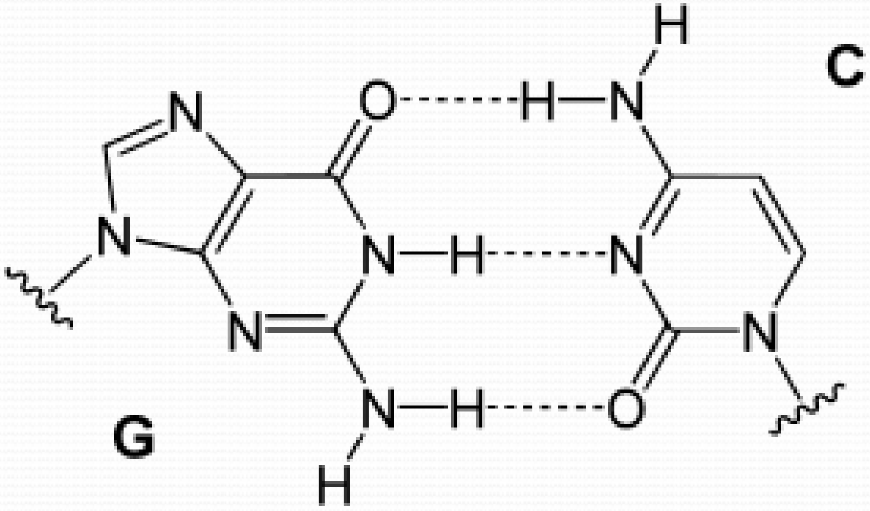
PCR, Immunoassays, RNA-Seq, and Mass Spectrometry

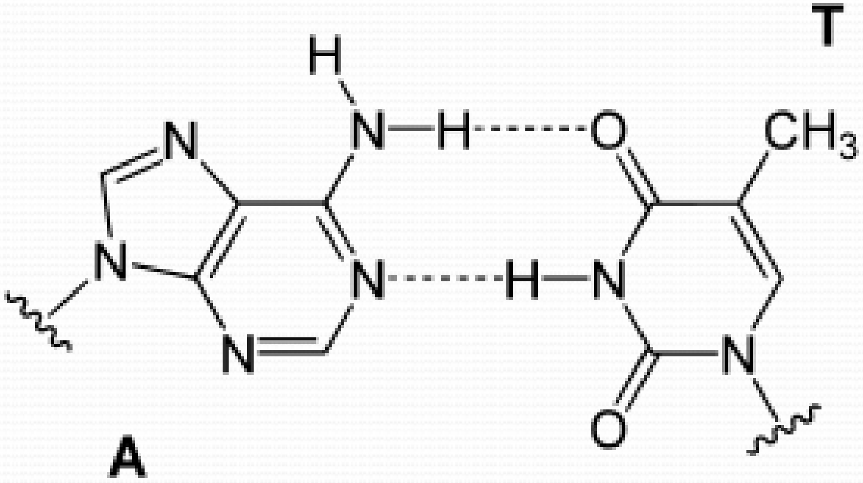
BIM 283

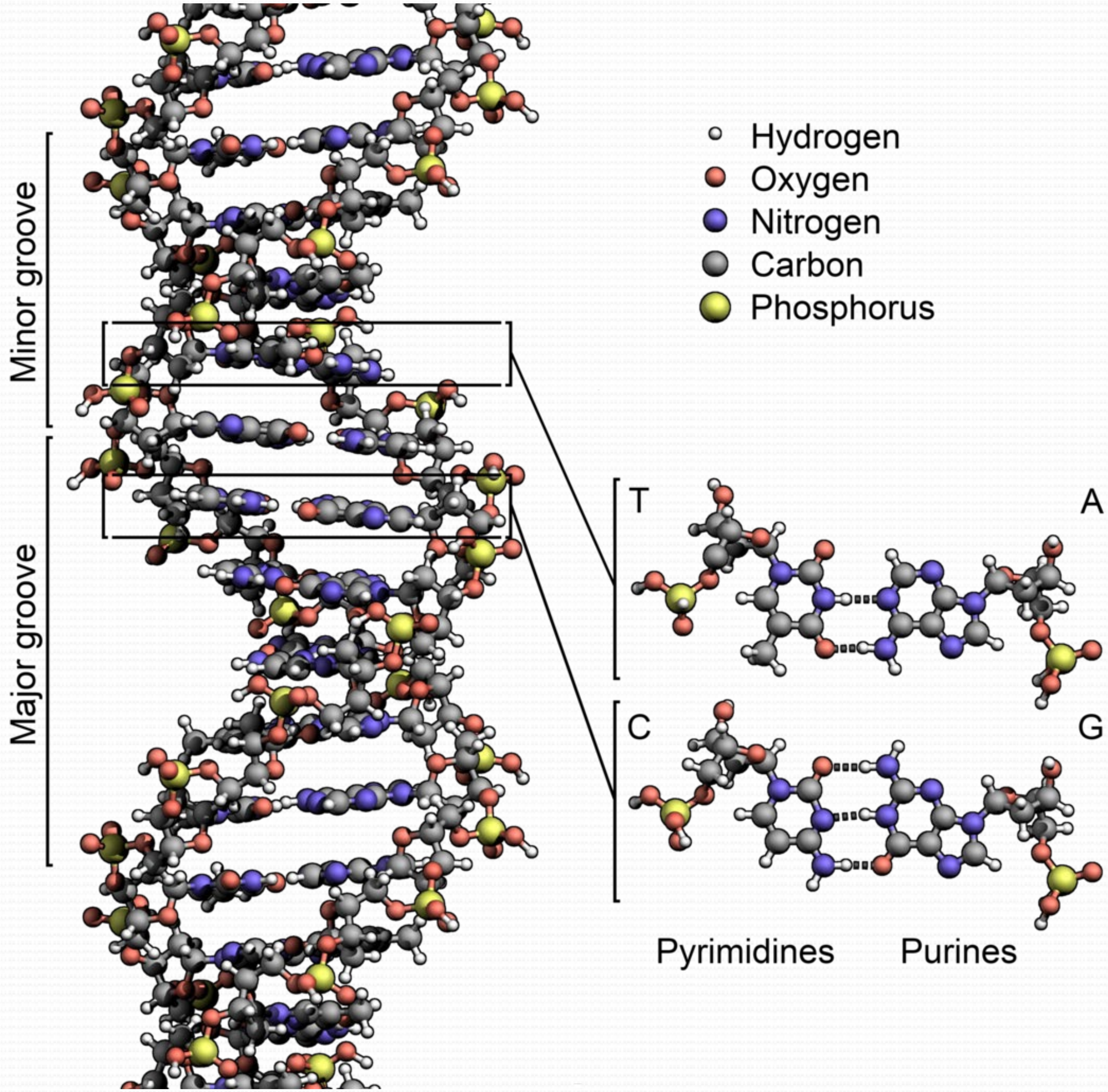
Advanced Design of Experiments for Biomedical Engineers

