3000788 Intro to Comp Molec Biol

Week 4: Metagenomics and microarray

Fall 2024

Sira Sriswasdi, PhD

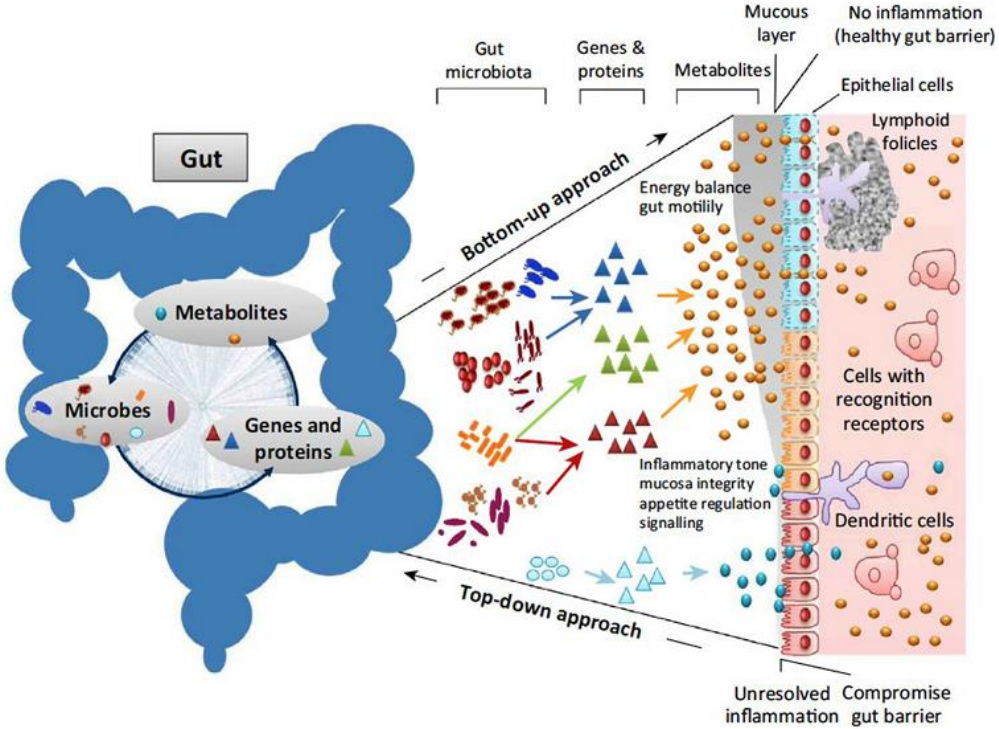
- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Part I: Metagenomics

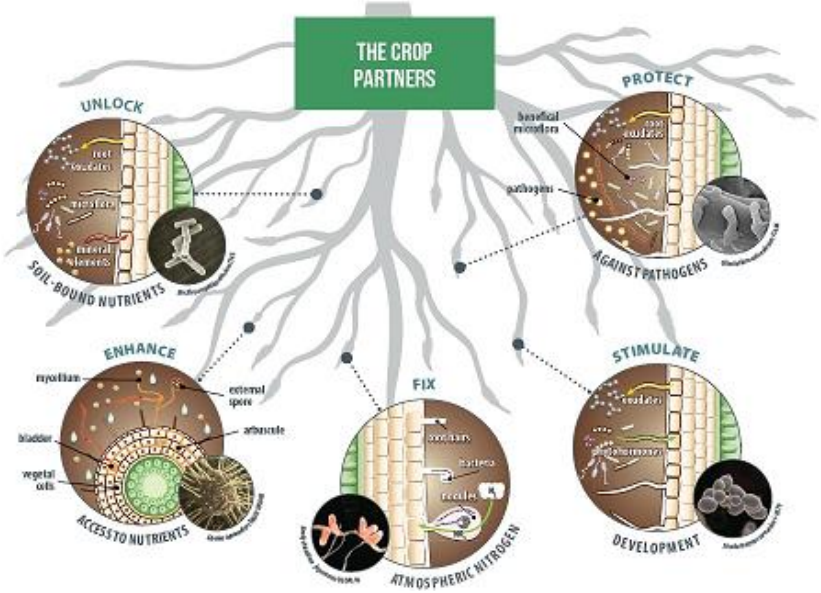
- Mixture data – sequencing reads from multiple species
- Environmental samples → Monitoring & discovery
- Capture species that cannot be isolated / cultured
- Challenging to process – alignment and assembly

Microbiome and meta-omics

Microbiomes



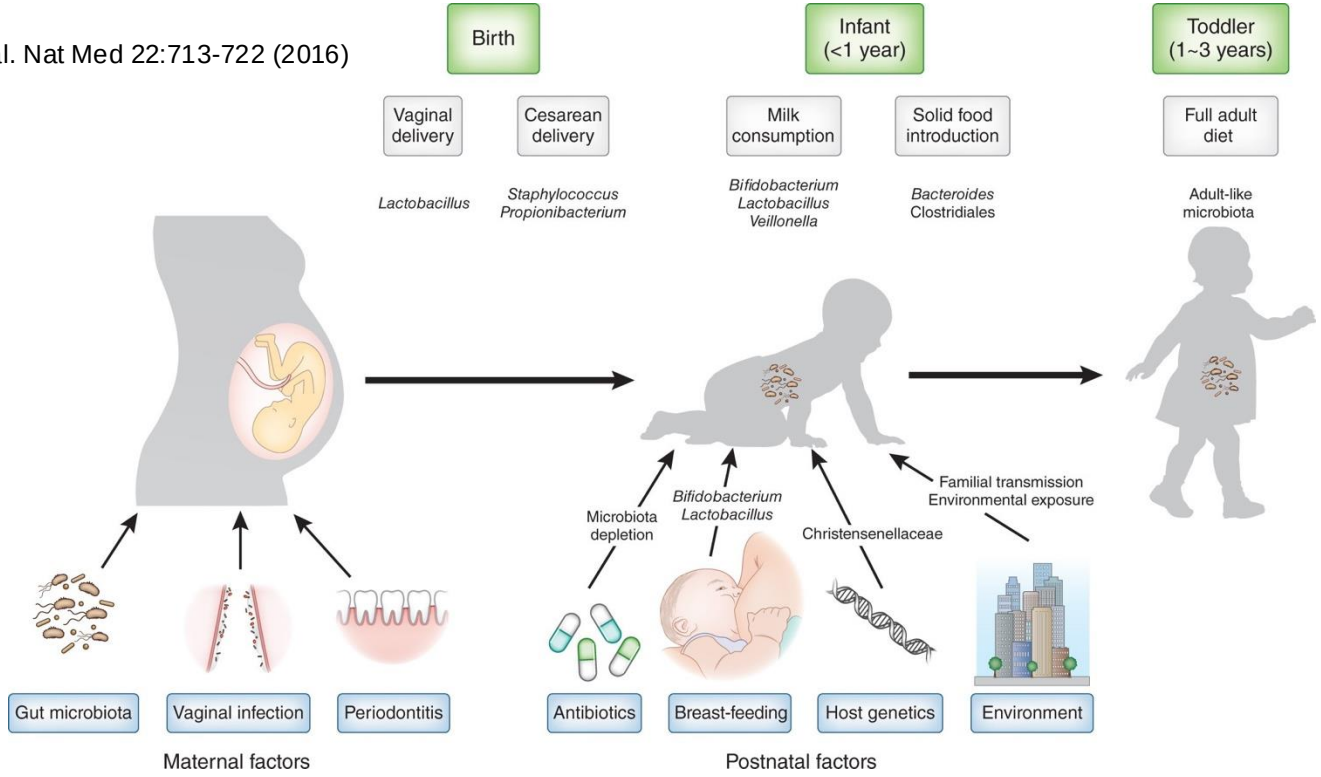
Moya and Ferrer, Trends in Microbiology 24: 402-413 (2016)



http://www.lallemandplantcare.com/en/our-solutions/rhizosphere-inoculants/

Microbiome is dynamics

Tamburini, S. et al. Nat Med 22:713-722 (2016)



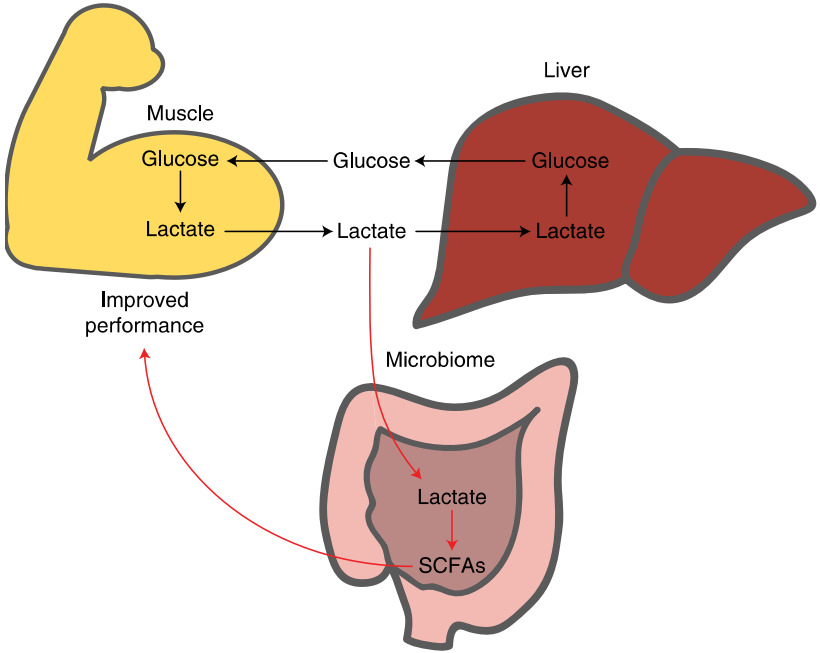
Mostly metagenomics

- Metatranscriptomics:
  - Difficult sample preparation
  - RNA are fragile
  - Challenging to determine whether a gene is ON or OFF in which subpopulations

- Metaproteomics:
  - Require reference protein database to interpret mass spectrometry data

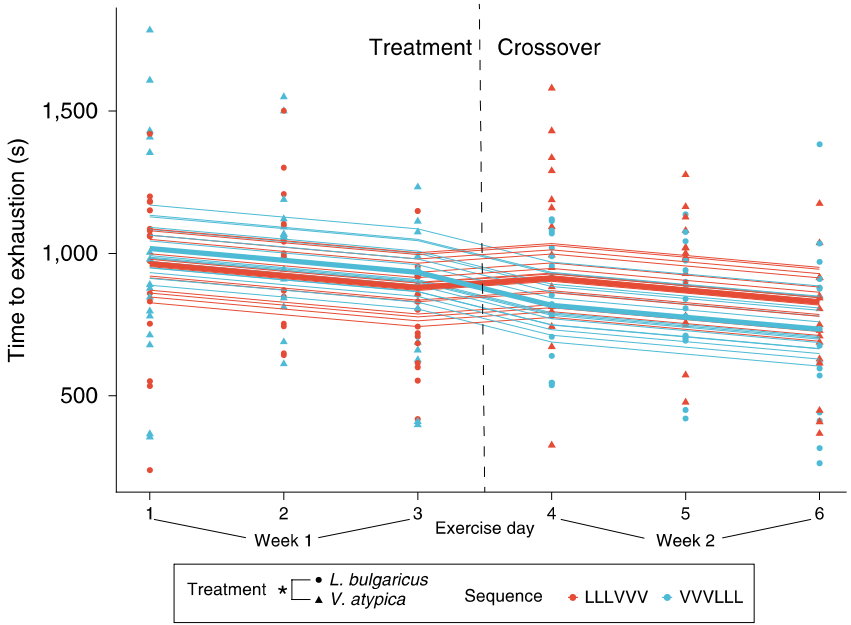
Applications of meta-omics

Lactate-utilizing bacteria in athelete guts

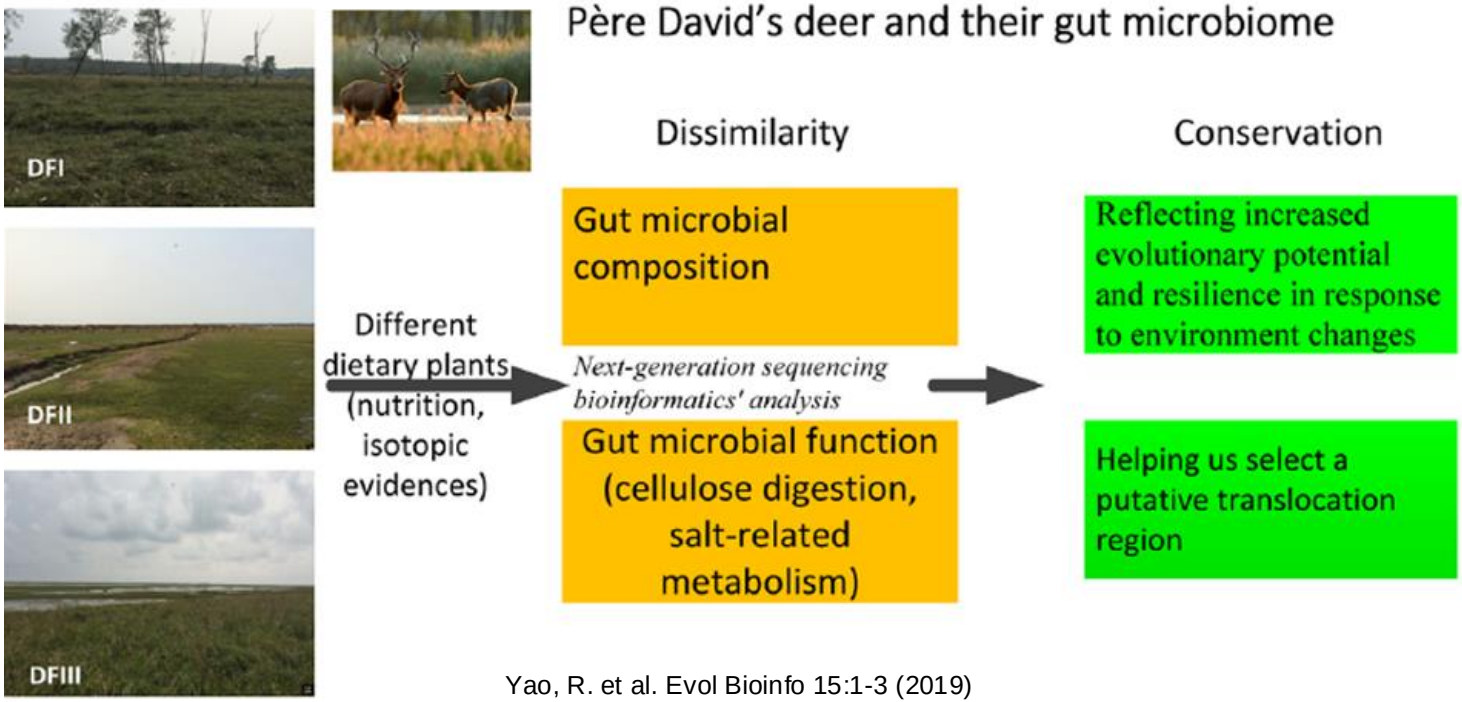


Scheiman *et al*. Nat Med 25:1104-1109 (2019)

