can express multiple RNA transcript isoforms
- Due to splicing
- Different isoforms can have different functions and tissue specificities
- Only some reads can detect

Paired-end sequencing with longer read length



Chhangawala *et al.* Genome Biology 16:131 (2015)

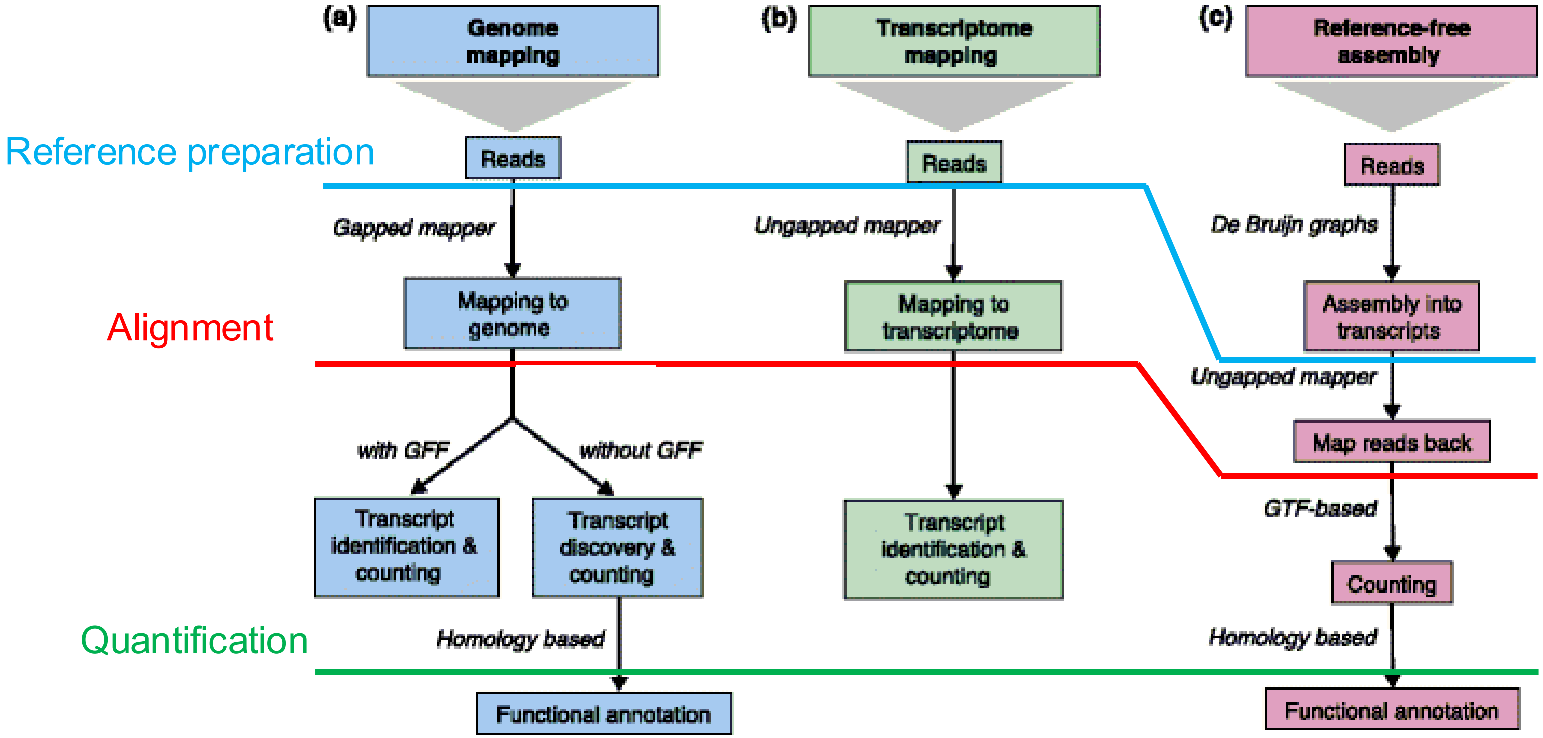
Complete isoform details with long-read data



https://www.genengnews.com/resources/tutorial/full-length-transcript-sequencing-no-assembly-required/