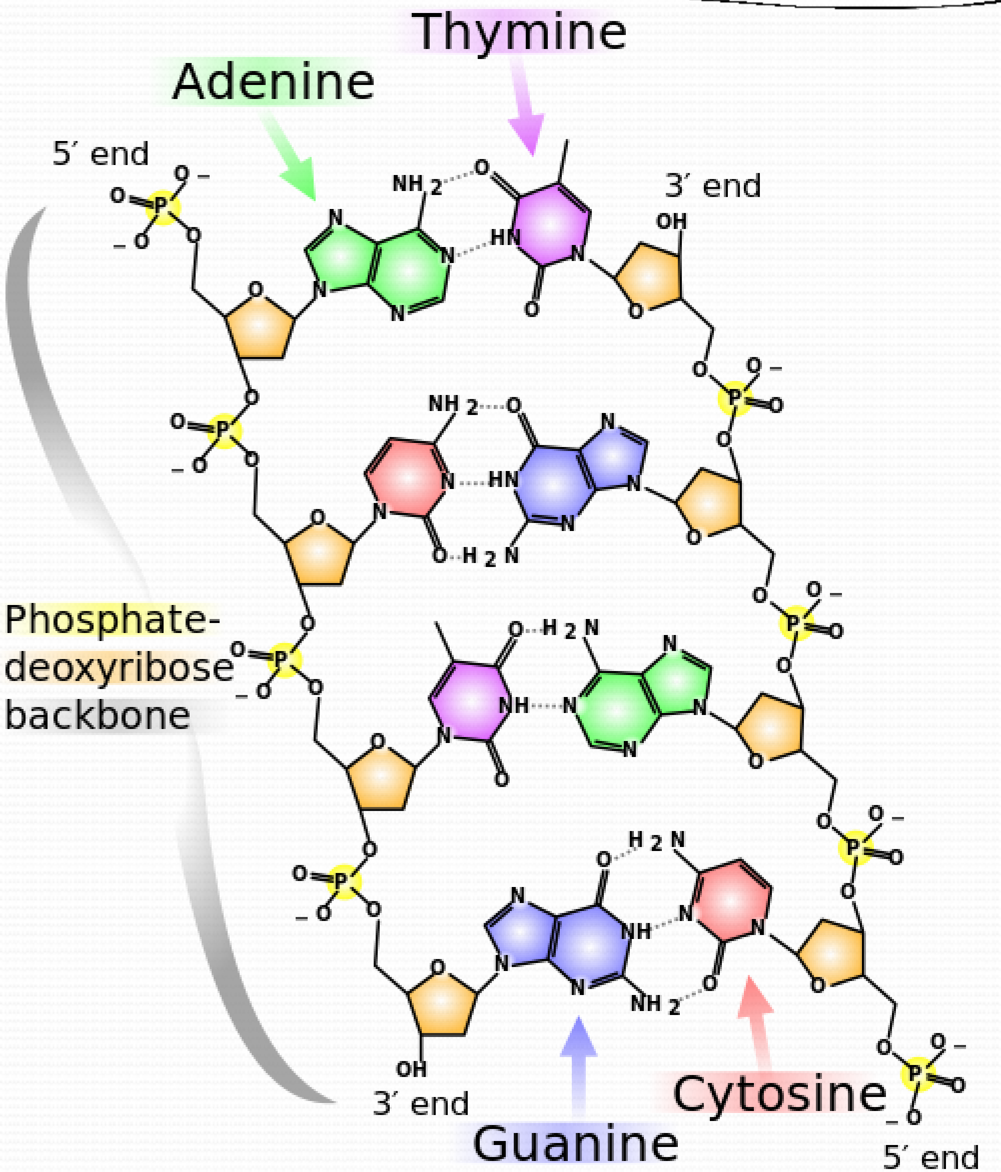
DNA and RNA

- DNA and RNA are polymers composed of four subunits.
- The four subunits in DNA are the nucleotides guanine, adenine, thymine, and cytosine (G, A, T, C) on a sugar (deoxyribose)/phosphate backbone.
- Pairs of nucleotides bind well to each other by hydrogen bonds, with GC forming one pair and AT the other.
- A and G are purines, C and T are pyrimidines. Uracil (U) is a pyrimidine that takes the place of thymine in RNA
- For a given strand of DNA the complementary strand (cDNA) has the matching nucleotides, so
  GGTCACTG matches
  CCAGTGAC


C
G


T
A


Hydrogen
Oxygen
Nitrogen
Carbon
Phosphorus
Minor groove
Major groove
T
A
C
G
Pyrimidines
Purines


Thymine
Adenine
5′ end
3′ end
Phosphate-
deoxyribose
backbone
3′ end
Guanine
Cytosine
5′ end
NH 2
O
HN
N
OH
O −
P
− O
H 2 N
NH

Transcription

- DNA can be transcribed to RNA (which has a different sugar on the backbone) using RNA polymerase
- RNA can be reverse-transcribed to DNA to allow DNA assays to be used on RNA
- DNA transcription in eukaryotes involves editing such as removal of introns.
- Reverse transcription is used in the lab for example for applying PCR to RNA.

PCR for Measurement

- PCR is a method of measuring the copy number of a particular DNA sequence in a sample.
- It uses an enzyme (DNA polymerase) that copies one strand of DNA to its complement, and this is done with both strands, so the total copy number is approximately doubled.
- Primers bind to a specific sequence, so only DNA with that sequence is amplified.
- A cycle of temperature changes should result in approximately doubling the copy number.
- The read-out is obtained by a fluorescent dye

- In 20 cycles, the amount of the analyte should be increased by a factor of 2^{20} = 1 million.
- In 40 cycles, the increase is about 2^{40} = 1 trillion.
- If we perform a fixed number of cycles, then the range of measurement is rather narrow, between the number of molecules that barely crosses the threshold, to one that maxes out the assay.
- Quantitative PCR = Real-Time PCR establishes a threshold of brightness, and the cycle at which it passes that threshold is the measurement. This is usually interpolated between the last cycle dimmer than the threshold and the first cycle brighter than the threshold.

qRT-PCR

- Reverse Transcription (RT) PCR quantifies RNA by using a reverse transcriptase to create the equivalent DNA sequences.
- RT-PCR should not be used to mean “Real-Time” PCR because it can be confused with “Reverse Transcription” PCR.
- qRT-PCR uses the cycle threshold method after reverse transcription

Sensitivity of PCR

- Very specific and very sensitive.
- A mere 25 copies of a transcript will be amplified to detection in qRT-PCR.
- Suppose we need 70 ng of RNA per run, and suppose we have a sample of RNA with an average copy number of 50 copies per 70 ng.
- The actual copy number is Poisson with a mean of 50, and (therefore) a standard deviation of $\sqrt{50} = 7.07$. Almost all 70 ng aliquots will have a copy number of $50 \pm (3)(7.07) = 50 \pm 21$ or between 29 and 71 copies, and thus will be detected.
- So the detection limit is about 25 copies and the minimum detectable value is about 50 copies.

Quantitative PCR

- We can get relative quantitation because the fluorescence is proportional to the quantity present at the end of the cycle.
- The copy number present at the end of cycle k is roughly 2^k times the copy number at the start.
- To obtain absolute copy number estimates, when those are needed, we can make a calibration curve or use standards along with each sample.

$z_0 =$ initial copy number

$Z_k =$ copy number at the end of cycle k

$X_{ik} =$ copies from i$^{\text{th}}$ copy in cycle $k-1$ (1, 2,...)

$$\mathrm{E}(\mathrm{X}_{ik}) = m = 1+\rho$$

$$V(X_{ik}) = \rho(1-\rho)$$

$$Z_k = \sum_{i=1}^{Z_{k-1}} X_{ik}$$

$$\mathrm{E}(Z_k) = z_0 m^k$$

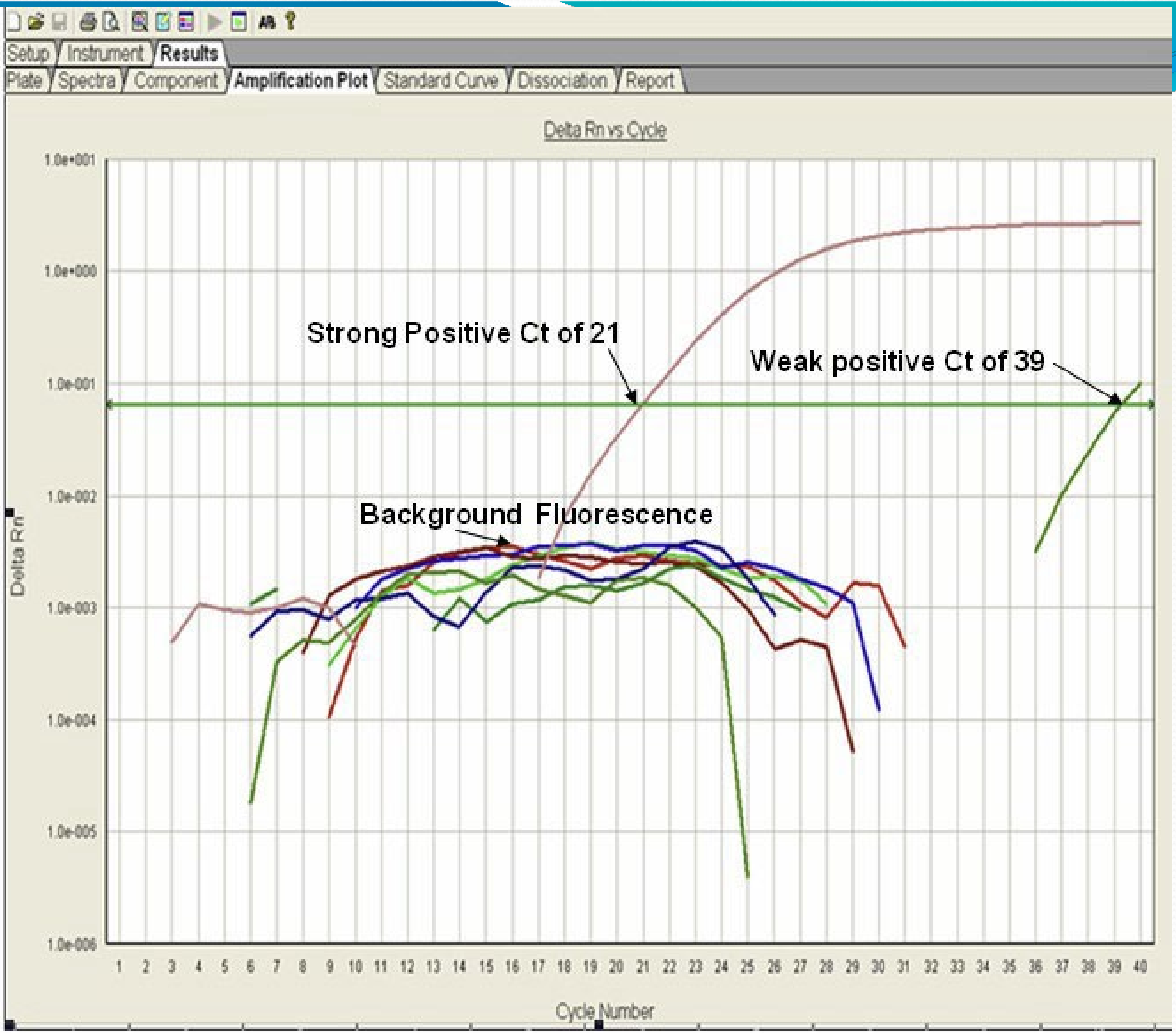
$$\mathrm{Var}(Z_k) = z_0^2(1-\rho)m^{k-1}(m^k-1) \doteq z_0^2(1-\rho)m^{2k-1} \quad \text{for } k \text{ large}$$