Increased athletic ability in mice with transplanted microbiome

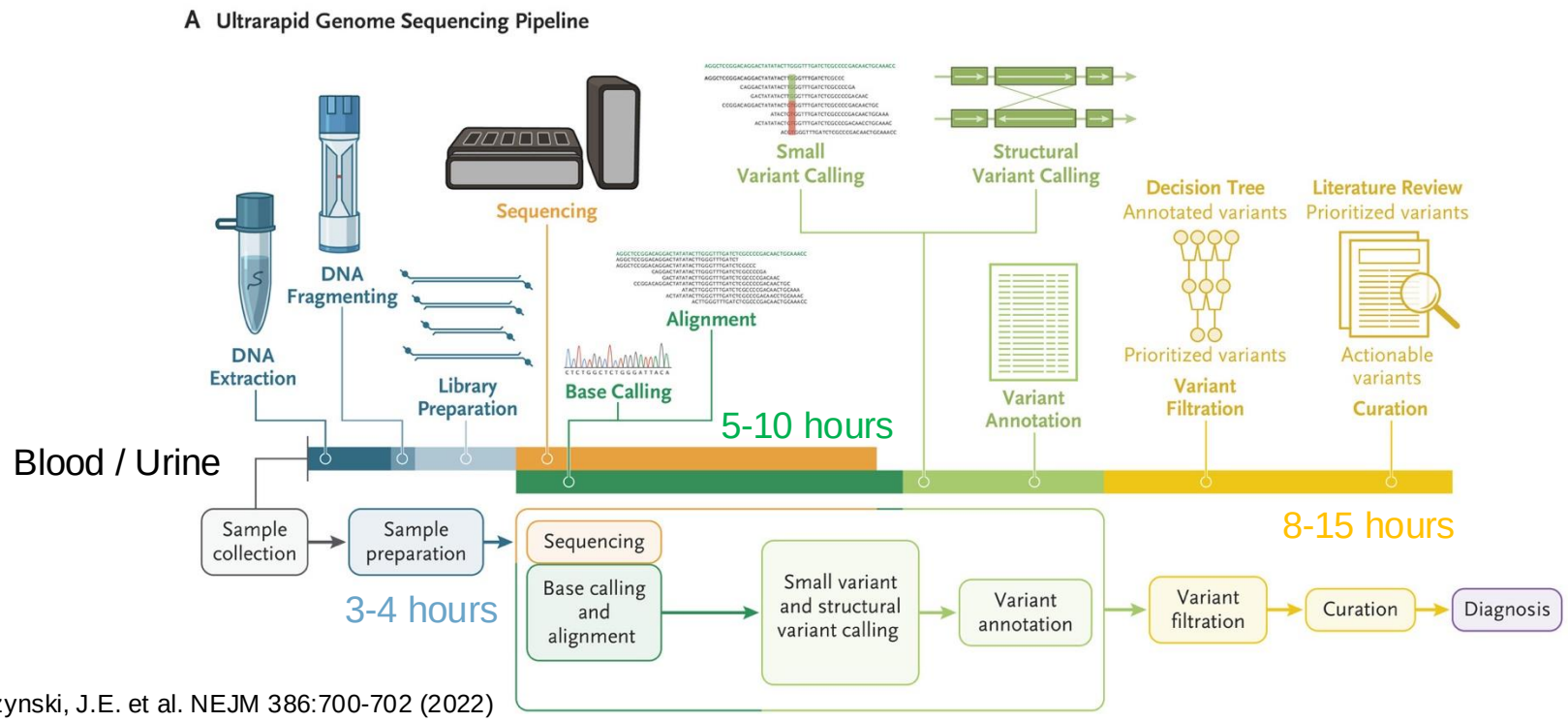


Wildlife conservation



Yao, R. et al. Evol Bioinfo 15:1-3 (2019)

Rapid pathogen detection for clinical decision



Gorzynski, J.E. et al. NEJM 386:700-702 (2022)

Research questions in meta-omics

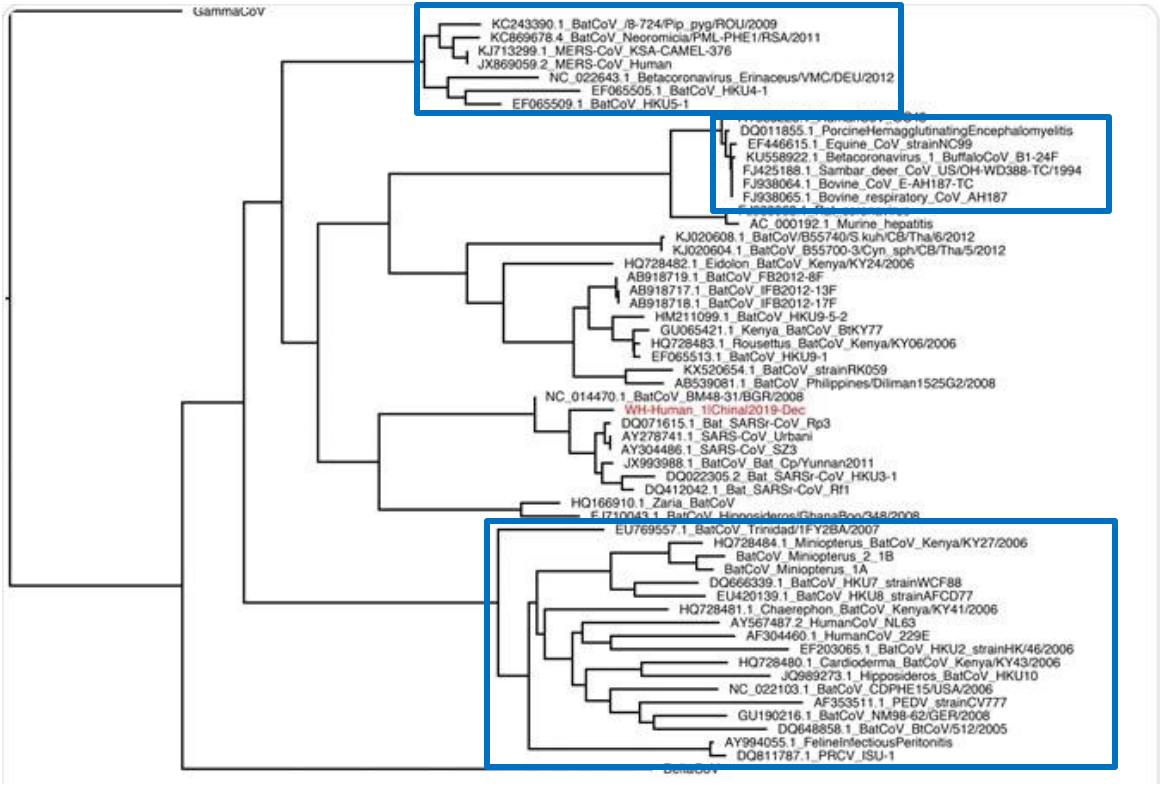
- Health
  - Host-pathogen association
  - Drug resistance genes
  - Gut microbiome, cancer microbiome
- Ecology
  - Change in microbiome due to human actions
    - Factory and hospital wastes
    - Global warming
  - Microbiome of extreme conditions
- Agriculture = pathogens and yield
- Surveillance

Challenges in meta-omics

- Grouping of DNA from the same organism – facilitate assembly
- Gene operon structure is required for functional interpretation
  - Read assembly
- Small genomic differences across species and sub-species
- Presence of plasmids

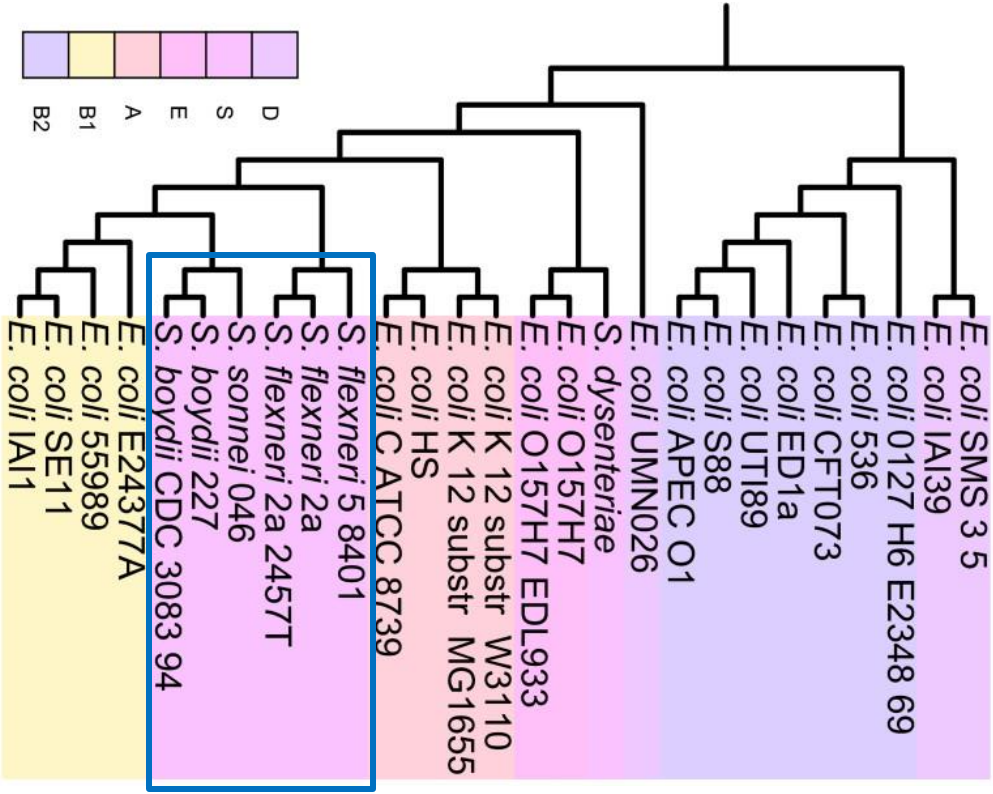
Operational texonomic unit (OTU)

Cluster of sequences with high similarity



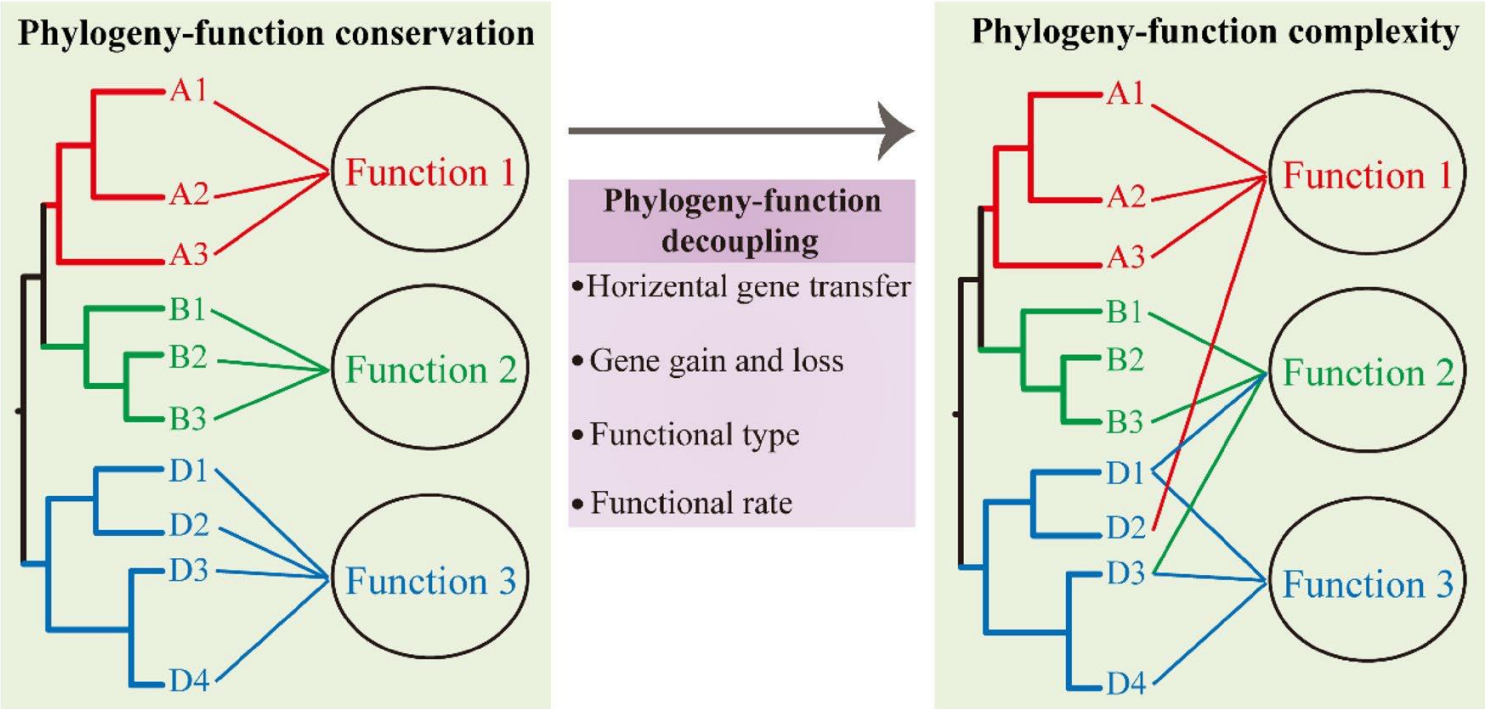
Blurry boundary between species

- Genus *Shigella* are pathogens that evolved from an *E. coli* ancestor
- 80-90% similarity to some *E. coli* clades
- Definition of taxonomy may require both genotypes and phenotypes



Skippington and Ragan, BMC Genomics 12:532 (2011)

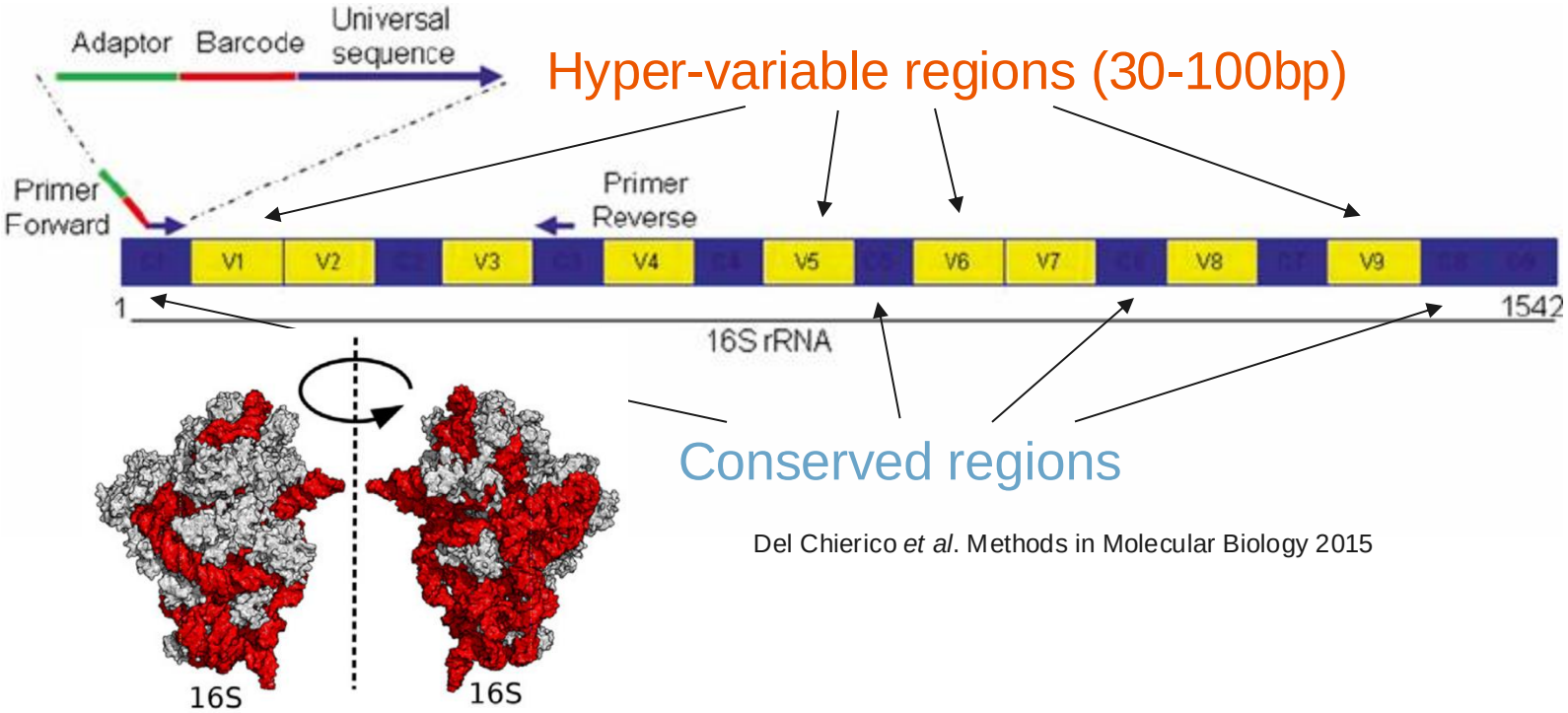
Functional view of taxonomy



Liu *et al.*, Marine Life Science & Technology 1: 112-128 (2019)

Taxonomy analysis via rRNA loci

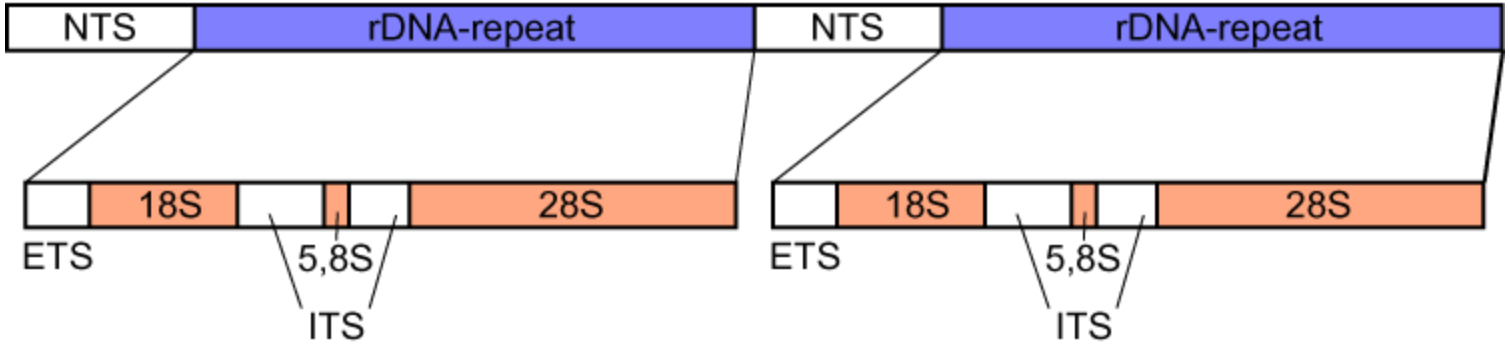
16S rRNA in prokaryotes



Del Chierico *et al*. Methods in Molecular Biology 2015

Ramazzotti and Bacci, Metagenomics: Perspectives, Methods, and Applications,103-119 (2018)

