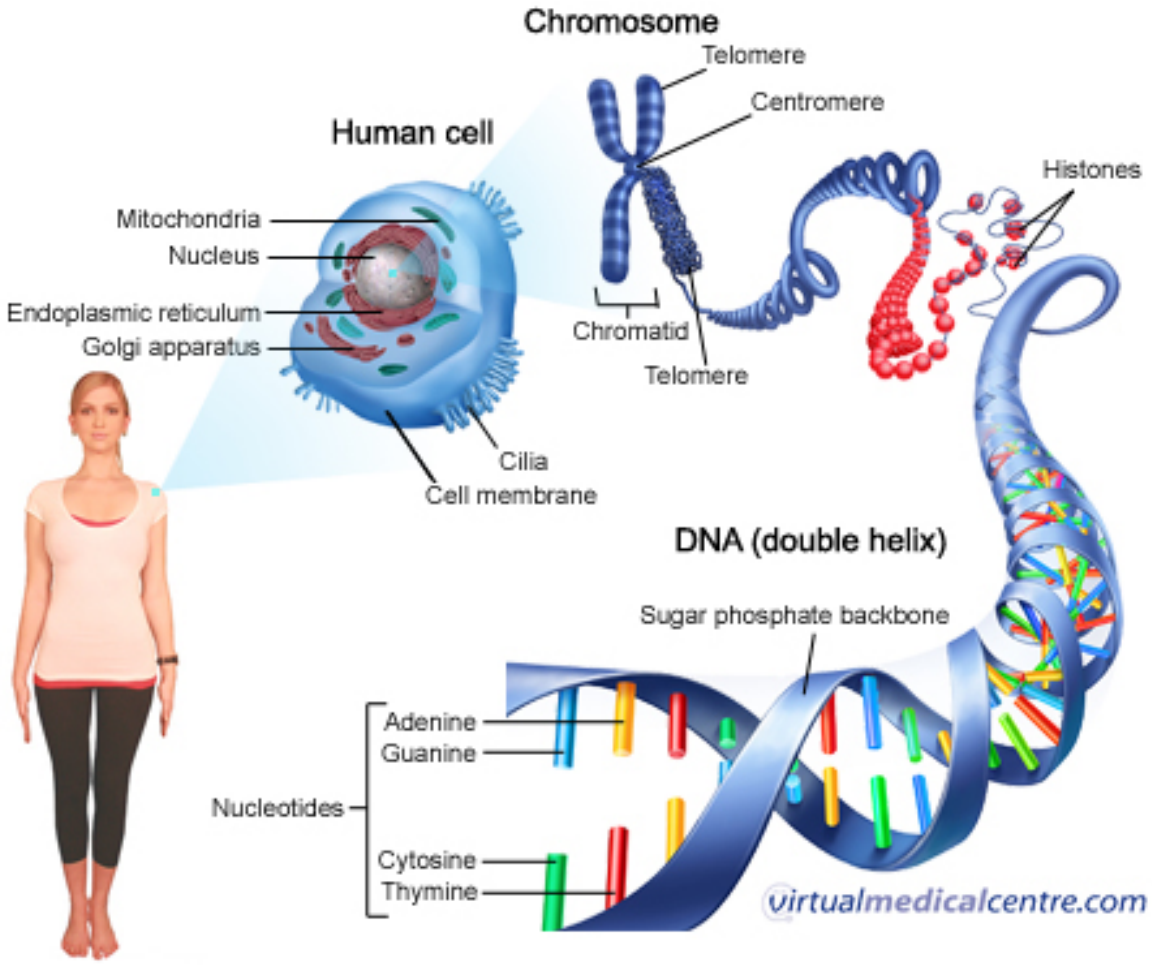
The Effects of Probe Sequencing On Microarray Gene Expression Measurements

Randy Williams, Pedro Muñoz-Agrinsoni, Kevin Kupiec
Dr. Rafael Irizarry
Ryan Sun
July 23rd , 2015