$$\mathrm{SD}(Z_k) = z_0 m_{k-1/2}\sqrt{1-\rho}$$

$$\mathrm{CV}(Z_k) = m^{-1/2}\sqrt{1-\rho} = \sqrt{\frac{1-\rho}{1+\rho}}$$

ρ	1	0.99	0.95	0.90
CV	0	0.071	0.160	0.229

rho	0.95											
				10		9		8		7		Total
5	77.4%	10	59.9%	20							46.3%	20
	20.4%	9	31.5%	19							24.4%	19
	2.1%	8	7.5%	18	63.0%	18					18.6%	18
	0.1%	7	1.0%	17	29.9%	17					6.9%	17
			0.1%	16	6.3%	16	66.3%	16			2.8%	16
			0.0%	15	0.8%	15	27.9%	15			0.8%	15
			0.0%	14	0.1%	14	5.1%	14	69.8%	14	0.2%	14
			0.0%	13	0.0%	13	0.5%	13	25.7%	13	0.0%	13
			0.0%	12	0.0%	12	0.0%	12	4.1%	12	0.0%	12



Strong Positive Ct of 21

Weak positive Ct of 39

Background Fluorescence

Interpreting qRT-PCR

- A threshold level is set that is enough above background that it would not happen by chance
- The response Ct is the cycle at which the signal exceeds the threshold
- We usually interpolate. If the threshold is 100, and the signal at cycle 20 is 88, and the signal at cycle 21 is 169, then we could interpolate linearly, but interpolation on the log scale is better, because the signal rises exponentially.

$$\text{Threshold} = 100$$

$$S_{20} = 88$$

$$S_{21} = 169$$

$$C_t = 20 + (100 - 88)/(169 - 88) = 20.15$$

$$\ln(S_{20}) = \ln(88) = 4.417$$

$$\ln(S_{21}) = \ln(169) = 5.130$$

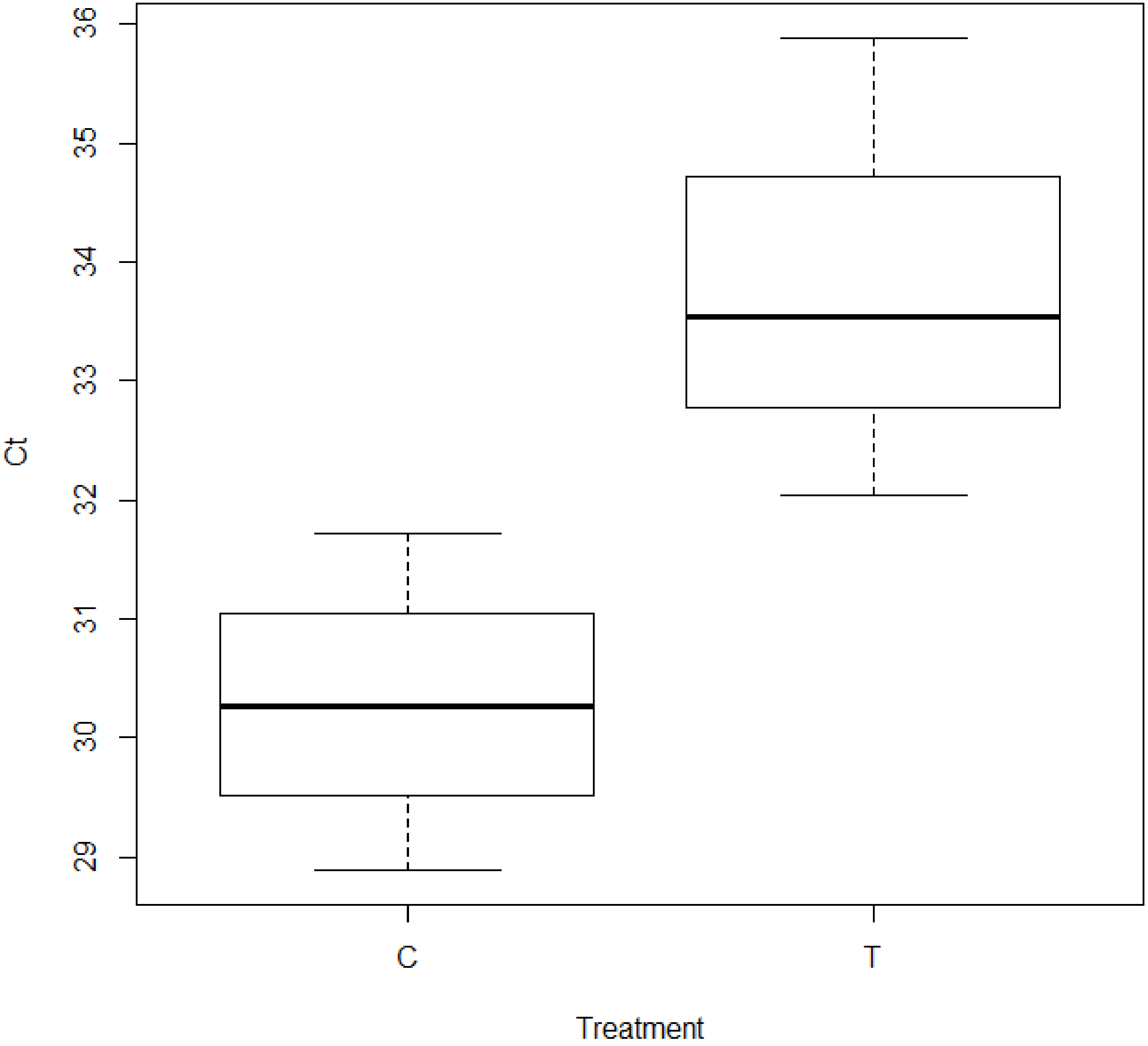
$$\ln(100) = 4.605$$

$$C_t = 20 + (4.605 - 4.417)/(5.130 - 4.417) = 20.20$$

- The higher the Ct value, the lower the original copy number
- Ct is on the log scale, so this is often a good scale for ANOVA and regression
- If there is a control in each tube or well, we can then analyze Ct(sample) – Ct(control)
- If we need to know the actual copy number (as for HIV), we need a calibration curve, which may be run on a periodic basis.

Example

- Eight HIV patients are assayed, four with a new treatment and four controls with current standard of care.
- The Ct values for the treated patients are 32.03, 35.89, 33.57, and 33.51
- The Ct values for the control patients are 30.14, 31.72, 28.88, and 30.38.
- Lower Ct values mean higher copy number in the original sample, which is more copies of the virus.