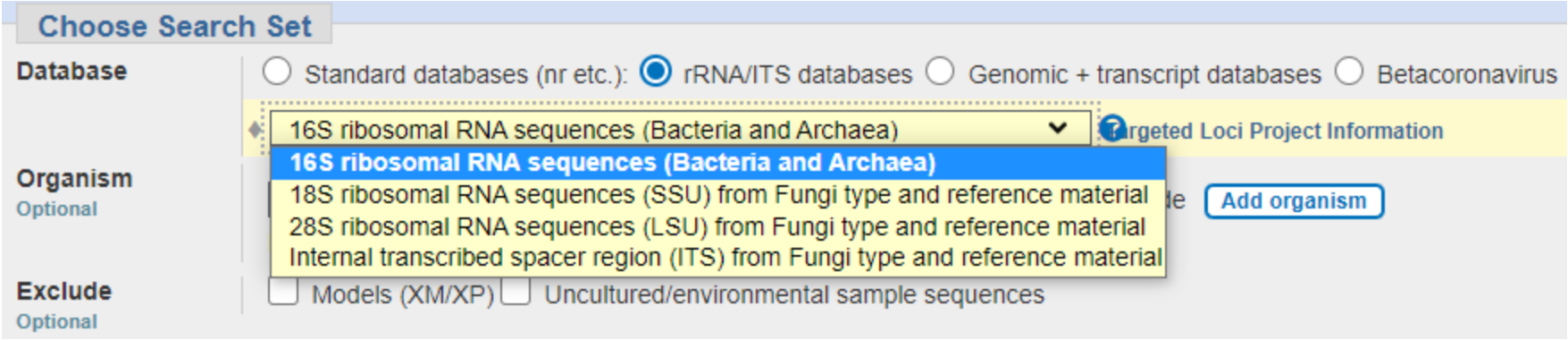
Internal transcribed spacer (ITS)



wikipedia.com

- Located between rRNA repeats
- ITS1 and ITS2
- 400-1000 bp
- Phylogenetics analysis of fungi and algae

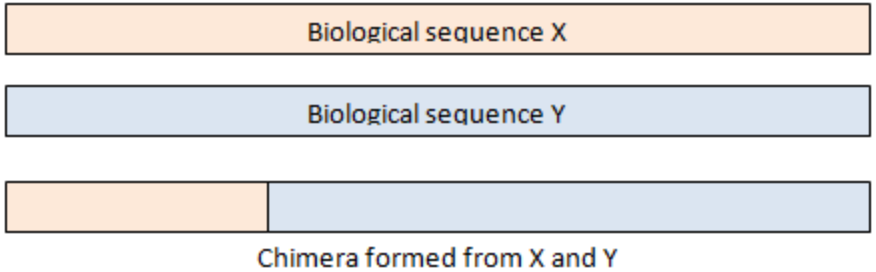
rRNA BLAST

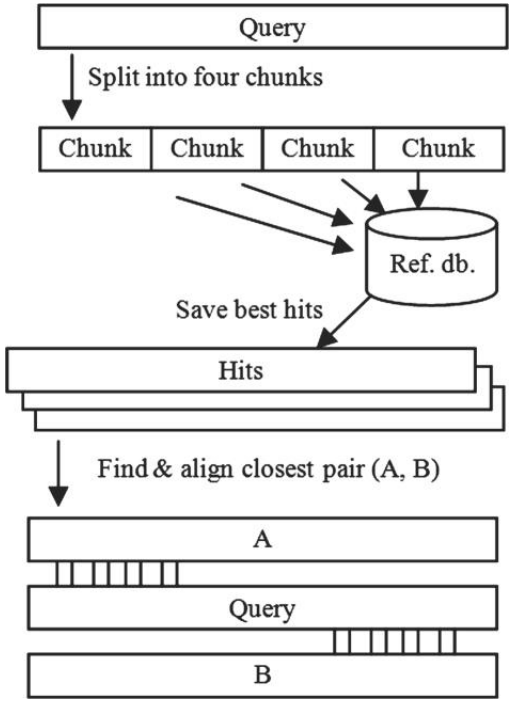


- Endpoint of rRNA amplicon analysis is taxonomic assignment
- Abundance profiles of taxa can be correlated to environment condition or disease status

Some analyses of rRNA data

Chimeric reads in amplicon sequencing

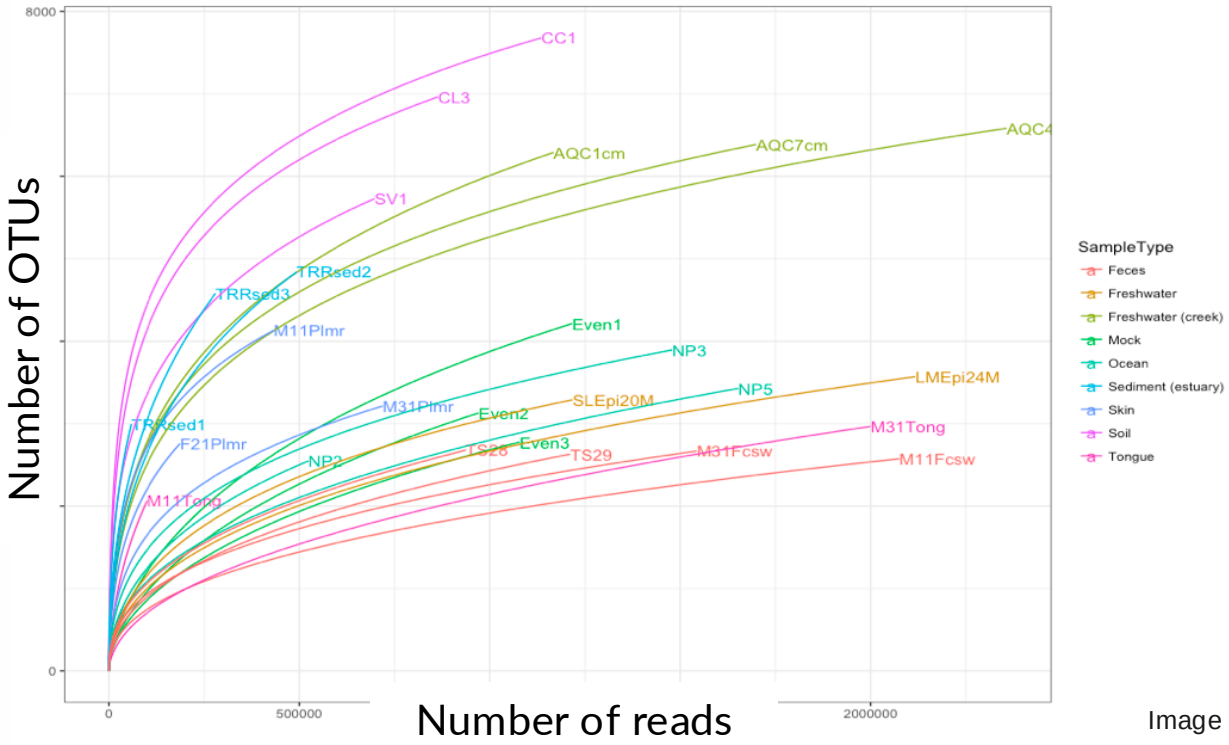




- Produced during PCR amplification
- Detected by alignment different portion of the reads to rRNA databases
- Mismatch of hits = chimeric reads

Edgar *et al.* Bioinformatics 27:2194-2200 (2011)

Rarefaction curve for evaluating depth of sequencing



- Subsample some reads and count the number of OTUs
- If depth is enough, not many addition OTUs will be found as the number of reads increases

Image from PhyloSeq GitHub

Complexity of microbiome composition

Diversity indices/ Parameters	Description	Formula
Shannon diversity index (H)	Estimator of species richness and species evenness: more weight on species richness	$H = -\sum_{i=1}^{s}(p_i \ln p_i)$ where s is the number of OTUs and p_i is the proportion of the community represented by OTU i.
Simpson's index (D)	Estimator of species richness and species evenness: more weight on species evenness	$D = \frac{1}{\sum_{i=1}^{s} p_i^2}$ where s is the total number of species in the community and p_i is the proportion of community represented by OTU i.

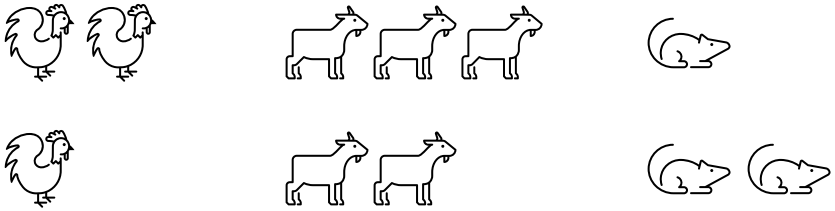
Kim, B.-R. et al. J Microbiol Biotechnol 27:2089-2093 (2017)

- Richness = number of distinct species
- Evenness = no dominant species

Comparing microbiome composition

Sample 1

Sample 2

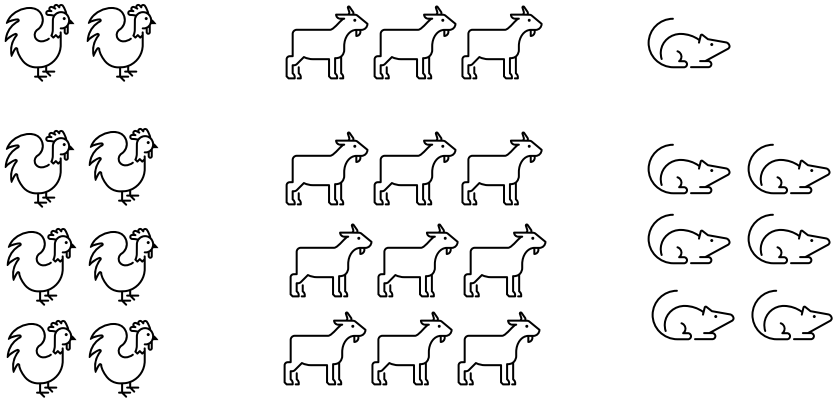


- S_1 = # of individuals in sample 1 = 6
- S_2 = # of individuals in sample 2 = 5
- Overlap = 1 + 2 + 1 = 4
- Bray-Curtis dissimilarity = $1 - \frac{2 \times \text{Overlap}}{S_1 + S_2} = 1 - \frac{8}{11} = \frac{3}{11}$

Impact of sequencing depth

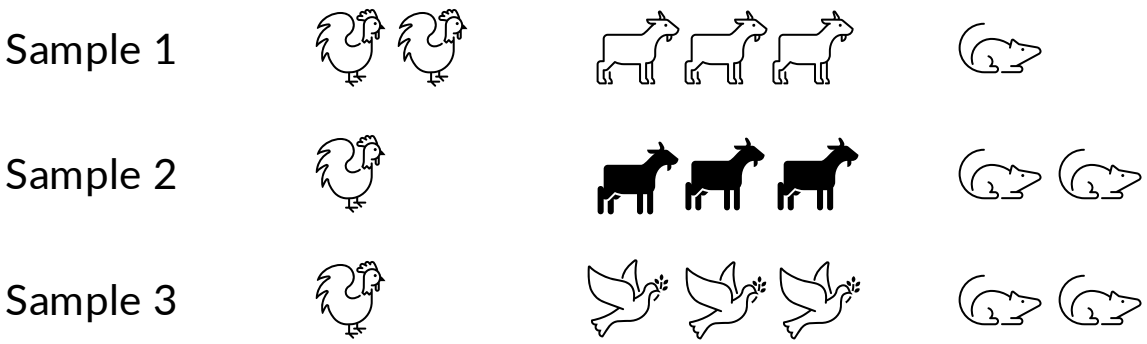
Sample 1

Sample 2



- Bray-Curtis is suitable between samples with similar sequencing depths

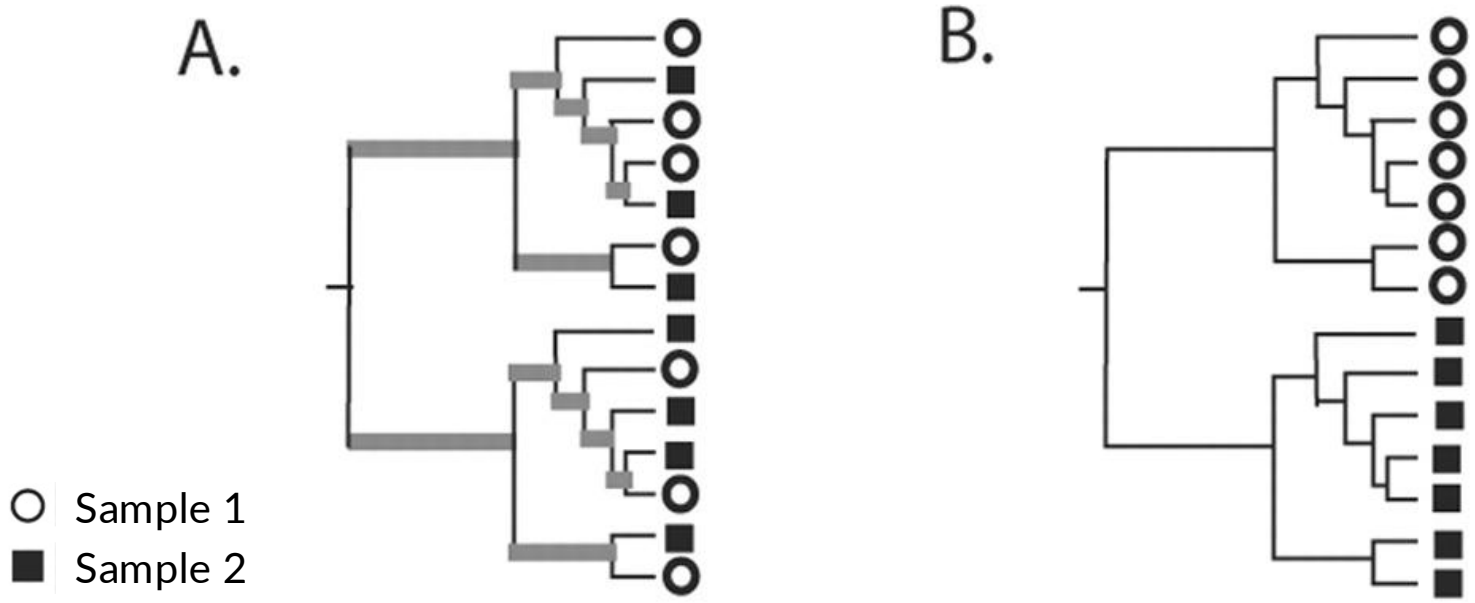
Impact of taxonomic similarity



- Bray-Curtis does not take into account taxonomic similarity

UniFrac distance

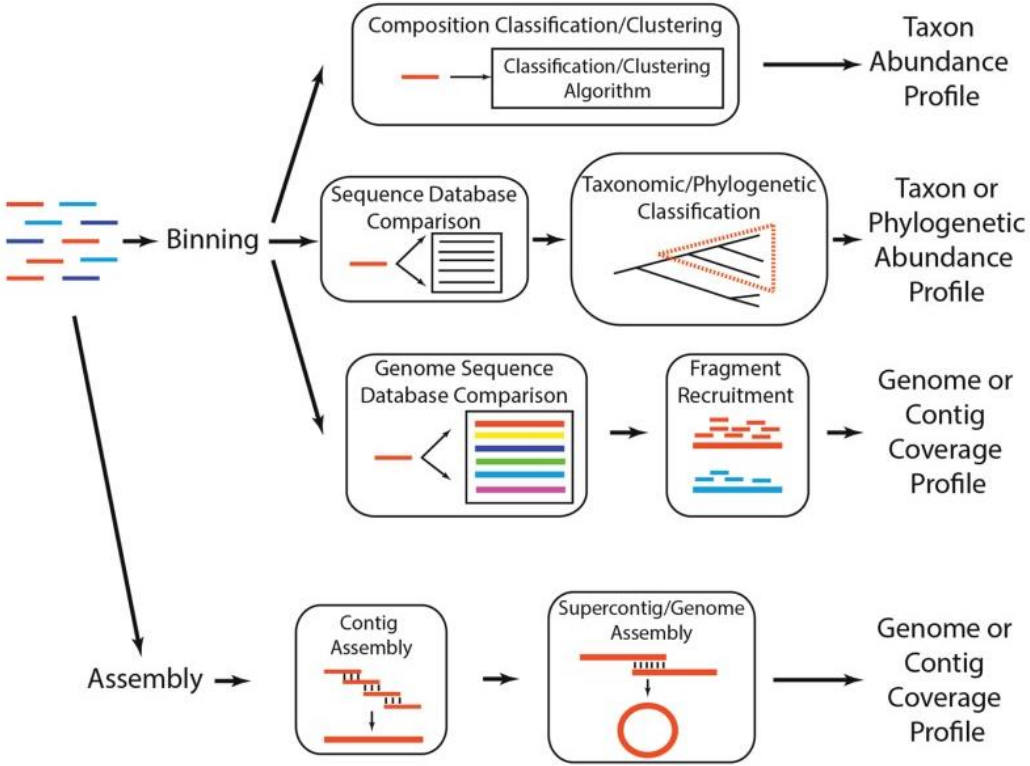
Lozupone, C. and Knight, R. Applied and Environmental Microbiology 71 (2005)