RNA-seq data processing

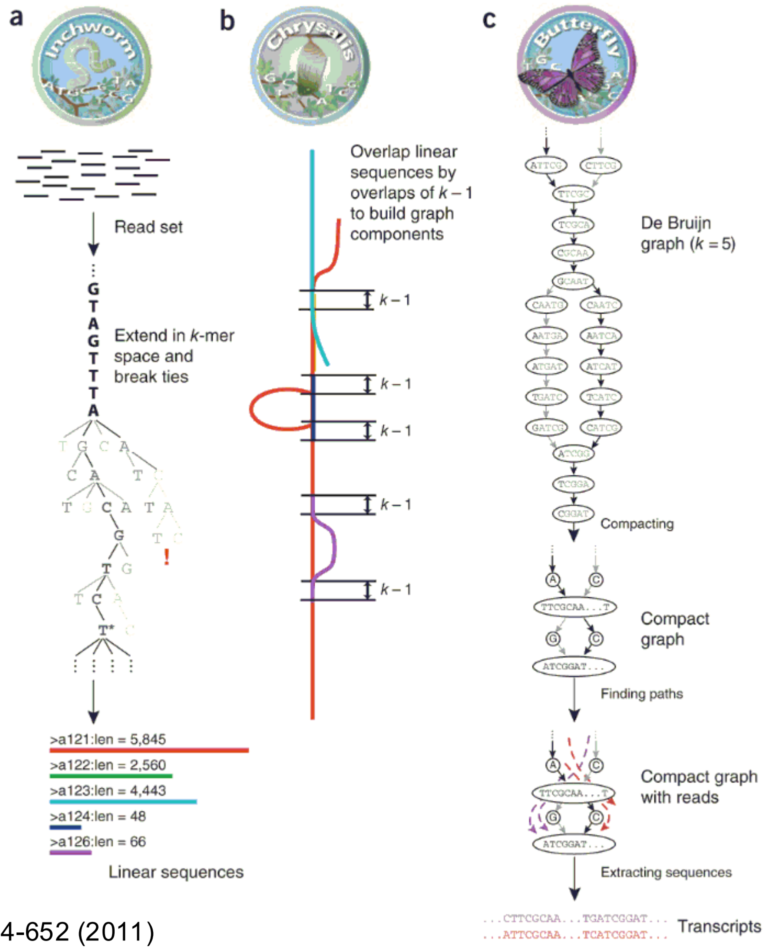
RNA-seq analysis pipelines



Conesa *et al.* Genome Biology 17:13 (2016)

De novo transcript assembly

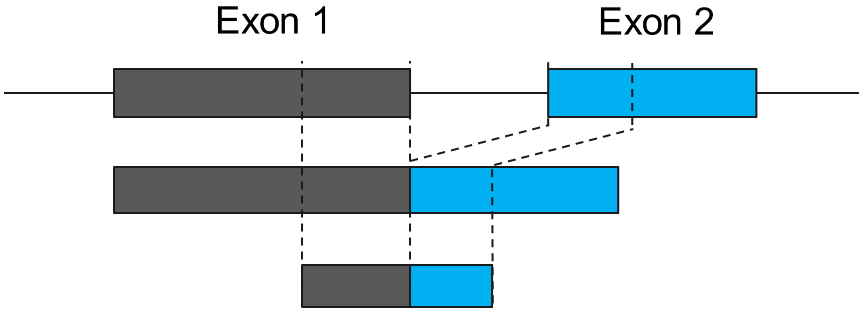
- For non-model organisms with no reference genome nor transcriptome
- Detect new isoforms
- Used as transcript database for re-alignment and quantification
- **Trinity**



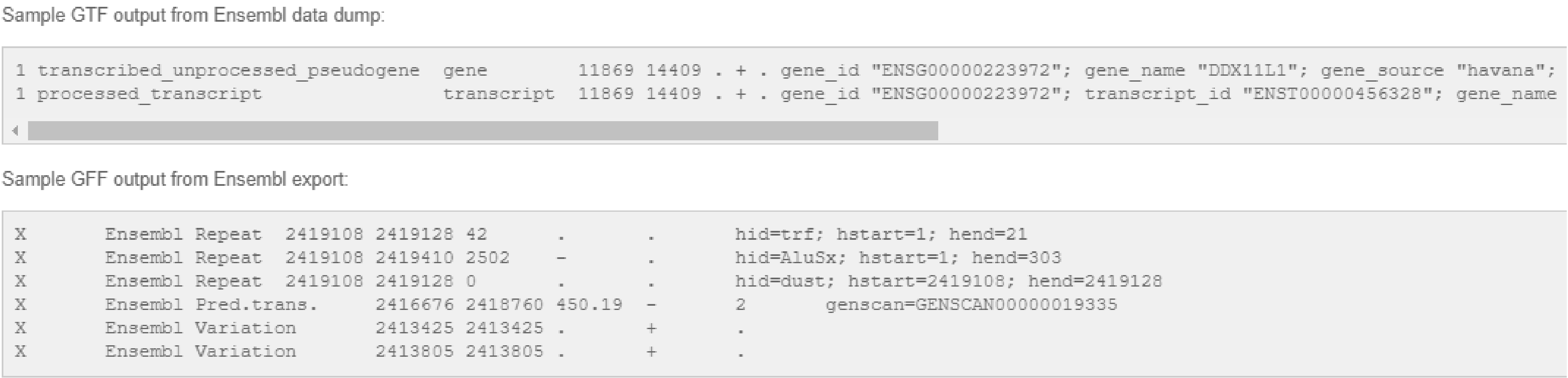
Grabherr *et al.* Nat Biotech 29:644-652 (2011)

Alignment to reference genome or transcriptome

- Reference transcriptome
  - **Fast**, cannot discover new isoform
  - **Ungapped**, *k*-mer-based alignment
  - **salmon / kallisto**
- Reference genome
  - **Slow**, but can detect new isoforms
  - **Gapped alignment**, allow for intron
  - Can be guided by exon annotations
  - **STAR, HISAT2**



GTF/GFF genome annotation format



Sample GTF output from Ensembl data dump:

```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana";
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name
```

Sample GFF output from Ensembl export:

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

- Tab-separated text file
- Chromosome ID, object name, base pair positions, strand, and other annotation details

Transcriptomics technique summary

- RNA-seq can detect broad RNA molecules
- Nanostring provides the most accurate quantification
- Microarray is the cheapest platform

Next lecture’s agenda

- Rapid RNA-seq alignment with *k*-mer
- Gene expression units
- Negative binomial model for gene expression data
- Differential expression analysis

Any question?

- See you next time