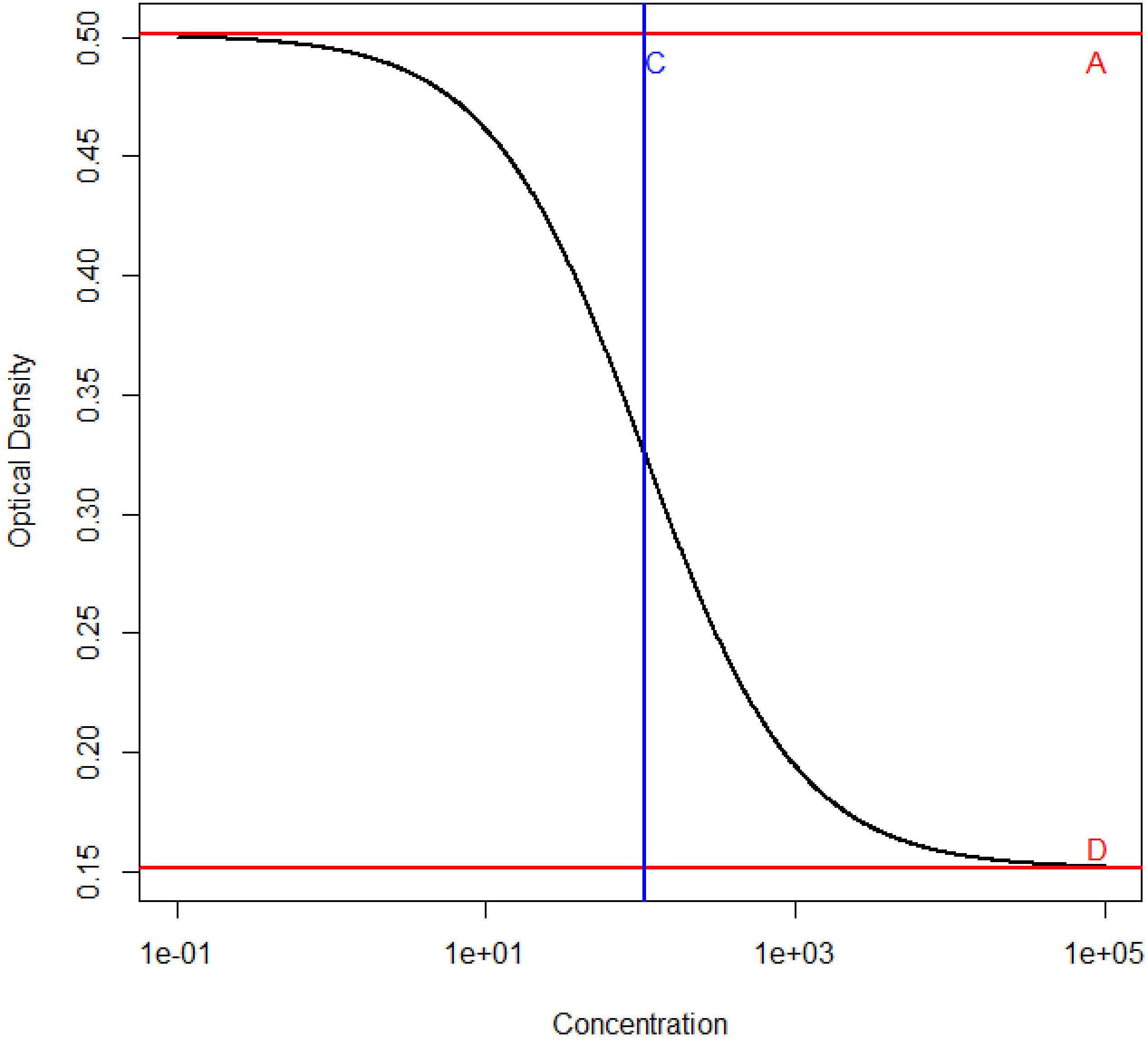

36
35
34
33
32
31
30
29
Ct
C
T
Treatment

```
> anova(lm(Ct ~ Treat))
Analysis of Variance Table

Response: Ct
          Df Sum Sq Mean Sq F value  Pr(>F)
Treat      1 24.064 24.0642  12.361 0.01258 *
Residuals  6 11.681  1.9468
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ELISA

- Enzyme-linked immuno sorbent assay (ELISA) is a test that uses antibodies and color change to identify a substance.
- Requires an antibody to the analyte.
- Depending on the assay type, binding results either in increased color or decreased color.
- Readout is through optical density/brightness at a particular frequency


Optical Density
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
C
A
D
1e-01
1e+01
1e+03
1e+05
Concentration

Four-Parameter (Log-)Logistic Curve

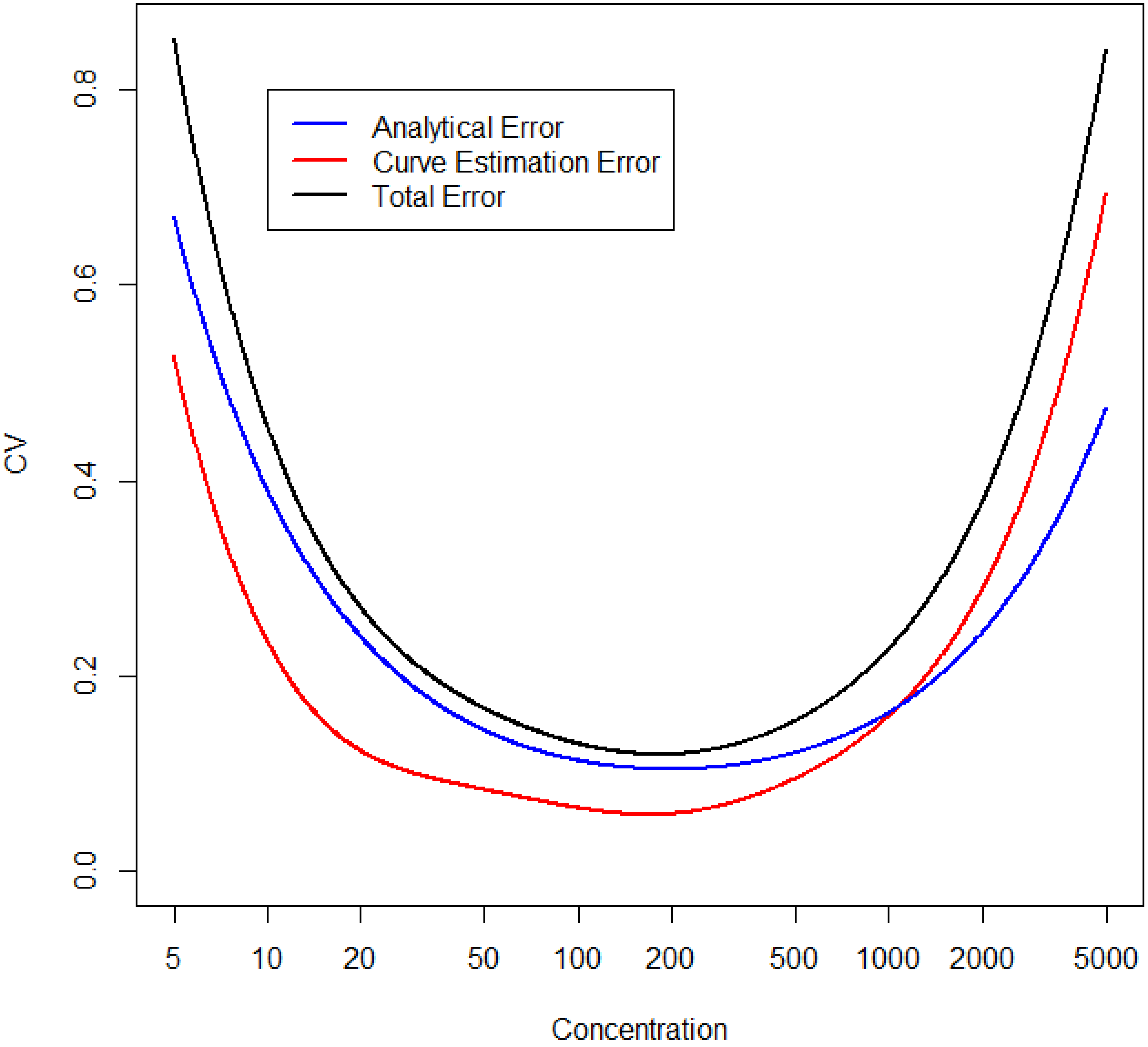
$$y = f(x) = \frac{A - D}{1 + (x / C)^B} + D$$

$$\frac{y - D}{A - D} = \frac{1}{1 + (x / C)^B}$$

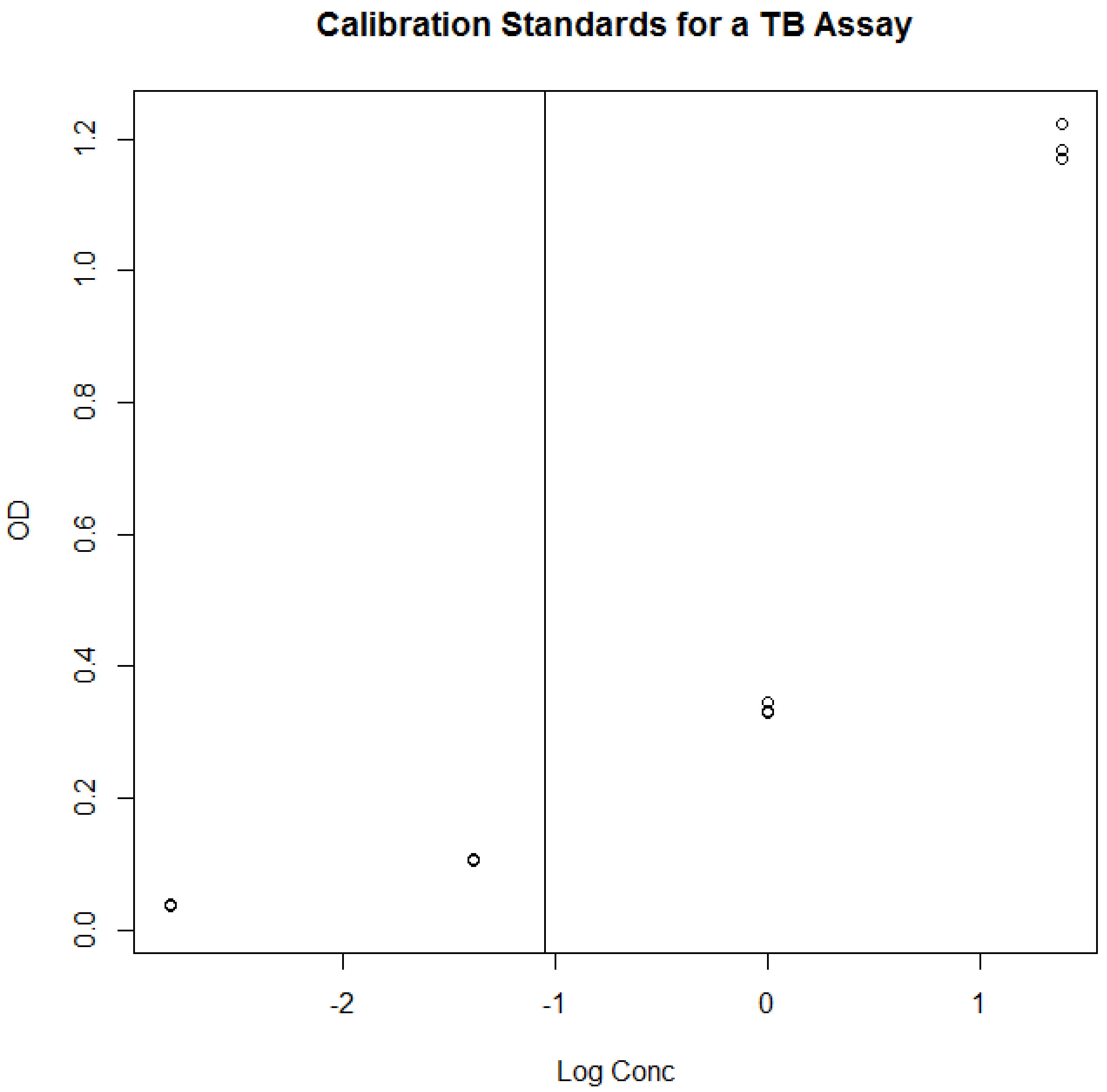
$$\frac{A - D}{y - D} - 1 = (x / C)^B$$

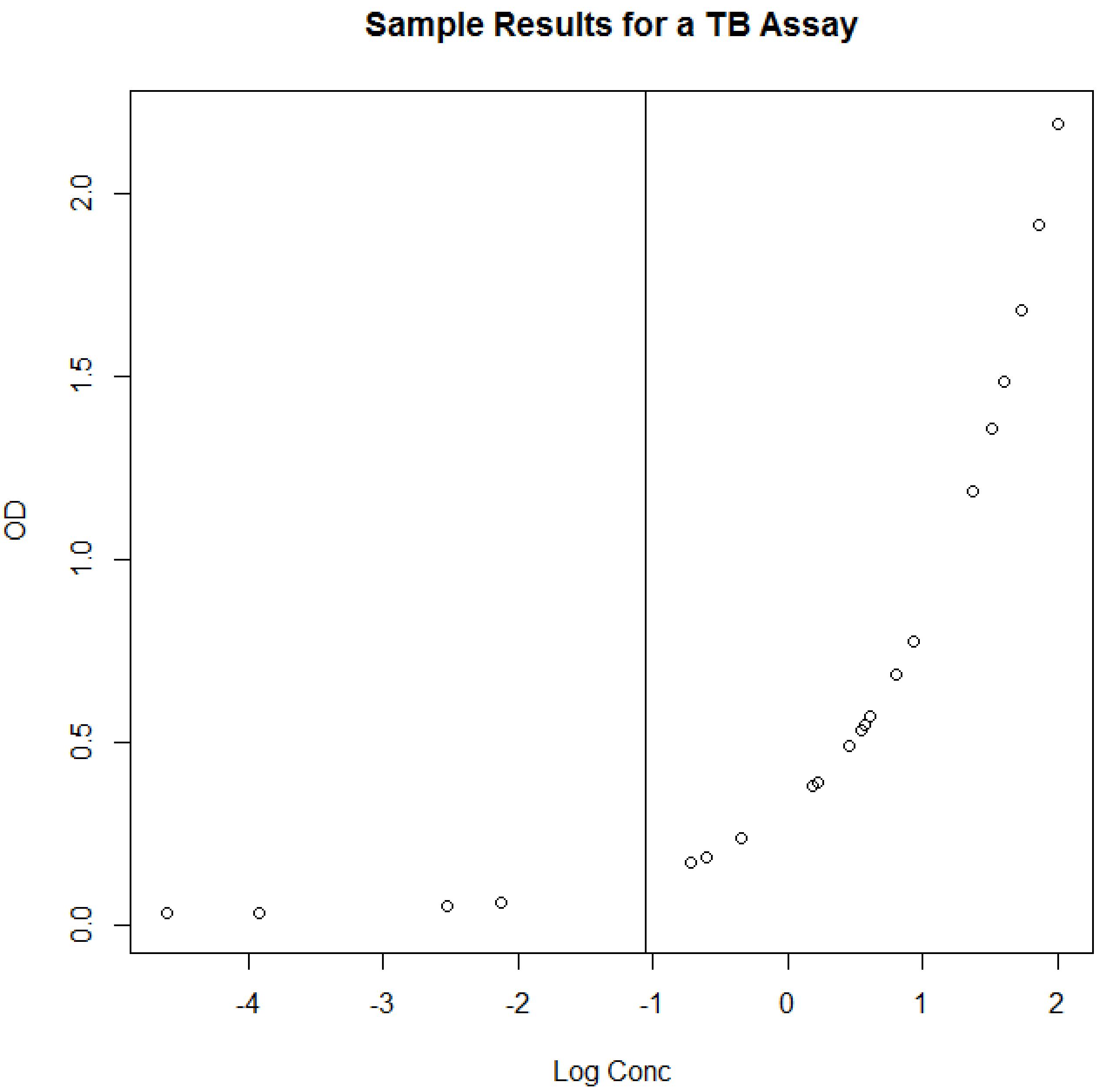
$$\ln\left(\frac{A - y}{y - D}\right) = B[\ln(x) - \ln(C)]$$

- ELISAs are often run on 96-well plates. Calibration valuesare in some of the wells so that the parameters of the 4-parameter log-logistic curve can be estimated.
- OD values above A are interpreted as an estimated concentration of 0. OD values below D are out of bounds high.
- In some forms of ELISA, the OD is low for low concentrations and high for higher ones.
- The assay is generally good in a region around the center, not so good at the ends


Analytical Error
Curve Estimation Error
Total Error
CV
0.0
0.2
0.4
0.6
0.8
5
10
20
50
100
200
500
1000
2000
5000
Concentration

- If the concentration is too low, then the assay will not yield useful results.
- If the concentration is too high, then a dilution of the sample can fall into the accurate range.
- Often, a dilution series is run. This way, if the original concentration is too high, dilutions at a factor of 10 or 100 will fall into the accurate range.


Calibration Standards for a TB Assay
OD
0.0
0.2
0.4
0.6
0.8
1.0
1.2
-2
-1
0
1
Log Conc


Sample Results for a TB Assay
OD
0.0
0.5
1.0
1.5
2.0
-4
-3
-2
-1
0
1
2
Log Conc

Luminex Assays

- Multiplexed bead-based ELISA-type assay.
- Several antibody bead types in each cell, labeled with different dyes that fluoresce at different frequencies.
- By default, calibration standards are included for all the analytes in separate cells.
- Calibrated values can be hard to analyze because values below and above the “good” range are reported just as High or Low, and those cannot be used in an analysis.
- I usually use the raw values

Example Panel (Millipore)