- UniFrac = fraction of shared phylogenetic branches between samples
- Can be weighted or unweighted by taxa abundances

Shotgun metagenomics

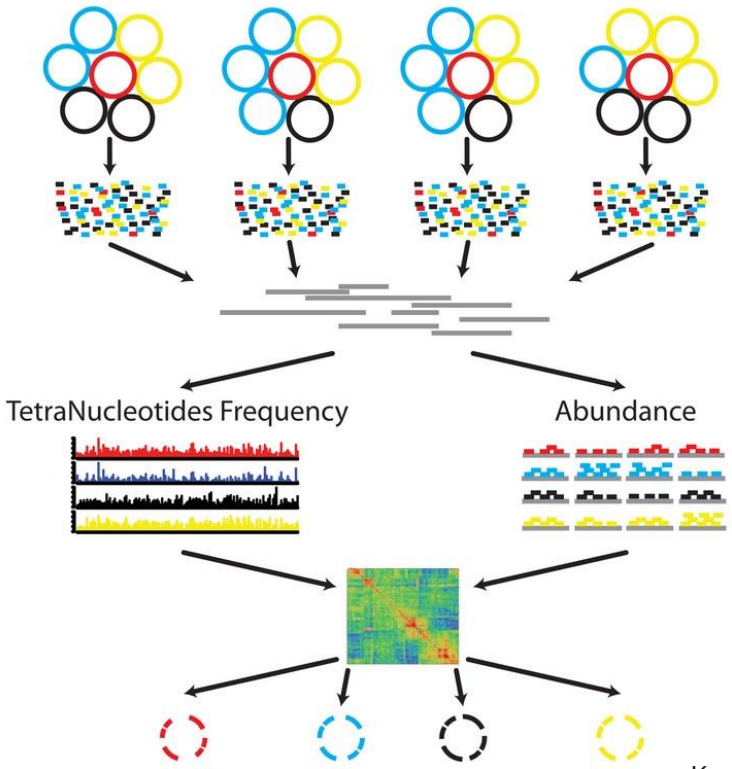
Key steps in shotgun metagenomics



- Dealing with contamination
  - Host DNA
- Binning = grouping DNA/RNA from the same host organisms together
- Direct assembly is possible for abundant species

Sharpton *et al*. Front. Plant Sci. 5:209 (2014)

Read binning strategies

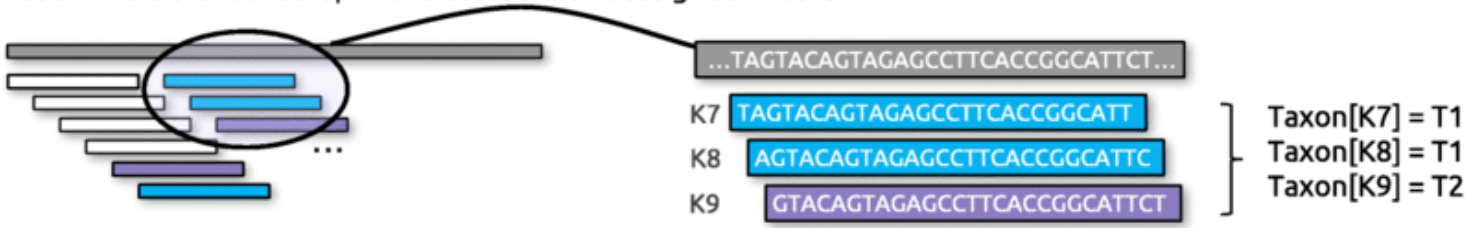


- Reads originated from the same species should have similar *k*-mer profile
- Pairs of reads originated from the same species should have highly correlated abundances across samples (because their abundances correlated with species abundance)

Kang *et al.* PeerJ (2015)

k-mer matching to predict taxonomy

A Read k-mers are looked-up in the database and assigned to taxa:



C K-mer count and coverage in taxonomic report show evidence behind classifications:

reads	kmers	dup	cov	taxID	rank	name
122	112	144	0.0004	11855	species	*Clostridioides difficile*
9650	7129	74.5	0.192	10632	species	Human polyomavirus 2
15	1570	1	0.0002	7643	species group	*Mycobacterium tb* complex

Bad classification with few k-mers

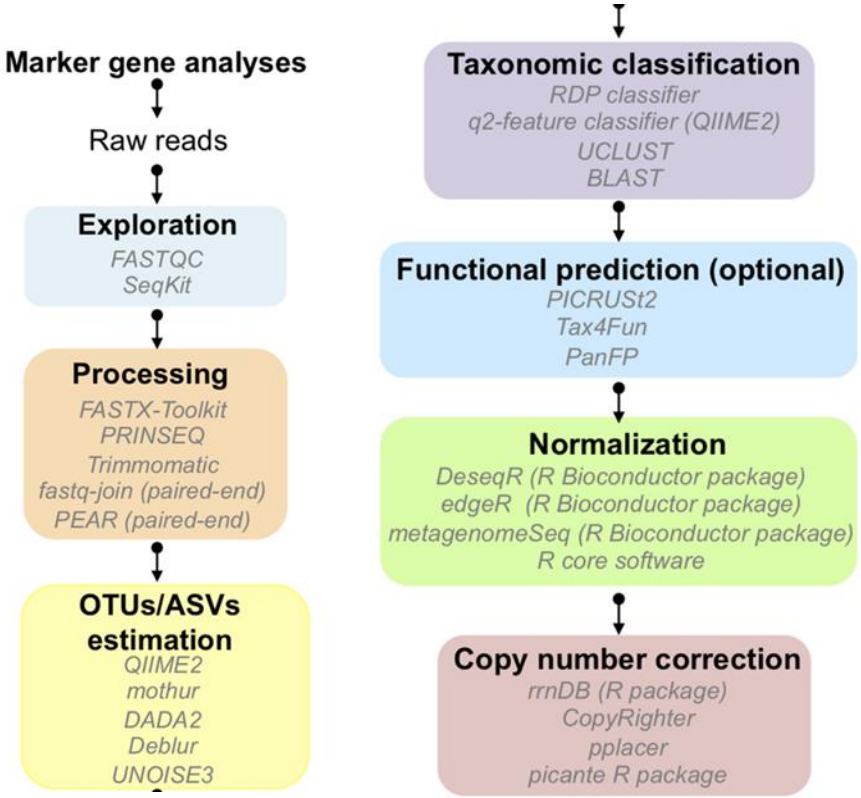
Good classification, reads cover genome

Number of distinct k-mers for taxon, and coverage of the taxon's k-mers

Breitwieser, F. et al. Genome Biology 19:198 (2018)

Metagenomics workflows

Amplicon analysis pipeline (16S rRNA)



k-mer, machine learning, or alignment

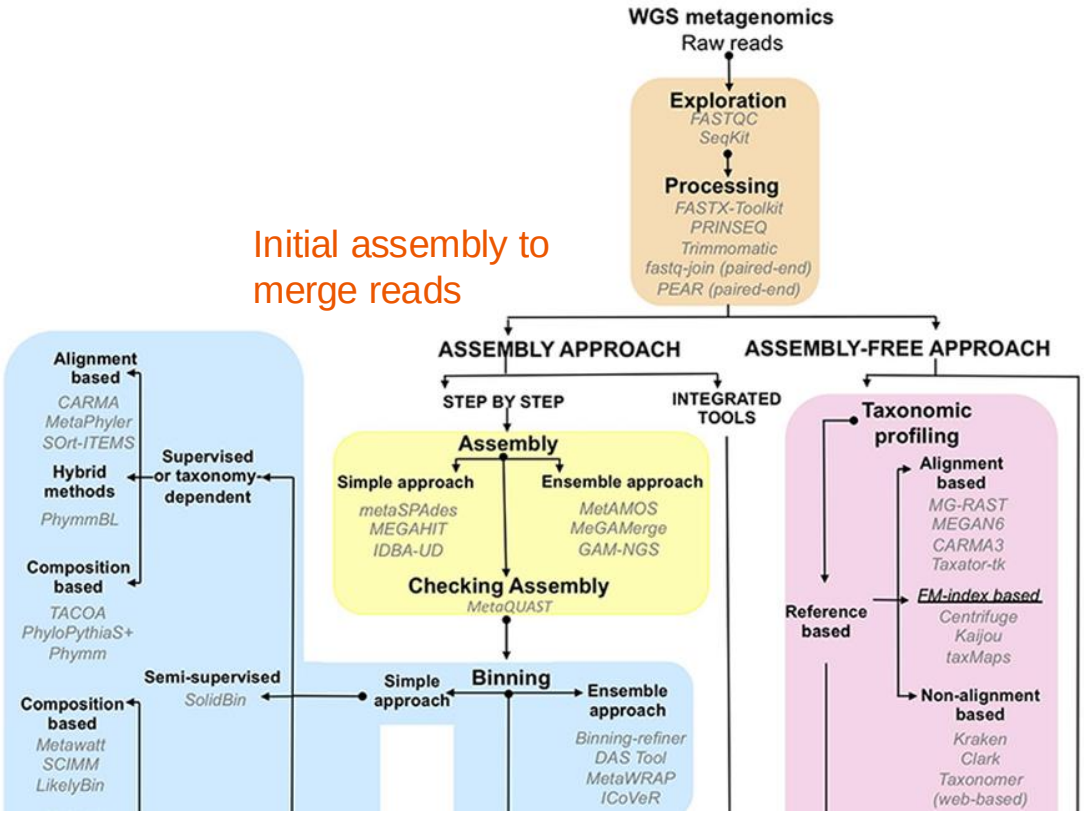
Remove noises, reads with high error, and chimera