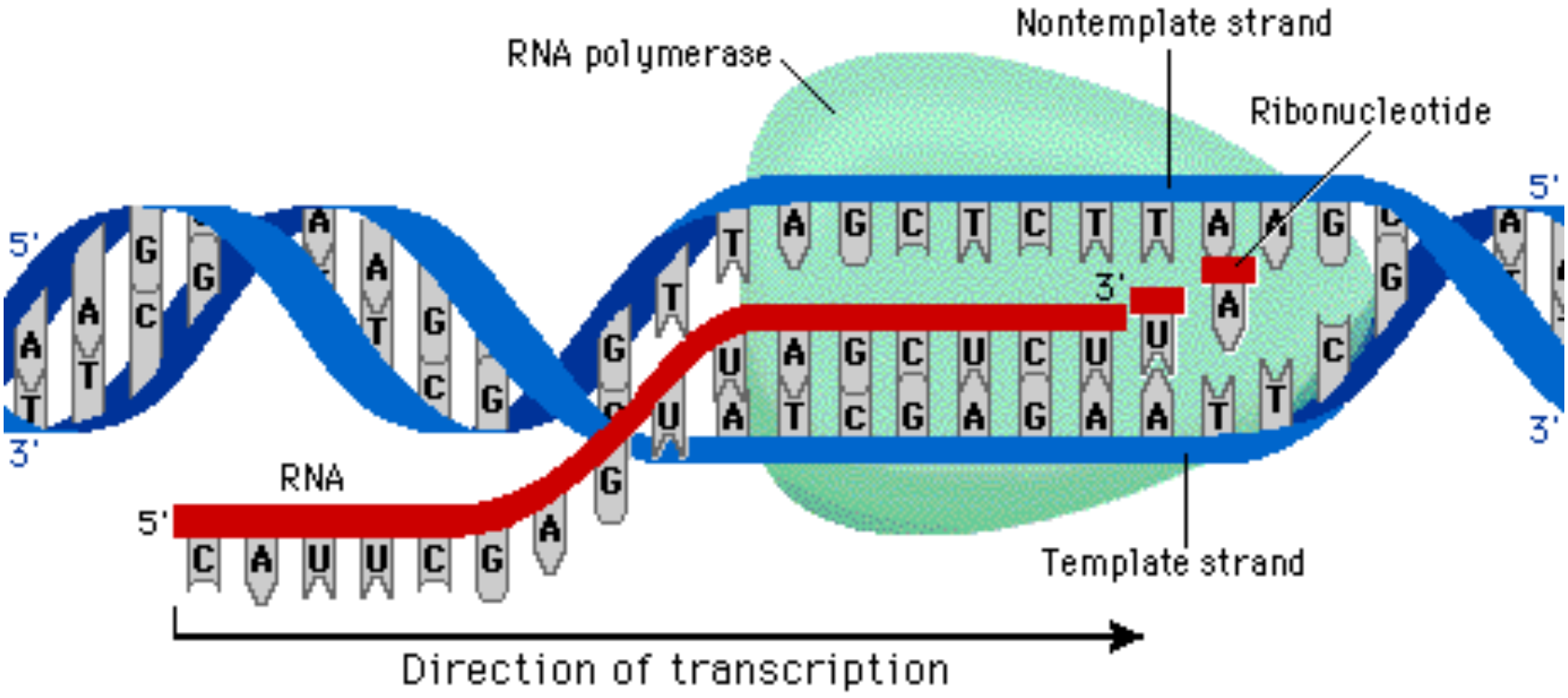
Background: Genes



Background: DNA & RNA

- DNA (Deoxyribonucleic acid) - molecule that carries and stores all genetic information on the functioning of living organisms
- Adenine (A), Cytosine (C), Guanine (G), & Thymine (T) are nucleotide base pairs attached to the "sugar phosphate backbone" that compose DNA
- RNA (Ribonucleic acid) - main function is to act as a messenger carrying instructions from DNA to synthesize proteins
- Differences: RNA is single stranded and it contains Uracil (U) instead of Thymine, an evolutionary energy conservation method due to its short-lived nature

Background: DNA & RNA (cont.)



Background: Probe Sequencing

- Probes are single-stranded sequences of DNA or RNA used to search for complimentary base pairs in a genome
- This process takes an oligonucleotide, short chains of DNA or RNA molecules, of length 25 that contain probes that perfectly match and represent a mRNA sequence by a probe set of length 11-20 probe pairs
- A mismatched probe is also used with the only difference being an intentionally mismatched 13^{th} probe used to measure non-specific binding

Background: Probe Sequencing (cont.)

Examples of Perfect Match and Mismatched Probes:

AACTGCTATCG
TTGACGATAGC

TTAGCGTGCAT
AATCGTACGTA

- Designed to hybridize only with transcripts from the intended gene
- Hybridization is the process of combining two complementary single-stranded DNA or RNA molecules and allowing them to form a single double-stranded molecule through base pairing.

Background: Microarray Data

- A microarray is an array of "immobilized single-stranded DNA fragments of known nucleotide sequence that is used especially in the identification and sequencing of DNA samples and in the analysis of gene expression."
- In our study, we focused on the intensity of gene expression that represents the amount of hybridization for each oligonucleotide probe

Our Question:

- Can we predict the effect a certain nucleotide base has on gene expression measurements?

Why study gene expression data?