- 42 Cytokine/Chemokine assays
- EGF, Eotaxin, FGF-2, Flt-3 ligand, Fractalkine, G-CSF, GM-CSF, GRO, IFN-α2, IFN-γ, IL-10, IL-12 (p40), IL-12 (p70), IL-13, IL-15, IL-17, IL-1Rα, IL-1α, IL-1β, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IP-10, MCP-1, MCP-3, MDC (CCL22), MIP-1α, MIP-1β, PDGF-AA, PDGF-AB/BB, RANTES, TGF-α, TNF-α, TNF-β, VEGF, sCD40L, sIL-2Rα

Location	Sample	1 IL-1b	10 IL-5	11 EGF	12 IL-6	13 IL-7
A1	blank	35.5	68	65	309	80
B1		12	32	12.5	160	43
C1	3.2pg/ml	146	219.5	55.5	347	135
D1		146	205.5	45	319	115
E1	16pg/ml	646	744.5	83	624	271
F1		564	782.5	63.5	625	236
G1	80pg/ml	2932	2850.5	274	1870	954
H1		2597	2680.5	285	1550	782.5
A2	400pg/ml	10578	9537	2089.5	7173	5728.5
B2		11399.5	8763	2487	8042	6685
C2	2000pg/ml	18343	22325	10573	21016	22340
D2		18146	22309.5	11662	21066	22446
E2	10 000pg/ml	18196.5	21775	14582.5	21159	25093
F2		18820	22690.5	16094	22608	25194
G2	QC1	2865	3660.5	408	2022.5	1113
H2		2448	3468.5	393	1843	989
A3	QC2	12462.5	13819	3396	12246	9638
B3		12803	15047.5	4122.5	12770	10391.5
C3	1	76	35.5	404	91	84
D3		68	33	393	81	76
E3	2	30	509	7914	1199	61.5
F3		32	587	8215	1417	70
G3	3	44.5	40	7536	137	61
H3		39.5	36	7328	154.5	74

Cells in the 96 well plate are indexed by
Rows A-H
Cols 1-12

Analysis of Luminex Data

- Use raw fluorescence data
- Take logs (after perhaps adding something to each value)
- In the illustrated data, the numbers are as small as 3, and as large as 25,000. One might add 20–50 before taking logs
- This provides relative quantitation, which is all that is required.
- Calibrated values would be used only when the absolute levels are useful

RNA-Seq

- Gene expression is the transcription of the DNA in a gene into mRNA, which (in many cases) is later translated into a protein.
- We can measure expression of a single gene with PCR or other assays.
- Gene expression arrays measure expression of many genes simultaneously using spots each of which contains a matching sequence to the gene sequence to be detected.
- RNA-Seq is more comprehensive (and expensive!)

RNA-Seq

- For RNA-Seq, the RNA in the sample is reverse transcribed into the corresponding DNA sequence.
- Then the DNA fragments are sequenced (in an NGS sequencer, usually Illumina)
- Each fragment is mapped to the reference genome
- The data to be analyzed are the number of fragments mapping to each gene in a table where the columns are samples and the rows are genes.

RNA-Seq

- This mapping can be complex
- We can choose to estimate isoforms or not (alternative splicing leading to different proteins)
- We can choose how to handle ambiguous reads (omit or spread across genes)
- We can then use statistical analysis to determine when there is significantly more expression in one condition or another.