Shotgun analysis pipeline

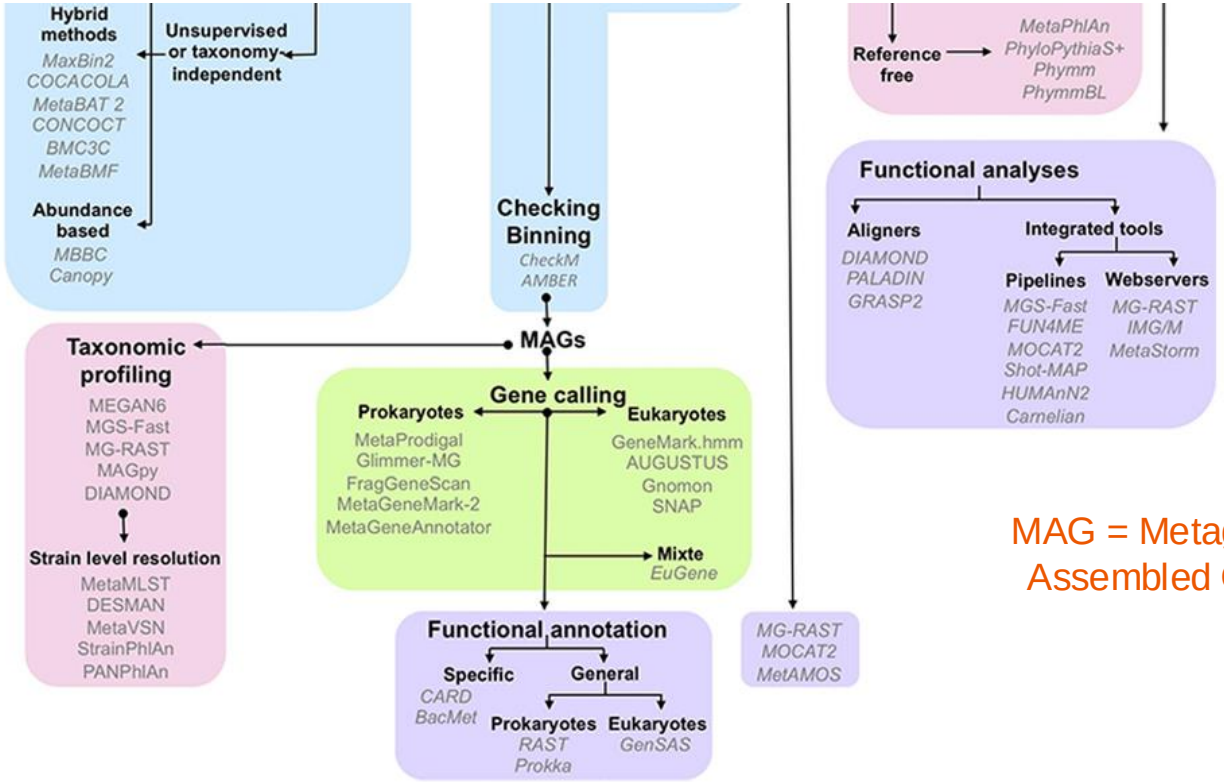
Initial assembly to merge reads

Direct taxonomic annotation

Read binning



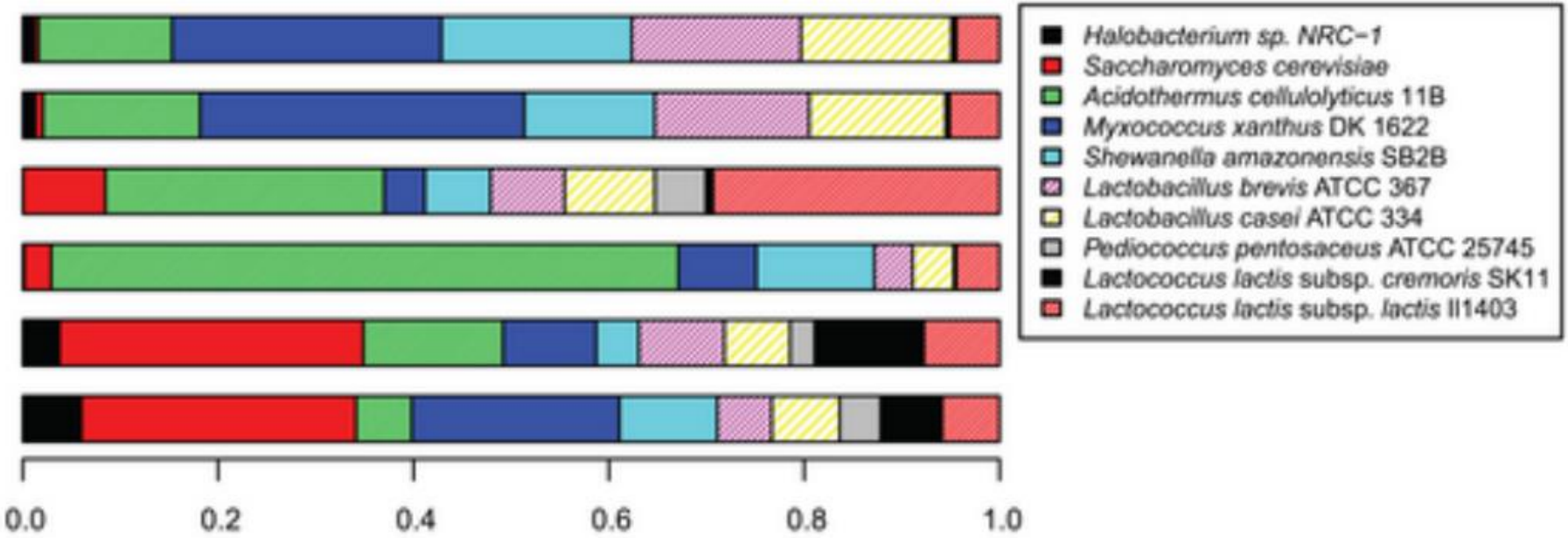
Shotgun analysis pipeline



MAG = Metagenomics Assembled Genome

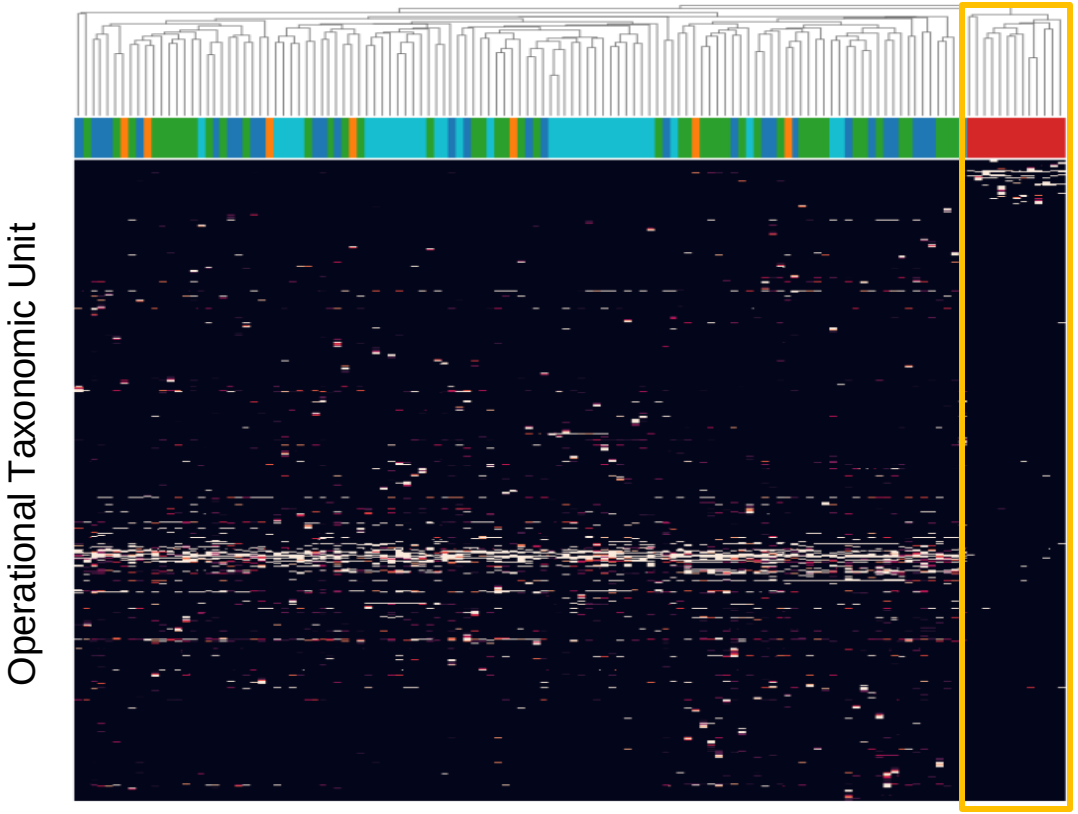
Variability in microbiome analysis

Same sample, different profiles



Source: http://www.cbs.dtu.dk/courses/27626

Batch effects



Red = same batch

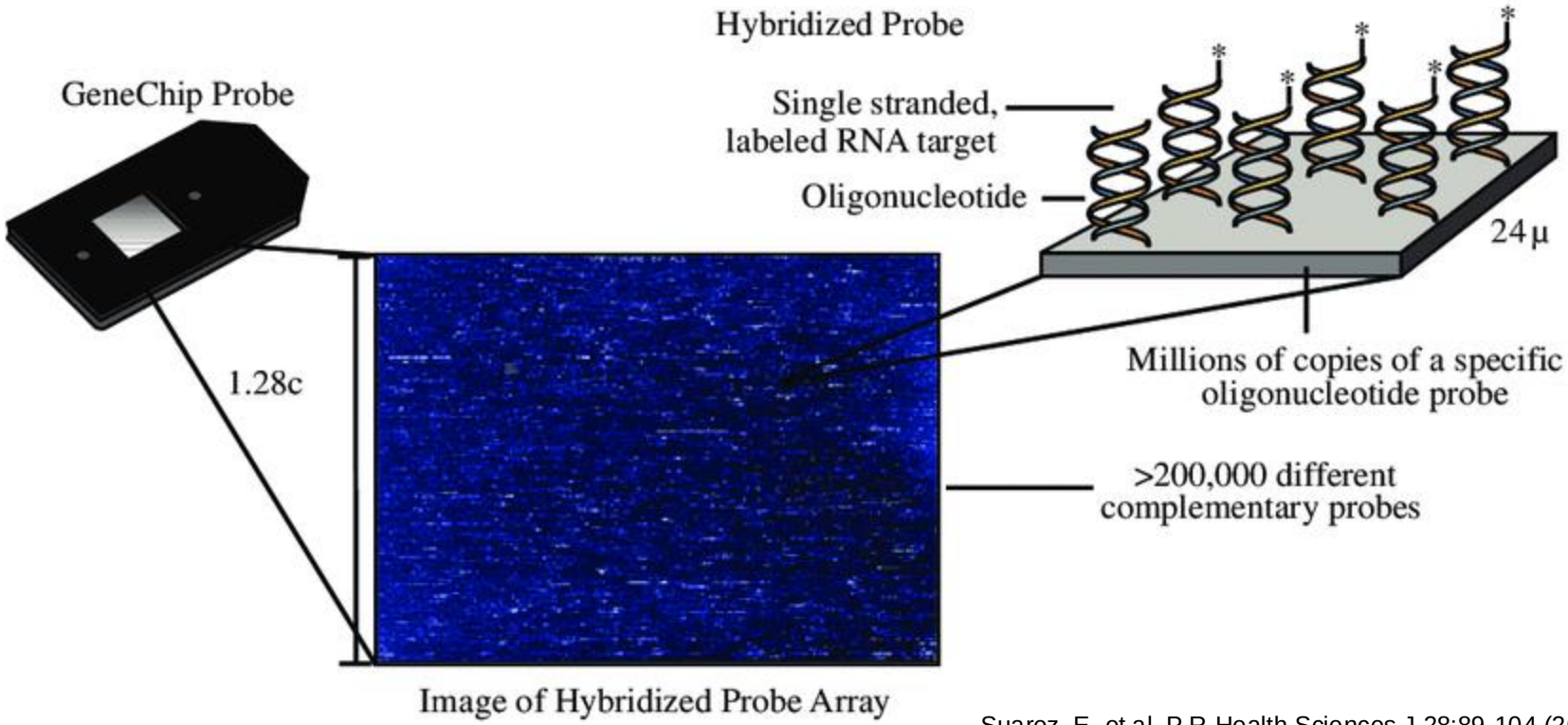
Any question?

Part II: Microarray and Nanostring

- Targeted transcriptomics
- Cheap and scalable
- A good illustration of how to apply statistics to biological data

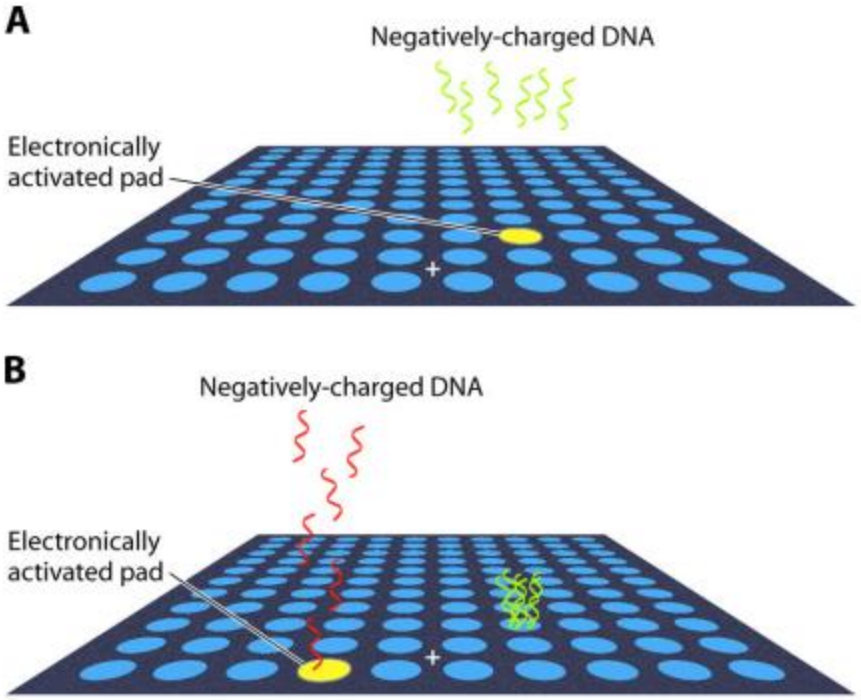
Oligonucleotide microarray

Microarray technology overview

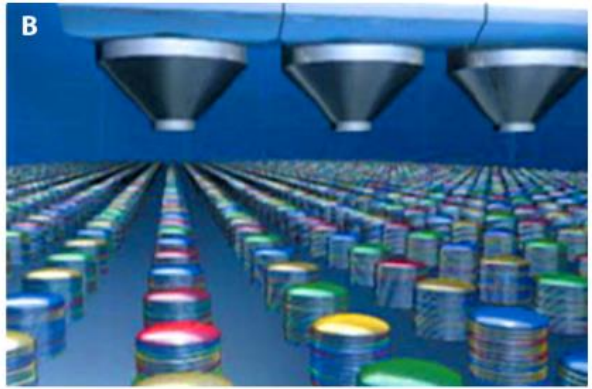


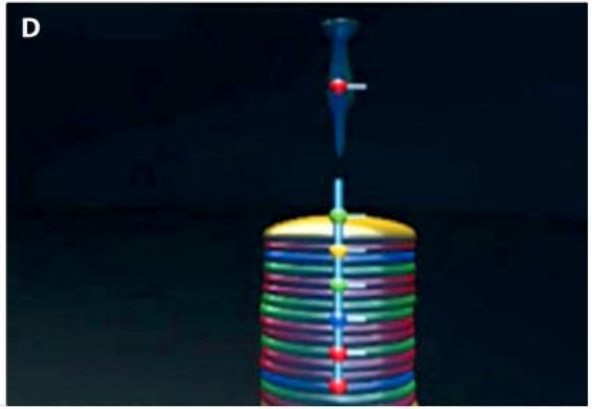
Suarez, E. et al. P R Health Sciences J 28:89-104 (2009)

Microarray fabrication

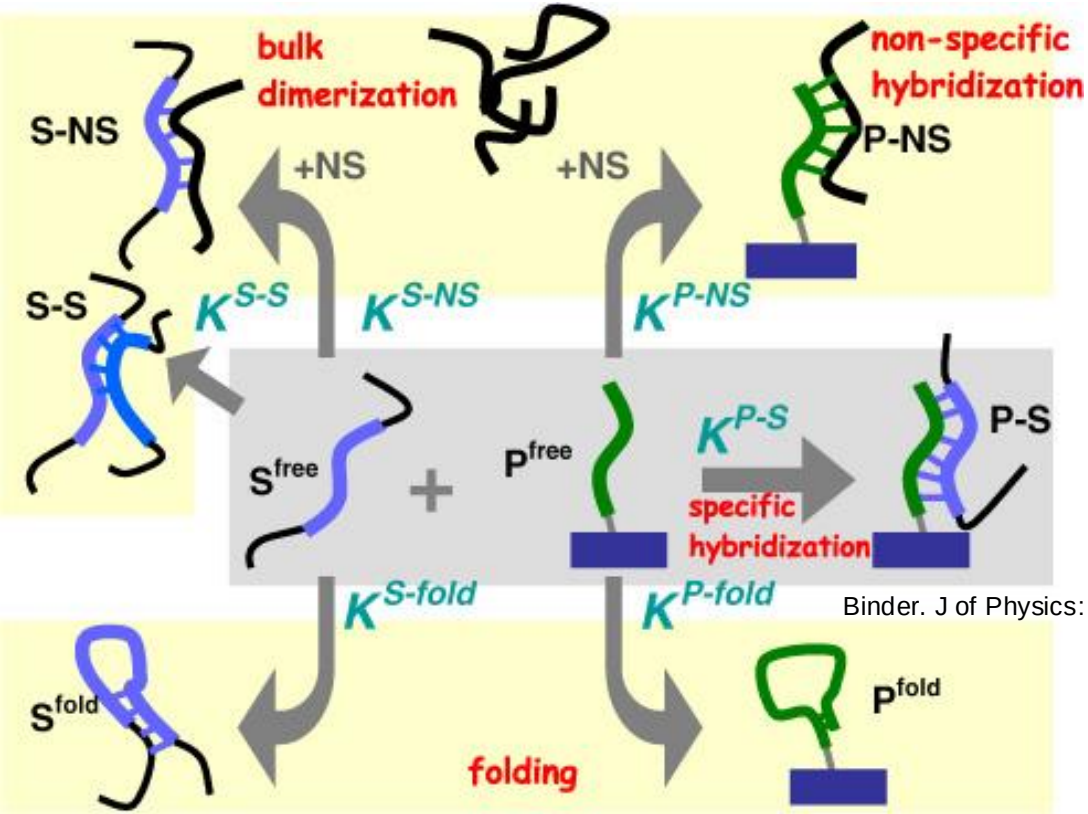


Miller, M.B. and Tang, Y.-W. Clin Microbiol Rev 611-633 (2009)



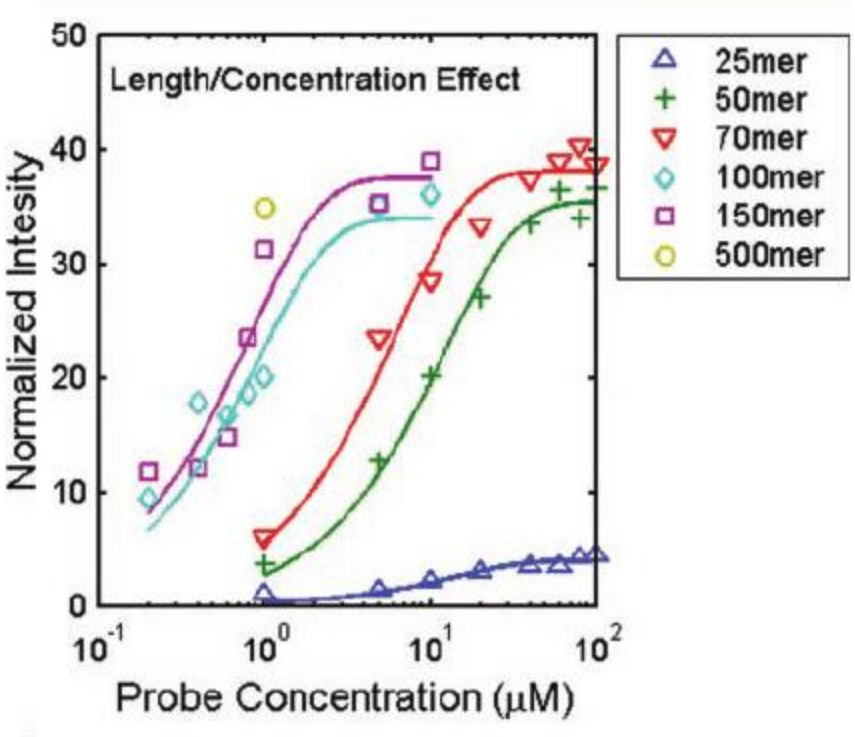


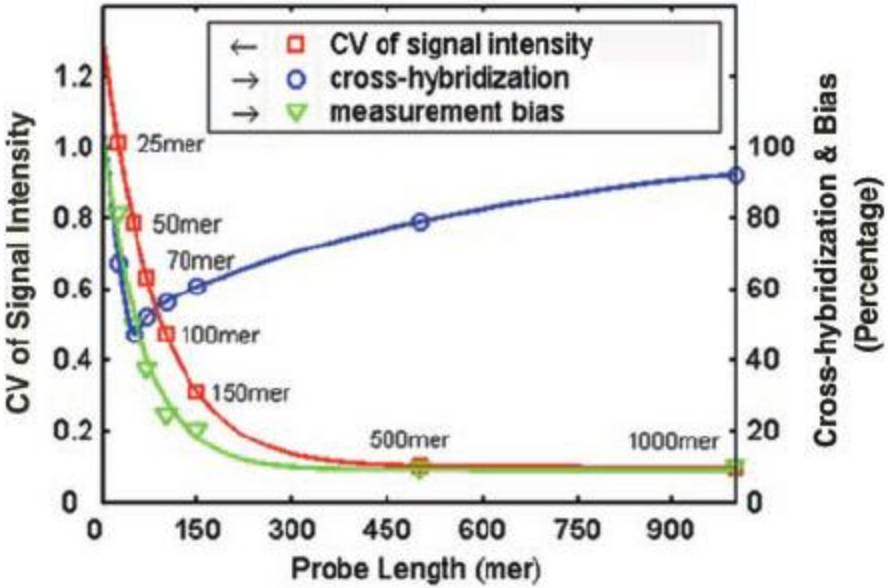
Unwanted probe interactions



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

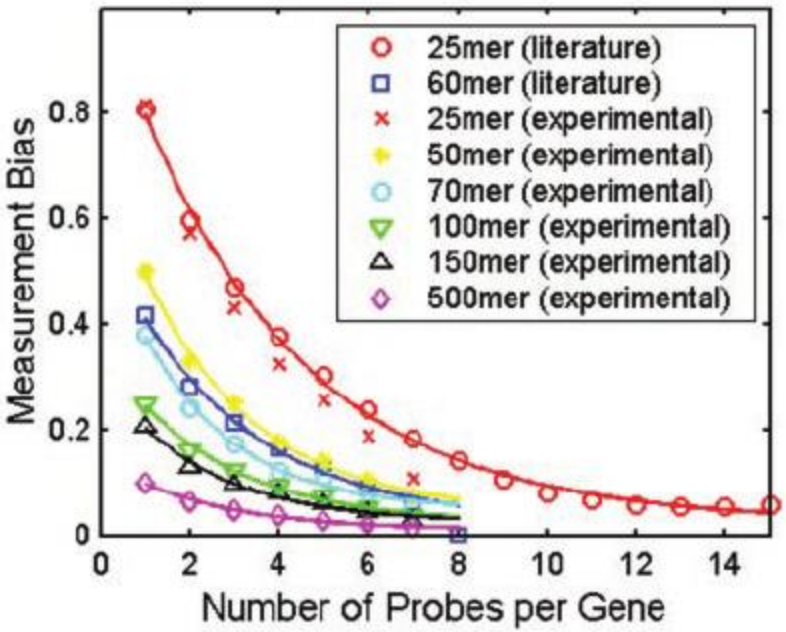
Impact of probe length

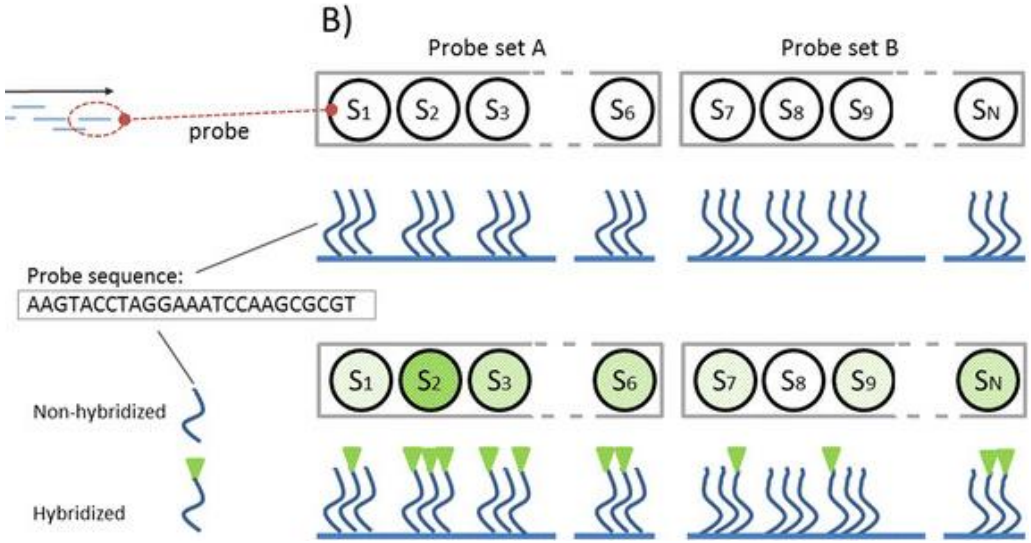




Chou, C.-C. NAR 32:e99 (2004)

Probe set = multiple probes per gene

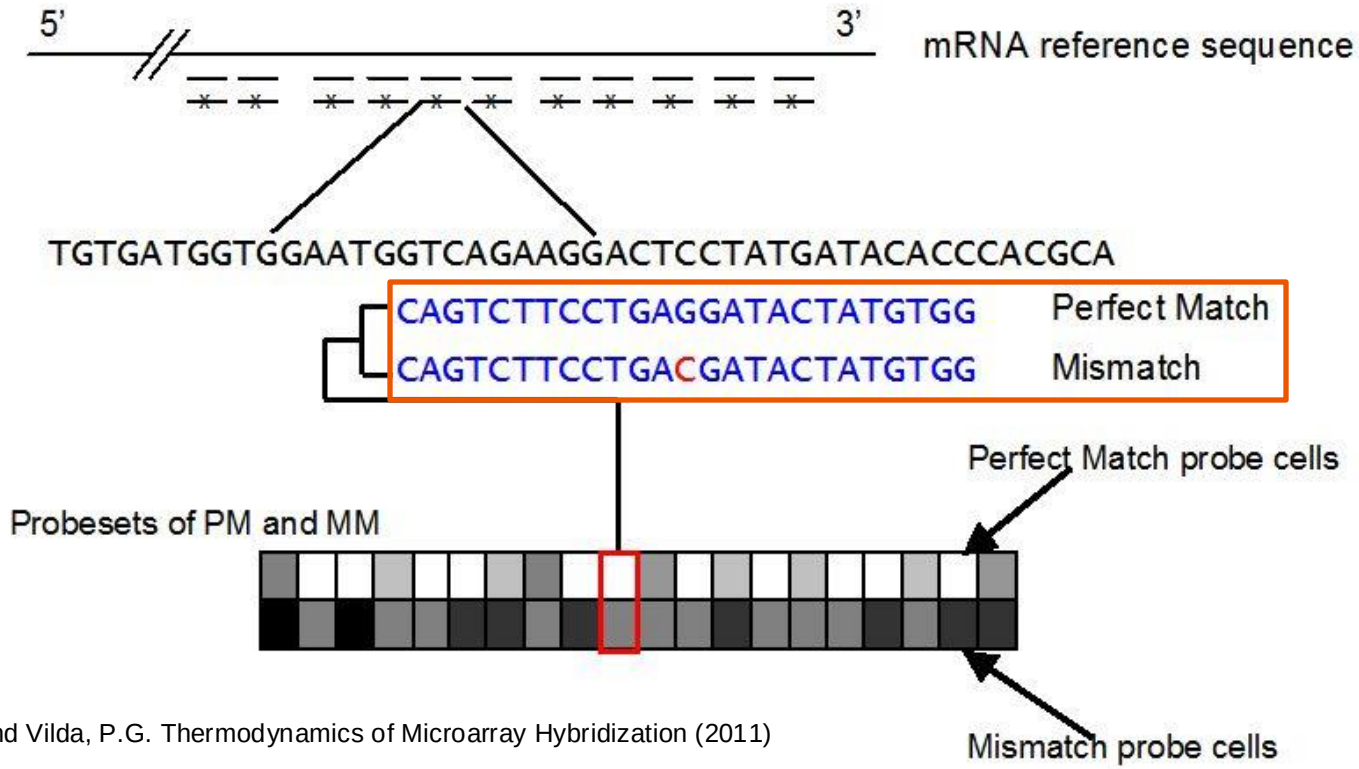




Chou, C.-C. NAR 32:e99 (2004)

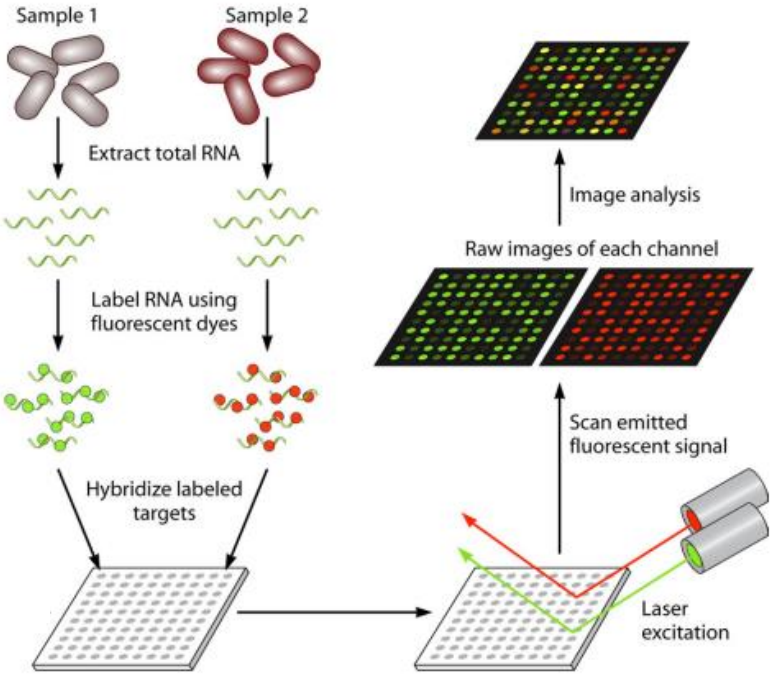
Jaksik, R. et al. Biology Direct 10:46 (2015)

Perfect match (PM) and mismatch (MM)



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

Multi-channel microarray



- 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

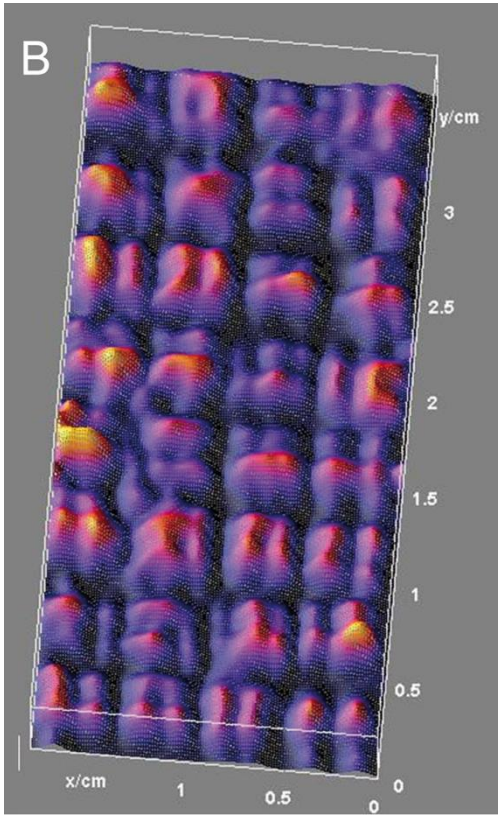
Miller, M.B. and Tang, Y.-W. Clin Microbiol Rev 611-633 (2009)

Key processing steps

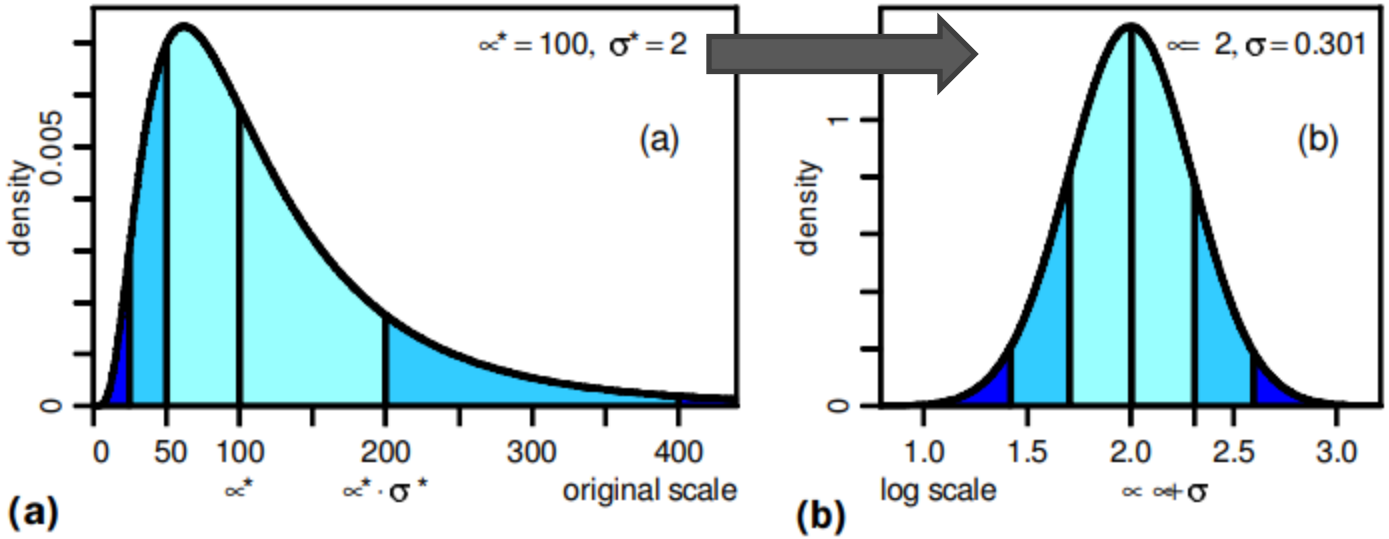
- Redefining probe set
  - BLAST to latest genome annotation

- Intensity correction
  - Model background using probe location & sequence
  - Perfect match (PM) vs mismatch (MM)
  - Global & local correction
- Outlier removal

- Probe set aggregation
- Log transform

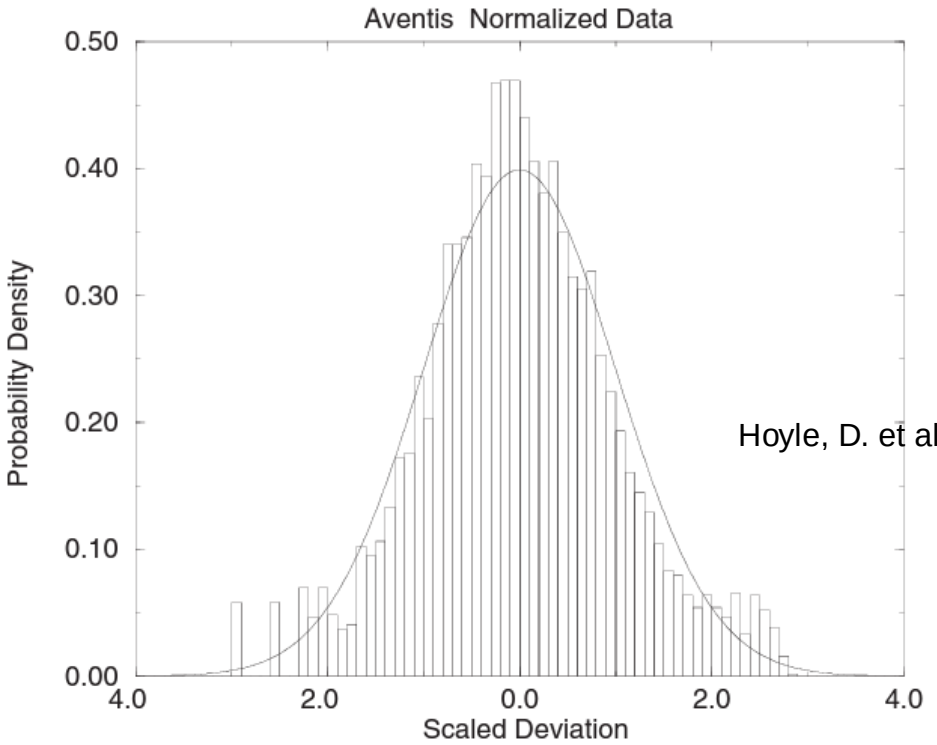


Log-normal distribution



Limpert, Stahel, and Abbt. BioScience 2001.