Analysis of RNA-Seq Data

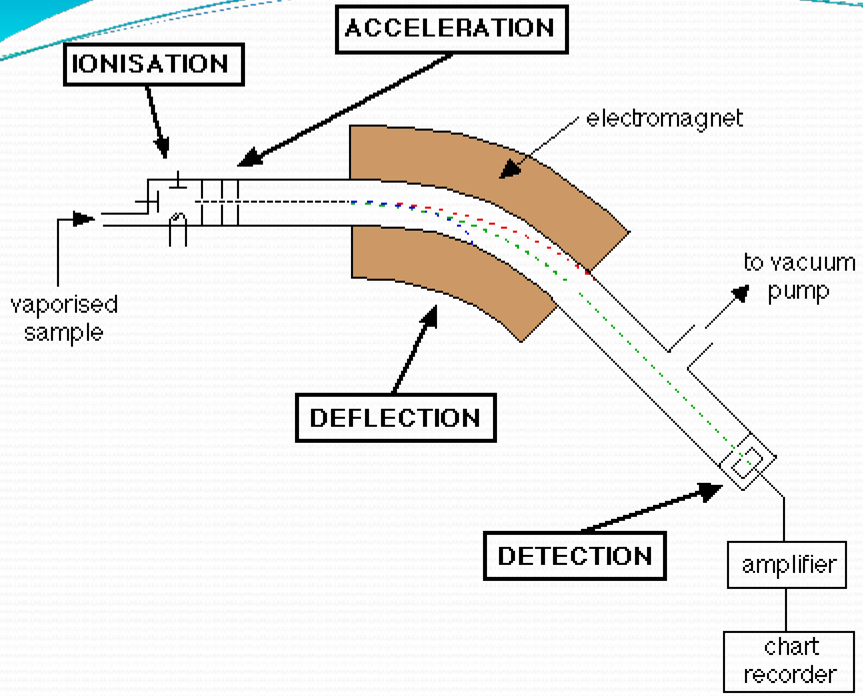
- For each gene/exon/isoform (we will say gene from now on), and for each sample, we have a count of fragments mapping to that gene.
- In principle, we need to test whether the counts from one group are significantly larger than another.
- Or we may have more than one factor or variable that could be associated.
- In practice, we may (probably) need to normalize the samples first and may need to import some information across genes.

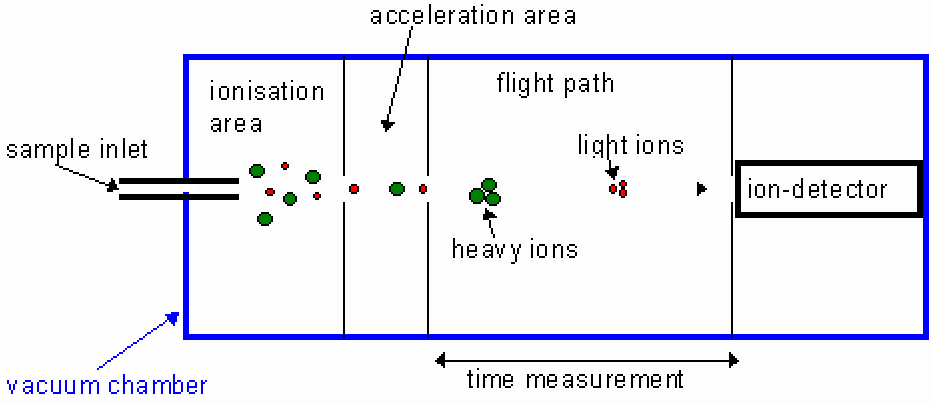
Mass Spectrometry

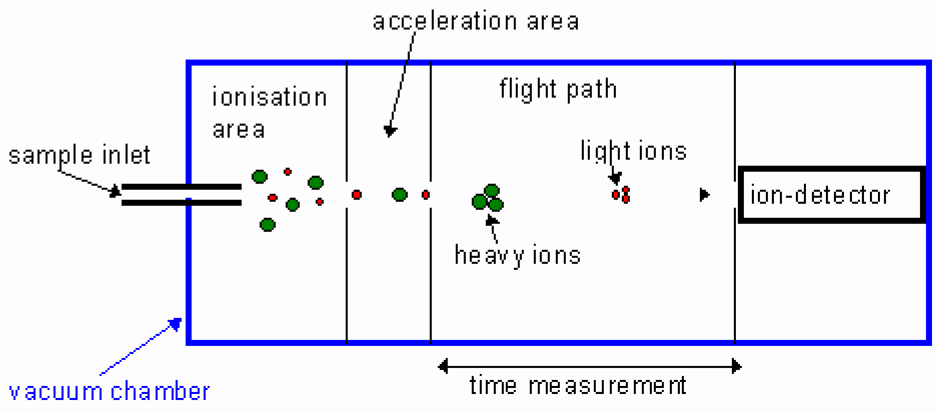
- Mass spectrometry (mass spec, MS) comprises a set of instrumental methods that can identify compounds by molecular weight and can potentially quantify by using the peak height or peak area.
- If the substance to be analyzed is not already a gas, it needs to be vaporized and also ionized so that the molecules are charged, most commonly by elimination of an electron.
- The behavior of the ion can be measured in terms of the mass-to-charge ratio (m/z) because a 100 Dalton+ molecule behaves in a magnetic field like a 200 Dalton ++ molecule; m/z is the same.

Physics

- One method is to detect the amount of deflection as the ions traverse an evacuated tube in a magnetic field. The amount of deflection at the end of the tube is smaller when m/z is larger. Varying the magnetic field yields a spectrum of total ion current at each field strength = at each m/z
- Time-of-flight MS is used for larger molecules. The ions are accelerated at the beginning of the tube, and the velocity larger for smaller m/z, so the time of flight is smaller. The spectrum is collected over time and the x-axis converted to m/z.
- Varieties of ion trap MS circulate the ions in a magnetic field and generate a signal that can be deconvolved with a Fourier transform.

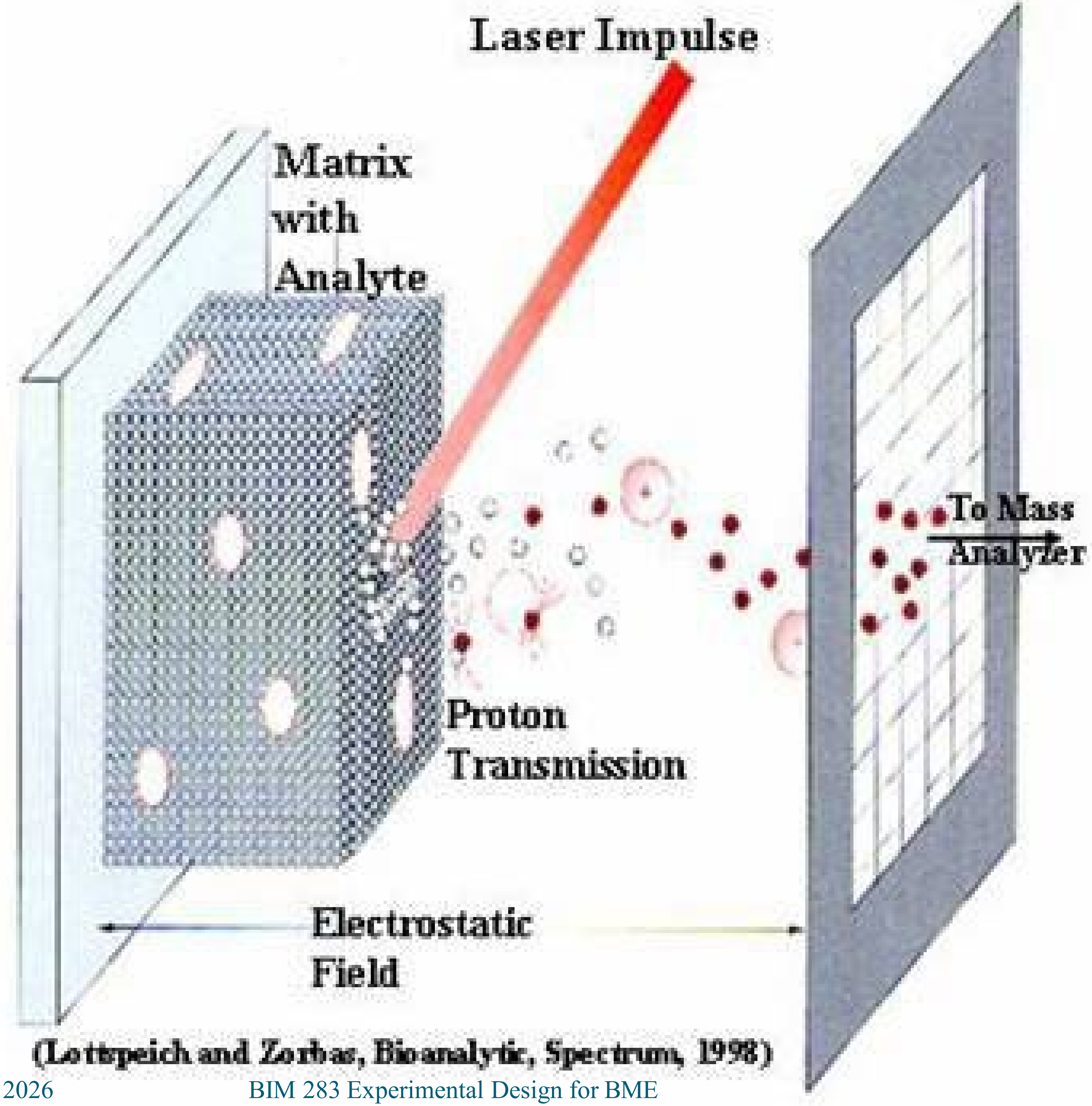

ACCELERATION
IONISATION
electromagnet
to vacuum pump
vaporised sample
DEFLECTION
DETECTION
amplifier
chart recorder


acceleration area
ionisation area
flight path
sample inlet
light ions
ion-detector
heavy ions
vacuum chamber
time measurement


acceleration area
ionisation area
flight path
sample inlet
light ions
ion-detector
heavy ions
vacuum chamber
time measurement

MALDI

- Matrix Assisted Laser Desorption Ionization
- Sample is embedded in matrix on a slide
- Laser energy absorbed by matrix, ionizes and vaporizes matrix and sample.
- Can fragment compounds
- Can doubly or triply ionize compound—Look for peak at half or a third of the expected m/z
- Results of different shots can be very different—multiple shots for each sample



(Lottspeich and Zorbas, Bioanalytic, Spectrum, 1998)