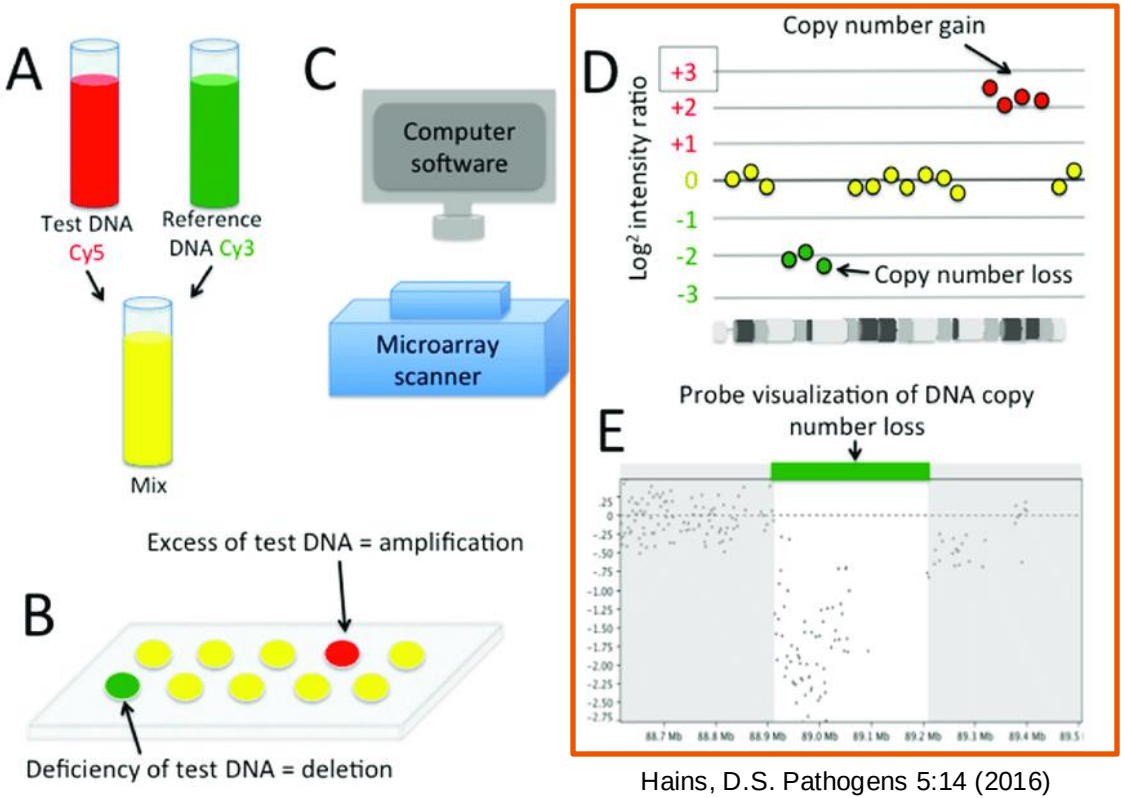
Microarray data are log-normal distributed



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

Beyond transcriptomics

Comparative genome hybridization (CGH)

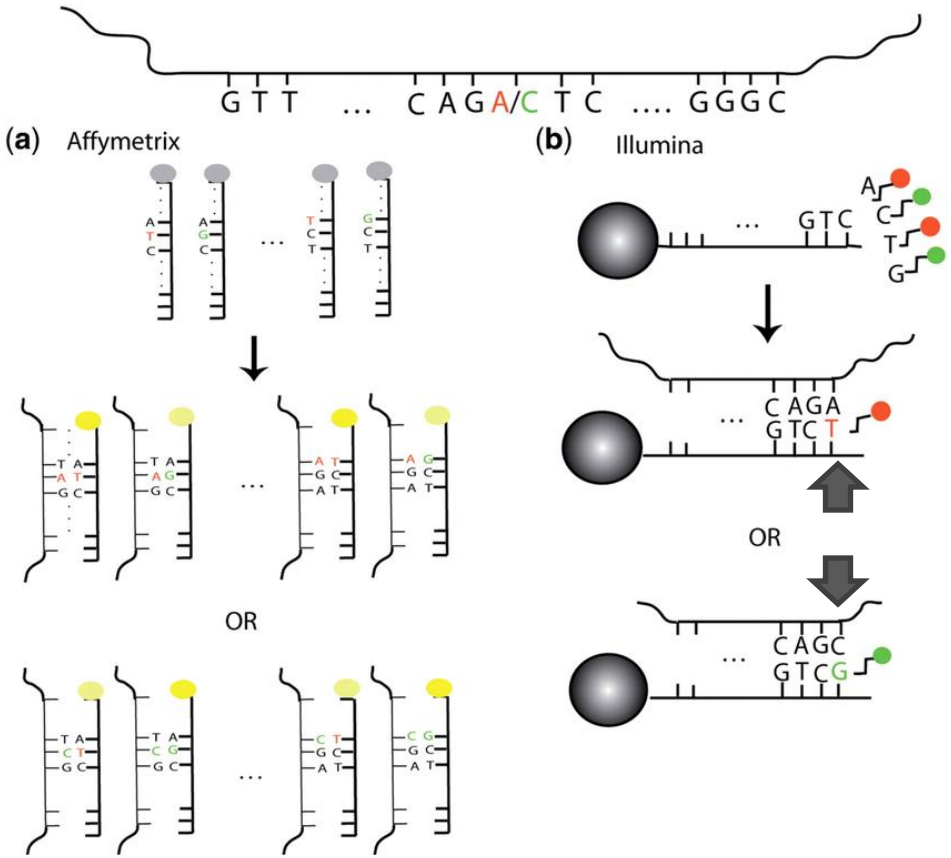


Hains, D.S. Pathogens 5:14 (2016)

SNP genotyping array

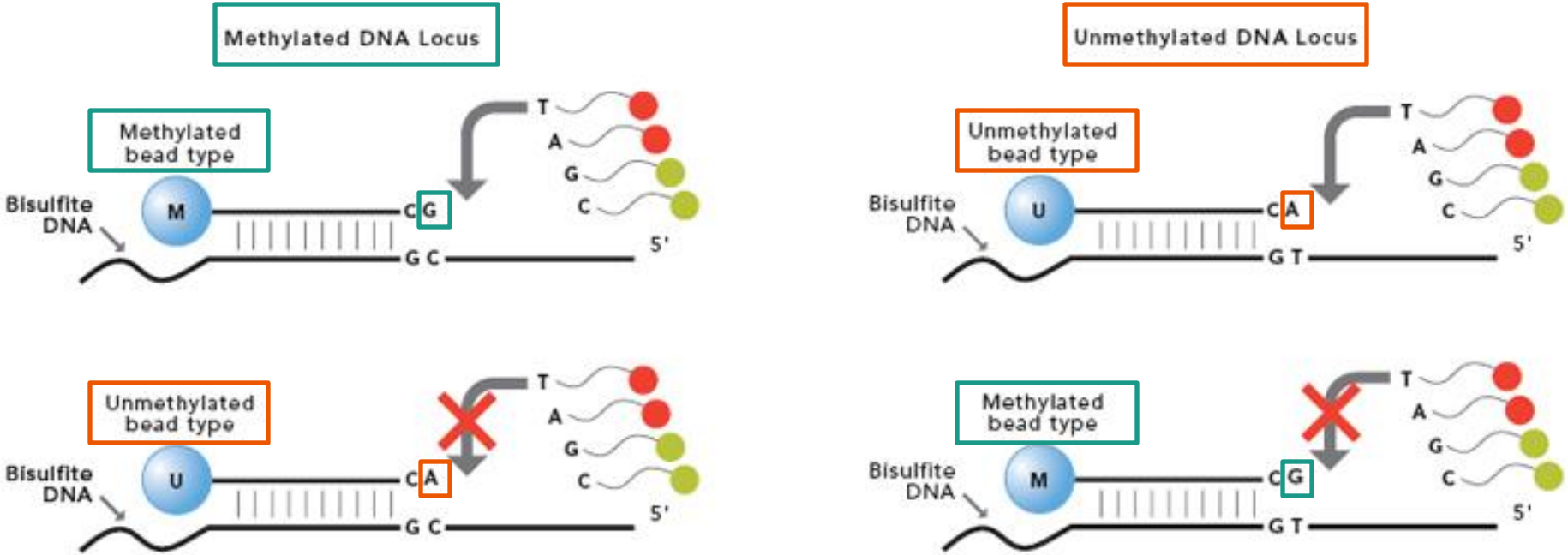
- Design probes for alternative SNPs at each position
  - Relative hybridization

- Single-nucleotide sequencing
  - Probe acts as primer
  - Match to the position up until right before the SNP
  - Incorporation of the next nucleotide determine the genotype