Electrospray Ionization

- Electrospray ionization uses electricity to disperse a liquid
- High voltage is applied to a liquid supplied through an emitter (usually a glass or metallic capillary).
- This leads to the formation of small and highly charged liquid droplets, which are radially dispersed due to Coulomb repulsion.
- This is often used between a separation stage (liquid chromatography) and the mass spec analyzer.

Separation Technologies

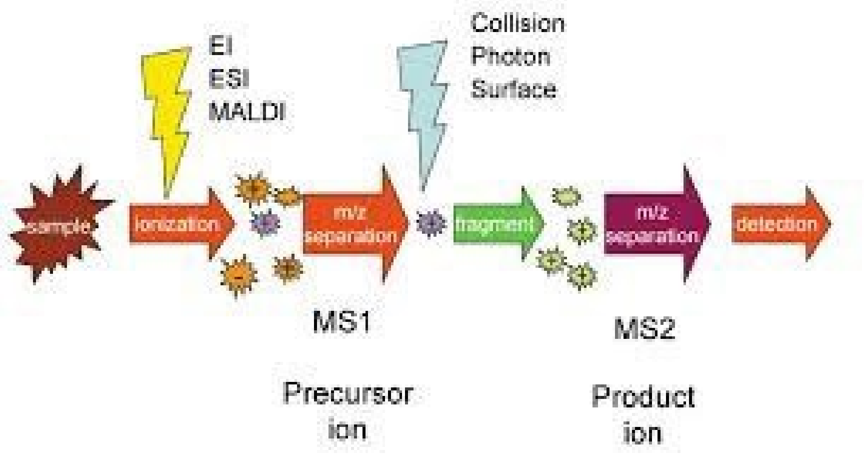
- Various chemical methods can be used to produce a sample that contains mostly the class of analyte of interest.
- These can be small molecule organic compounds, lipids, oxy-lipids, sugars, proteins, etc.
- Gas or liquid chromatography will separate the sample by various chemical properties and then aliquots will be serially analyzed by mass spec.
- This can help differentiate molecules that have the same weight because they elute at different times.

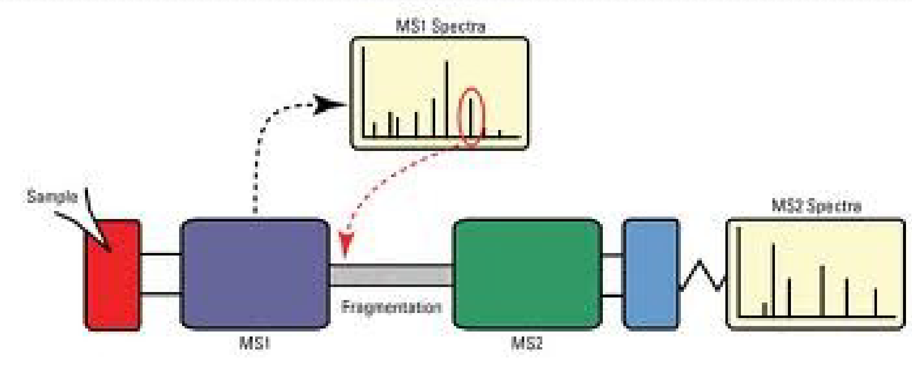
Metabolomics

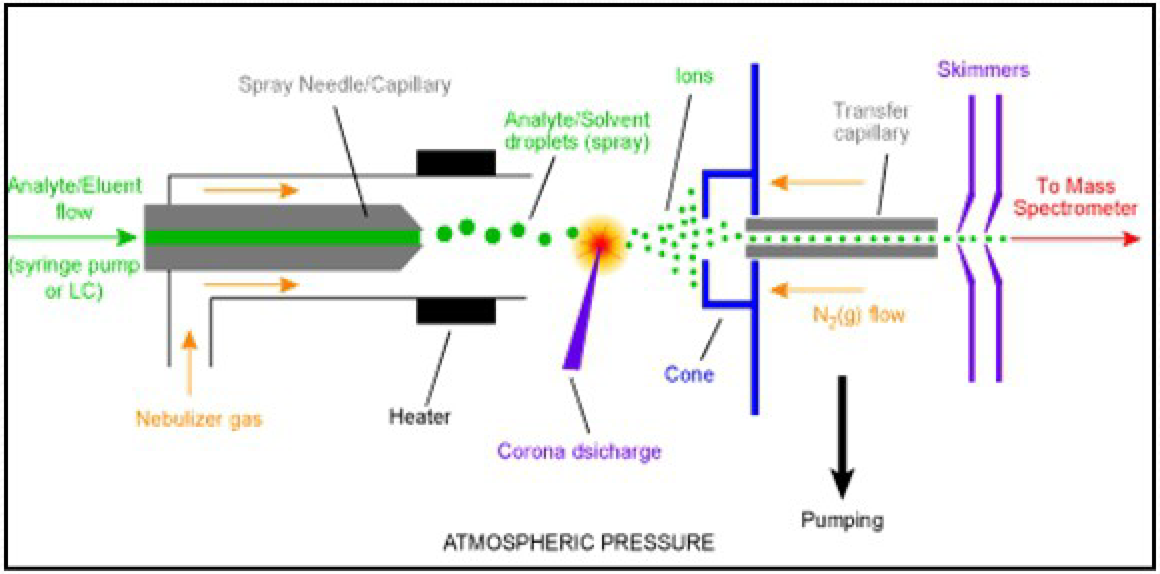
- Mass spectrometry (e.g., LC/ToF-MS can detect and measure relative amounts of small molecules much more easily than with proteins
- Lipids, saccharides, and others
- For example we can speciate 500 lipids including di- and tri-glycerides species
- For example, we can measure over a hundred compounds in the arachidonic acid pathway including prostaglandins, COX1, COX2, and LOX products, and CyP450 pathways eicosanoids.
- Can measure enzymes, substrate, and product

Proteomics

- Proteins can be very large molecules, up to several thousand amino acids or hundreds of thousands of Daltons.
- These must be broken up into much smaller peptides, usually chemically with a protease, before they can possibly be ionized.
- These peptides are then usually further fragmented in a second chamber and the spectrum of fragment sizes recorded. This is called tandem mass spec or MS/MS


EI
ESI
MALDI
Collision
Photon
Surface
sample
ionization
m/z separation
fragment
m/z separation
detection
MS1
MS2
Precursor ion
Product ion


MS1 Spectra
Sample
Fragmentation
MS1
MS2
MS2 Spectra


Spray Needle/Capillary
Ions
Skimmers
Analyte/Solvent droplets (spray)
Transfer capillary
Analyte/Eluent flow
(syringe pump or LC)
To Mass Spectrometer
N$_2$(g) flow
Cone
Nebulizer gas
Heater
Corona dsicharge
Pumping
ATMOSPHERIC PRESSURE

Processing Spectra

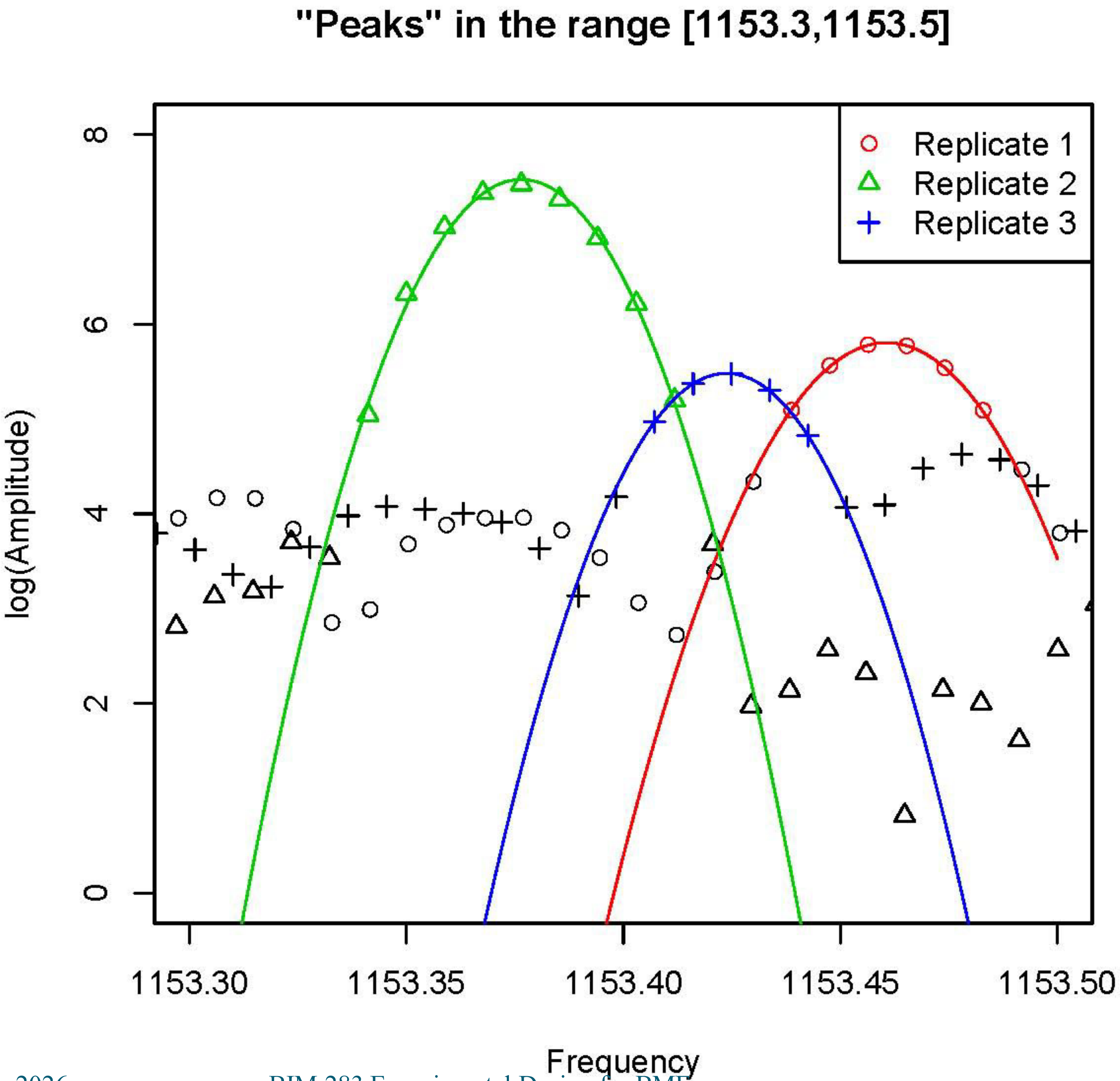
- Baseline estimation
- Peak identification
- Mass calibration
- Data transformation
- Normalizing across spectra
- Peak quantitation
- Identification of isotope and shoulder peaks
- FTICRMS R package

Baseline Correction

- In regions of the spectrum where no analyte is present, the signal should fluctuate around a baseline, which should be set to zero
- In peak regions, the estimated baseline forms the base for peak heights or area.
- For some technologies, the baseline tends to be relatively flat, but for others, such as NMR and FT-ICR MS, the baseline is curved or wavy.
- It is important to get the baseline right to find peaks and accurately measure them.

Peak Identification

- In some cases, a peak is just a point on the spectrum where the points to the left and right have lower magnitudes.
- But this can result in false multiple peaks from jagged signal.
- For FT-ICR MS, the peaks look quadratic on the log scale which allows accurate identification.


"Peaks" in the range [1153.3,1153.5]
log(Amplitude)
Frequency
0
2
4
6
8
1153.30
1153.35
1153.40
1153.45
1153.50
○ Replicate 1
△ Replicate 2
+ Replicate 3

Mass Calibration

- This is supposedly done by in-machine software using known large peaks
- In principle, the mass accuracy of FTICRMS at 1000 Daltons should be 0.001 Daltons or better
- In practice, as shown on the previous slide, the same compound can be as much as 0.1 Daltons apart
- Because the theoretical mass accuracy of FT-ICR is so high, this mass calibration process can be done very effectively.
- The accuracy of ToF MS and NMR is much worse.

Data Transformation

- Data from MS and most other technologies needs to be transformed, usually to a log scale to make analysis effective.
- Care needs to be taken at the low end .
  - Logs of zero and negative numbers are not defined
  - Signal fluctuating around a zero baseline will often be negative.
- Shifted log (add a constant), generalized log, and other methods can be used.

Statistical Analysis

- Use only peaks that are present in a minimum number of samples, impute peaks for samples in which peaks are not detected
- Appropriate statistical analysis per peak depending on the design (e.g., one-way ANOVA, two-way ANOVA, linear regression).
- Correct for multiple comparisons using Benjamini Hochberg False Discovery Rate control methods
- Determine signatures by a variety of methods: logistic regression, PAM, SVM

Proteomics

- Quantitative proteomics by MS/MS is much more difficult since we never actually see the signal from a whole protein (in contrast to ELISA or Luminex where we do).
- The process of analysis is to take each peptide whose mass is measured in stage 1 and assign it to a protein, if possible, by the fragment spectrum in stage 2
- For each protein (and for humans we have a kind of a list in advance) we count the unique peptides that map to that protein and use that as the score for the protein.
- This can then be analyzed with (for example) overdispersed Poisson regression.