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

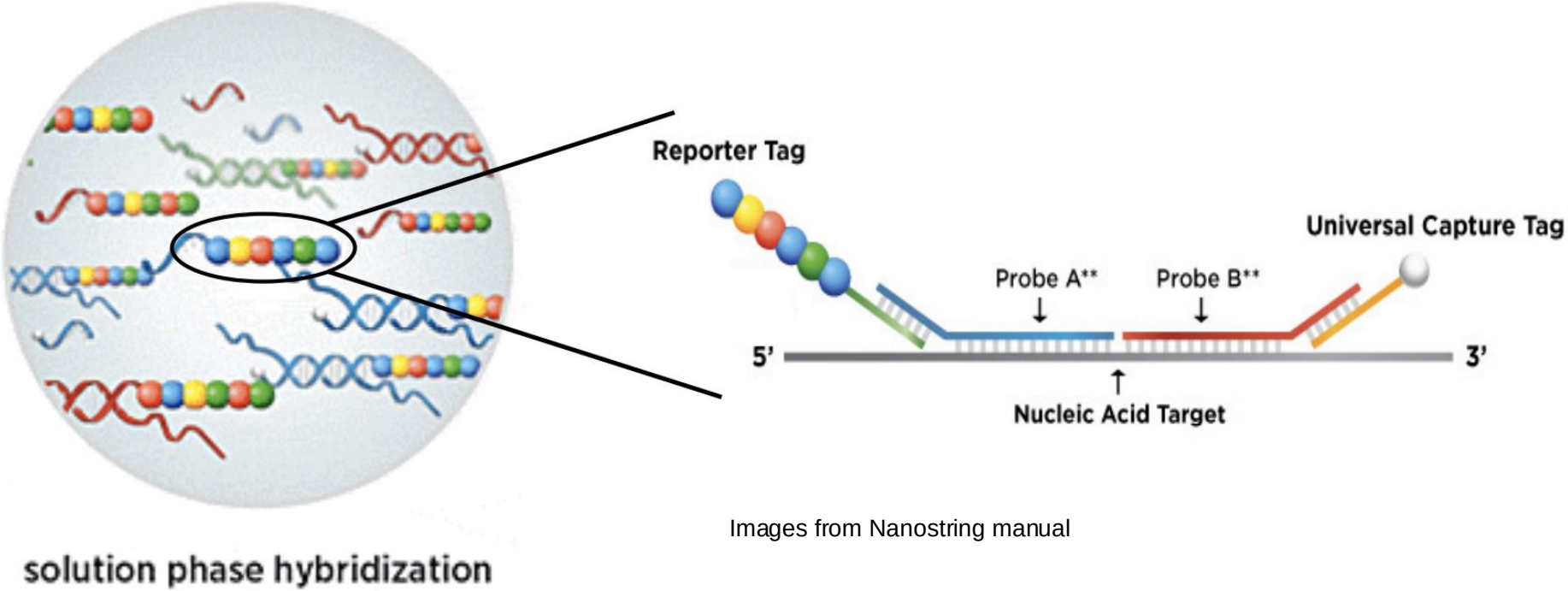
Methylation array



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

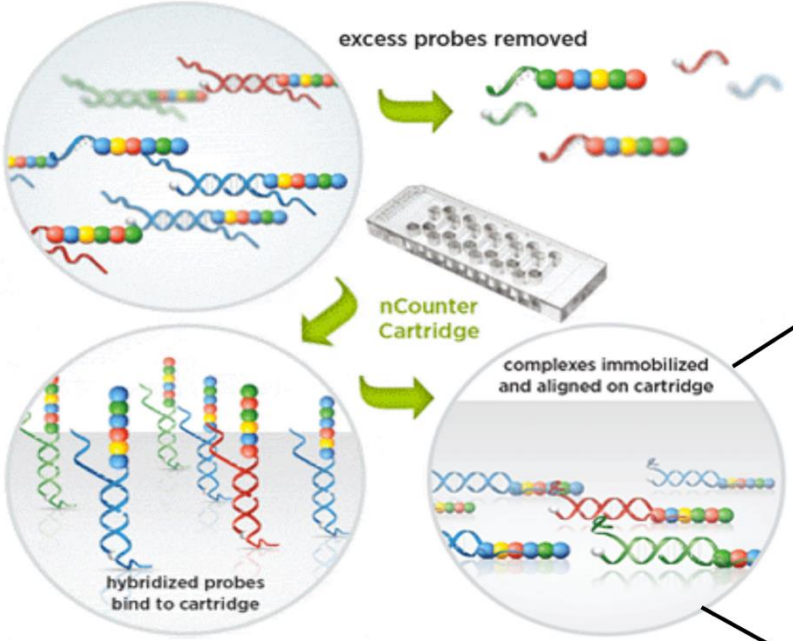
Nanostring

Transcript-specific probes & fluorescence barcodes

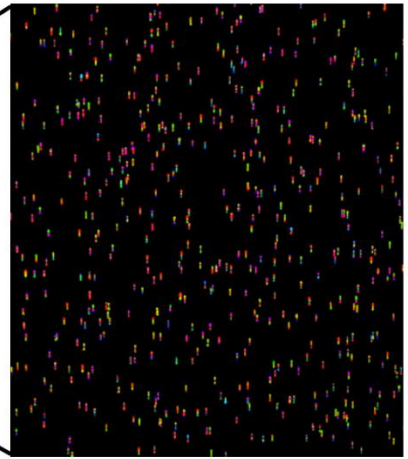


Images from Nanostring manual

Counting number of molecules



Barcode	Counts	Identity
	3	XLSA
	2	FOX5
	1	INSULIN



Images from Nanostring manual

Prebuilt barcode 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

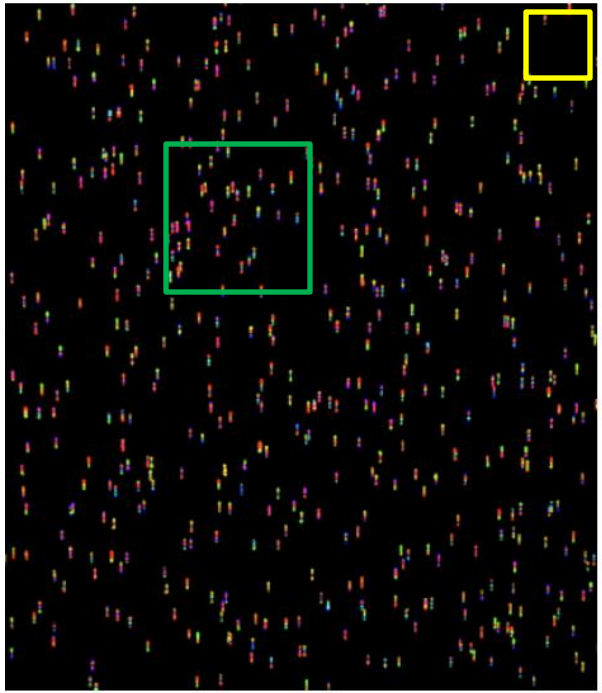
Canine IO

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

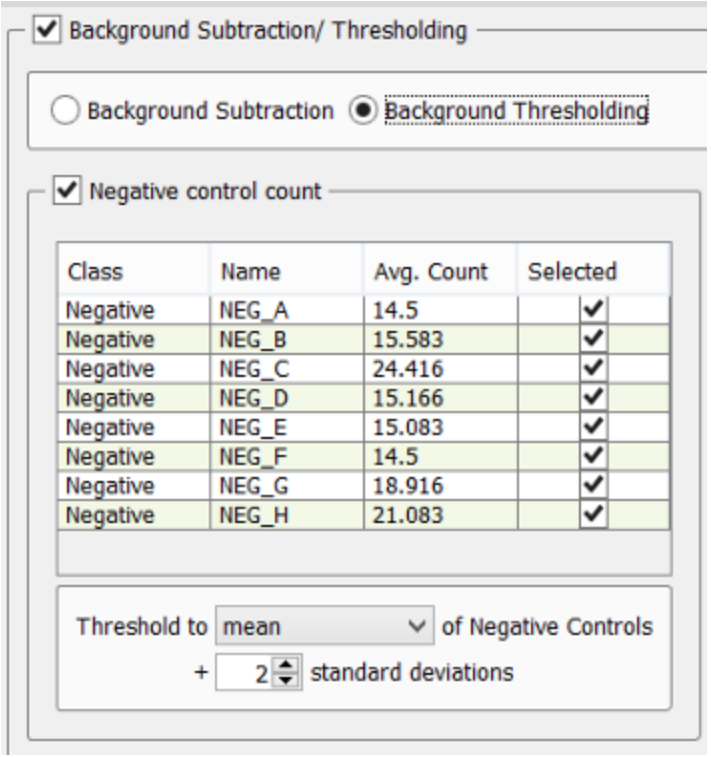
Nanostring quality control



- 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

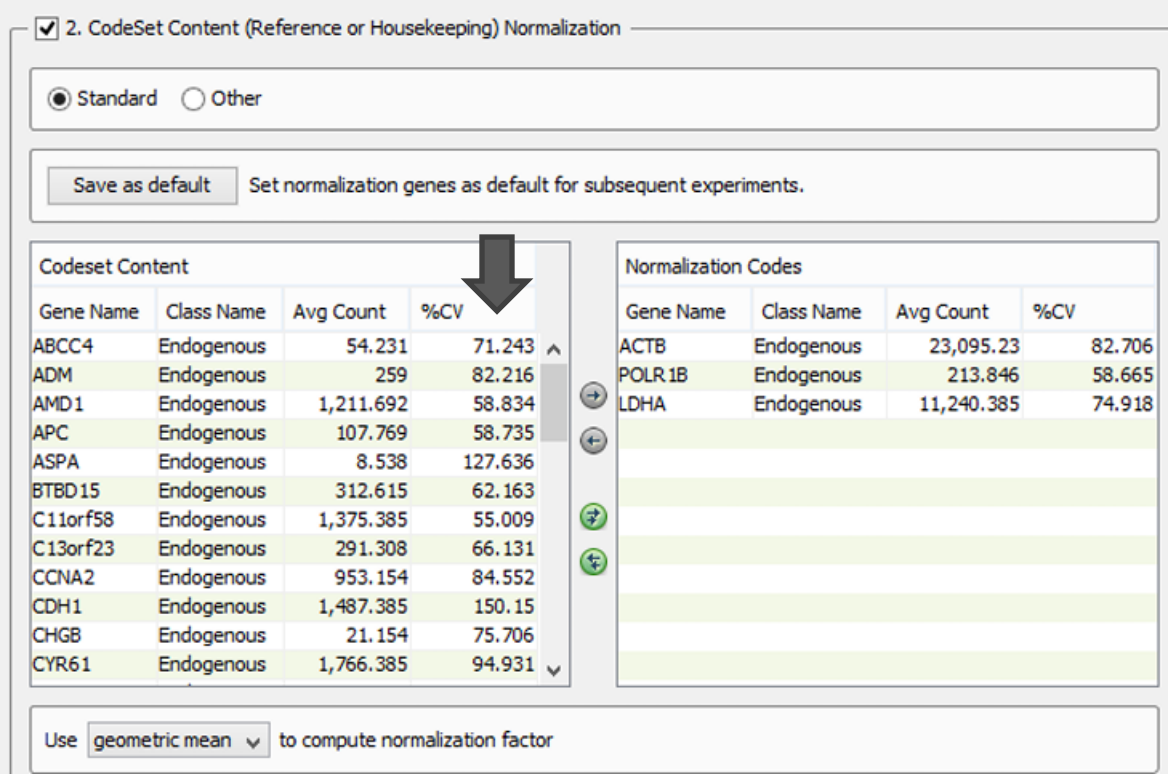
Nanostring data preprocessing

Negative and positive control

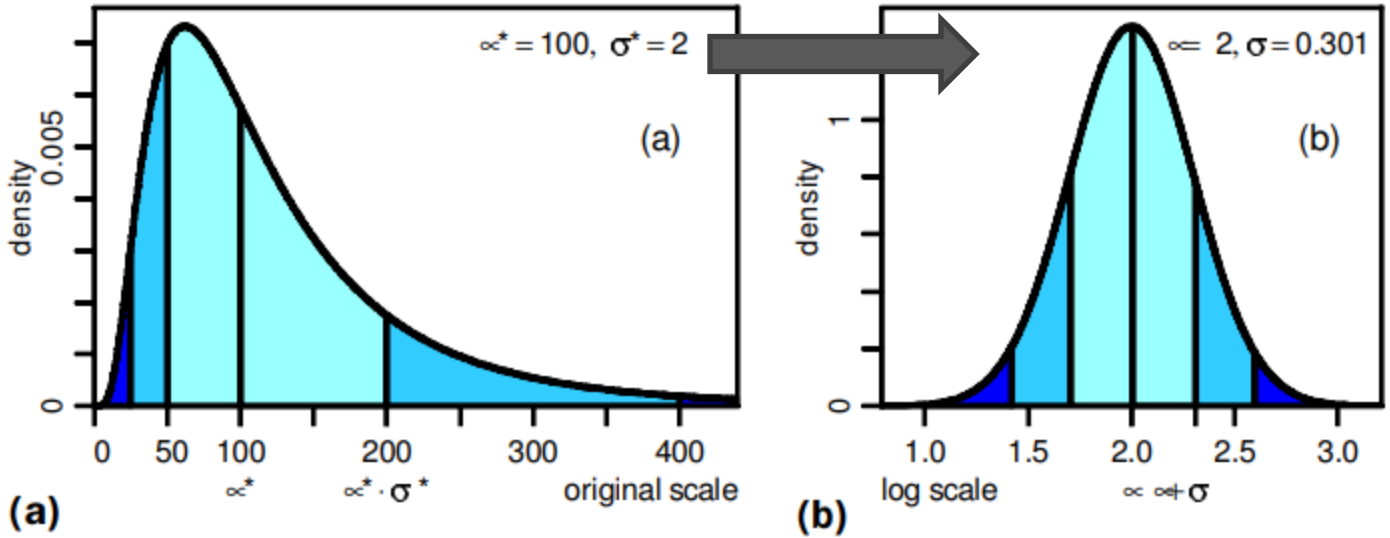


Raw Data					
			Sample 1	Sample 2	Sample 3
Positive	POS_A	ERCC_00117.1	24573	21007	21856
Positive	POS_B	ERCC_00112.1	6948	6414	6589
Positive	POS_C	ERCC_00002.1	2123	1826	1932
Positive	POS_D	ERCC_00092.1	432	363	425
Positive	POS_E	ERCC_00035.1	52	68	53
Positive	POS_F	ERCC_00034.1	49	38	52
		Geomean of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

Housekeeping control



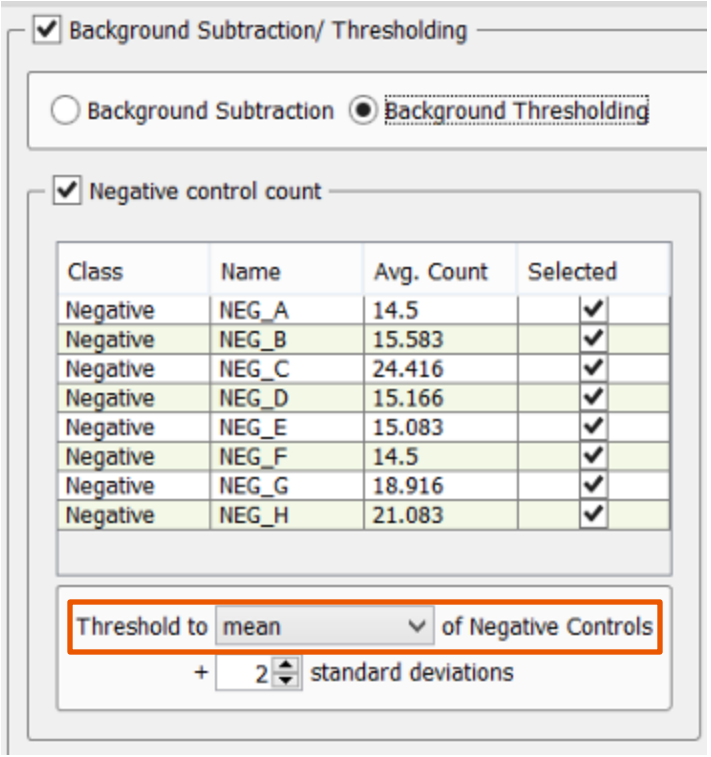
Arithmetic Mean vs Geometric Mean



Limpert, Stahel, and Abbt. BioScience 2001.

- $\frac{\log(x)+\log(y)}{2} = \log\left(\sqrt{xy}\right)$ → AM of log-transformed = GM of original data

Arithmetic mean of background noises



- Background noises are Normal
- Arithmetic Mean is ok

Geometric mean of positive controls

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 closer to Log-Normal
- Use GM to represent AM of log-transformed data
- Positive controls with known concentration serve as correction factor

Simple transcriptomics analysis

Transformed data can be analyzed with *t*-test

Control Treatment

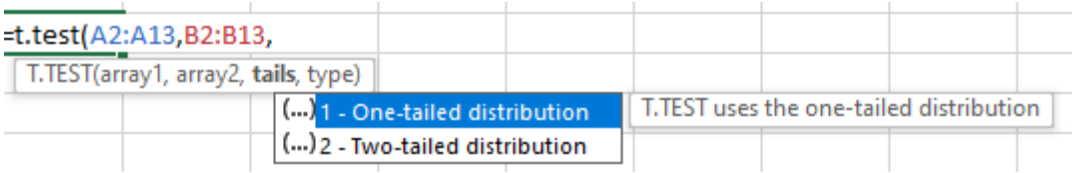
J15

	A	B	C	D	E	F	G	H
1	Acc ID	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	
2	NM_007818	67540.89	70924.09	80243.76	3501.2	5697.47	2426.72	
3	NM_001105160	811.93	801.36	740.71	128.67	104.42	101.33	
4	NM_028089	190.41	211.06	236.19	9.05	23.33	8.44	
5	NM_016696	66.77	57.56	101.09	750.9	659.84	491.89	
6	NM_013459	3.3	11.29	1.89	735.82	816.46	118.22	
7	NM_007809	45.34	36.12	51.02	245.27	372.13	335.67	
8	NM_009999	103.04	370.21	200.29	17.09	13.33	8.44	
9	NM_133960	7708.78	6976.38	6569.04	1731	1641.81	1853.55	
10	NM_027881	31.32	10.16	24.56	268.39	186.62	135.11	
11	NM_054053	31.32	24.83	19.84	323.68	428.78	116.11	
12	NM_007377	47.81	89.17	70.86	370.93	378.79	279.72	
13	NM_028064	703.95	689.62	662.29	214.11	168.85	144.61	
14	NM_008182	222.56	339.73	226.75	30.16	63.32	26.39	
15	NM_013661	12.36	11.29	8.5	97.51	77.76	71.78	
16	NM_007815	20613.09	25218.13	31540.46	5209.07	7680.3	6312.2	

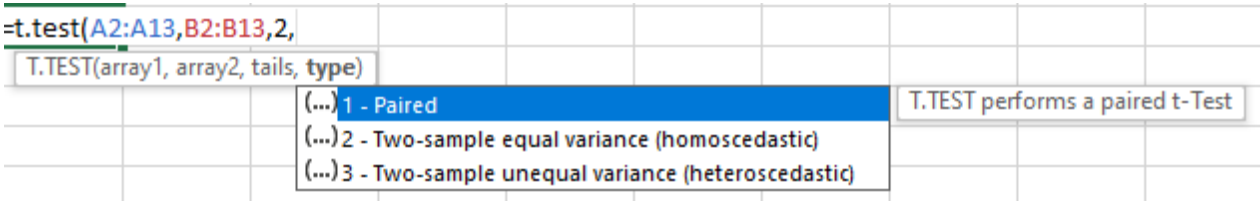
- Log transform
- Perform *t*-test on each gene
- Correct the p-values for multiple testing

http://homer.ucsd.edu/homer/basicTutorial/clustering.html

Choosing the right *t*-test



- Two-tailed tests whether the expression is higher or lower



- Use paired only for before & after treatment data of the same sample
- Otherwise, assuming unequal variance (Welch) is safer

Correction with Bonferroni method

- Divide the p-value cutoff by the number of test
- Adjusted p-value cutoff = 0.05 / 1000 = 0.00005

- Applying similar test 1,000 times will result in 0.05 tests on average with smaller p-value than 0.00005 just by chance

- Easy to calculate but lose power

False discovery rate (FDR) vs p-value

- **P-value** = probability of observing the same or more extreme (higher fold change) under the null hypothesis (that there is no differential expression)

- **False Discovery Rate** = probability that a detected differentially expressed gene was not differentially expressed

- P-value is easy to calculate (because it assumes no differential expression)
- But FDR involves alternative hypothesis

- There are ways to control FDR through p-value!

Benjamini-Hochberg procedure

- Valid under broad assumptions (independent tests, positively correlated tests, etc.)

- Given a series of tests with p-values, $p_1, p_2, \ldots, p_n$

- To control FDR to be within 0.05
  - Sort p-values from low to high, $p'_1, p'_2, \ldots, p'_n$
  - Find largest k such that $p'_k \leq 0.05 \times k / n$
    - For the smallest p-value, this is equivalent to Bonferroni
    - For other p-values, this technique gradually loosens the cutoff
  - Reject null hypothesis for tests corresponding to $p'_1, p'_2, \ldots, p'_k$

Correction method comparison

P-value	Bonferroni	B-H	B-Y
Smallest	0.0005	0.0005	0.0005
2nd smallest	0.0005	0.001	0.000667
3rd smallest	0.0005	0.0015	0.000818
4th smallest	0.0005	0.002	0.00096
5th smallest	0.0005	0.0025	0.001095

- There are 100 tests
- Target p-value or FDR cutoff = 0.05

Effect of correction

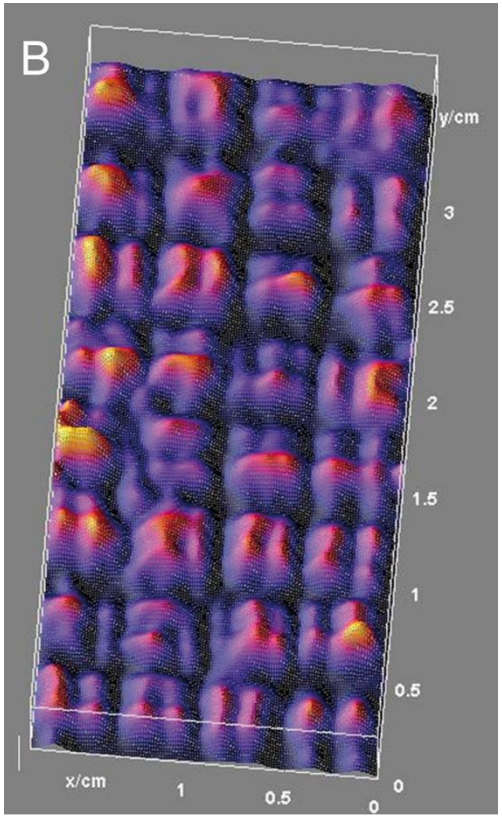
Gene	Sorted p-value	Rank	Benjamini-Hochberg	Result	c(rank)	Benjamini-Yekutieli	Result
Gene M	0.0000001	1	0.0005	Pass	1	0.0005	Pass
Gene S	0.0000035	2	0.001	Pass	1.5	0.00067	Pass
Gene A	0.00028	3	0.0015	Pass	1.83	0.00082	Pass
Gene C	0.0011	4	0.002	Pass	2.08	0.00096	Fail
Gene P	0.06	5	0.0025	Fail	2.28	0.0011	

Linear effect model

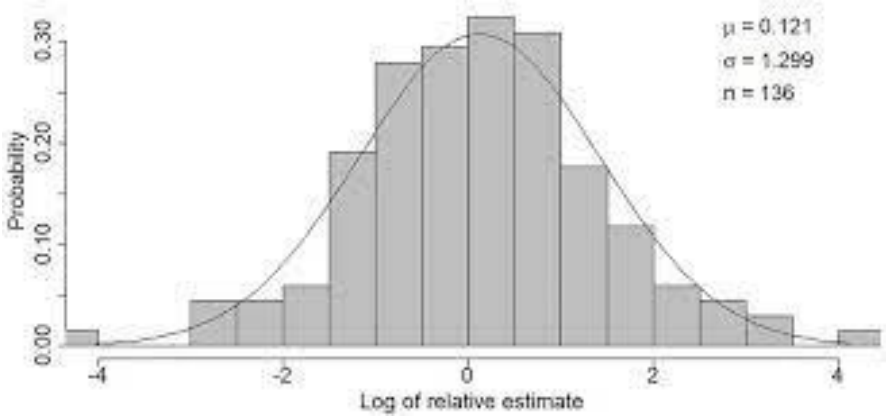
Background noise correction models

- Null hypothesis
  - Background noise is normally distributed and is the same over the entire array

- Linear effect model
  - Background noise is normally distributed with mean depending on (x, y) positions and a fixed variance



Fitting normally distributed data



- Probe intensities: $n_1, n_2, \ldots, n_k$
  - Fitted mean and variance: $\mu = \frac{\sum_i n_i}{k}$ and $\sigma^2 = \frac{1}{k-1}\sum_i (n_i - \mu)^2$
  - Likelihood: $\prod_i P(n_i | \mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^k e^{-\frac{1}{2}\sum_i \left(\frac{n_i - \mu}{\sigma}\right)^2}$

Linear effect model

- Position of probe *i* with intensity n_i is (x_i, y_i)

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

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

Some algebra exercises

- Setting partial derivatives to zero
  - $0 = \sum_i \big(n_i - (ax_i + by_i + c)\big)(-x_i)$
  - $0 = \sum_i \big(n_i - (ax_i + by_i + c)\big)(-y_i)$
  - $0 = \sum_i \big(n_i - (ax_i + by_i + c)\big)$

- Or equivalently
  - $a \sum_i x_i^2 + b \sum_i x_i y_i + c \sum_i x_i = \sum_i n_i x_i$
  - $a \sum_i x_i y_i + b \sum_i y_i^2 + c \sum_i y_i = \sum_i n_i y_i$
  - $a \sum_i x_i + b \sum_i y_i + ck = \sum_i n_i$

- Three linear equations with three variables ☺

Incorporating confounding variable

- Design matrix

Sample	Condition	Batch	Patient's Age
S1	Control	1	35
S2	Control	2	21
S3	Control	3	45
S4	Treatment	1	18
S5	Treatment	2	37
S6	Treatment	3	52

Any question?