- Gene expression reveals which genes are more active
- Active = gene that is being expressed in relation to a particular condition
- For example, if you can determine which genes are more active in a sick patient in relation to a healthy patient, you are able to pinpoint the specific gene associated with a particular disease or illness
- To attempt to find a more effective way in removing the effects, or noise, that the probe sequence produces on gene expression data

Genetic Data:

- Our microarray data is composed of a data frame of dimensions 604,258 x 618
- Each column represents a tissue sample from different parts of the human body including colon, cerebellum, kidney, lung, and others.
- Each row represents the probe intensity for each of the tissue samples
- We had to organize this data into matching probe sets of 11

Methods for Analysis:

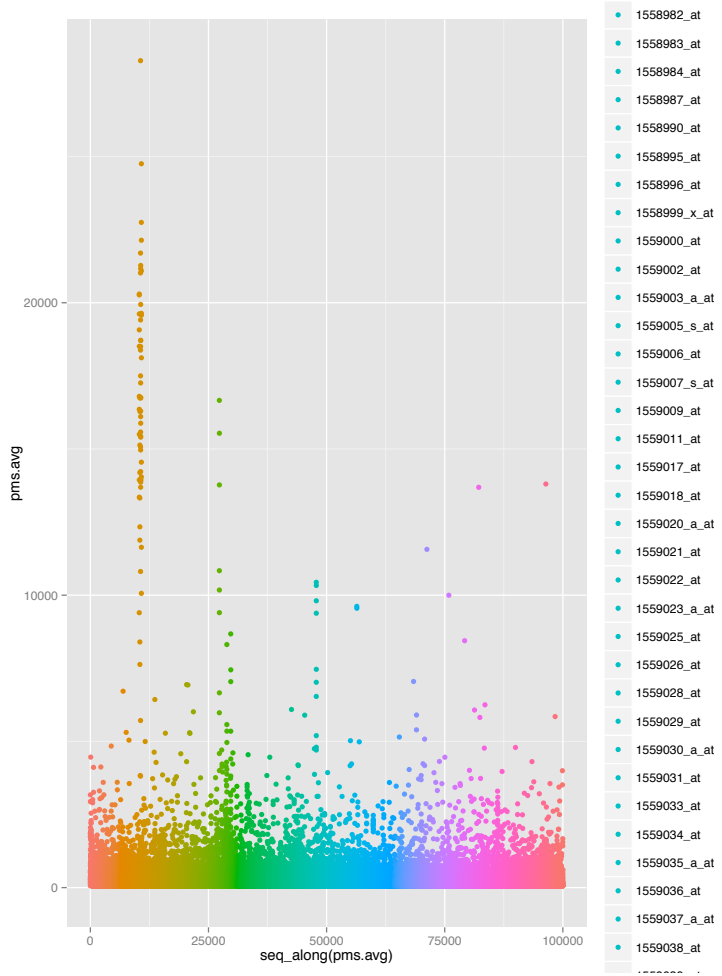
- Using "R" computer programming software, we had to download Bioconductor packages to read and interpret genetic data
- Trying to see how different 25 base sequences affect intensity in the gene expression data
- The data had probe sets of size 11 (sometimes 16) analyzing the same gene, each probe has a different sequence of nucleotides

Methods of Analysis (cont.):

- We wanted to see if the number of A's, T's, C's, & G's have an effect on the intensity of gene expression
- To do this, we represented the data graphically, then constructed linear models to determine whether a relationship exists

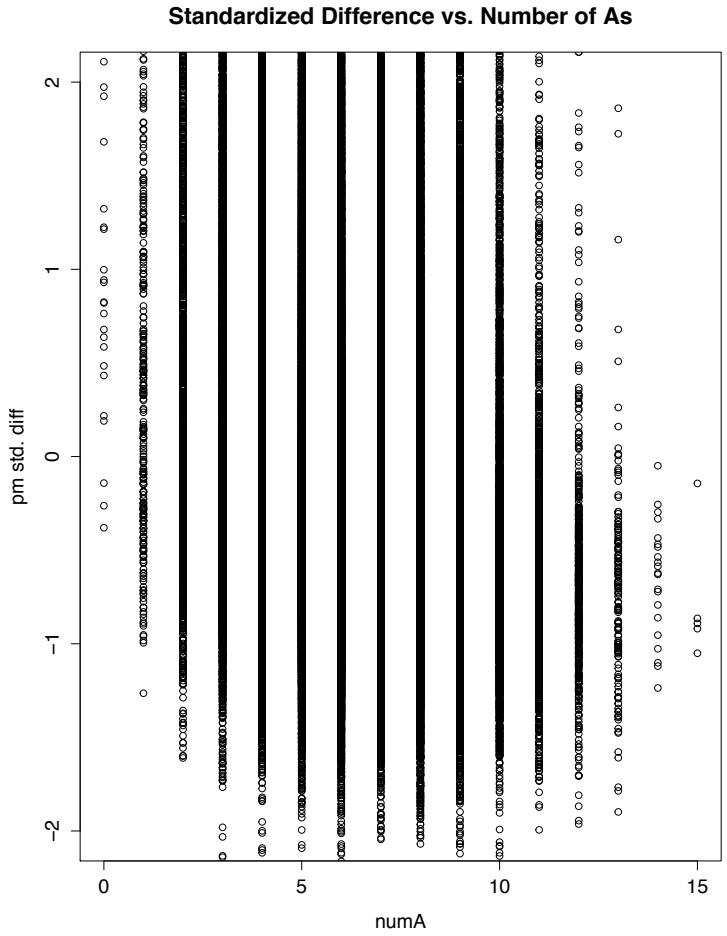
Graphical Results:

Probe Sequence Index vs. Average Intensity

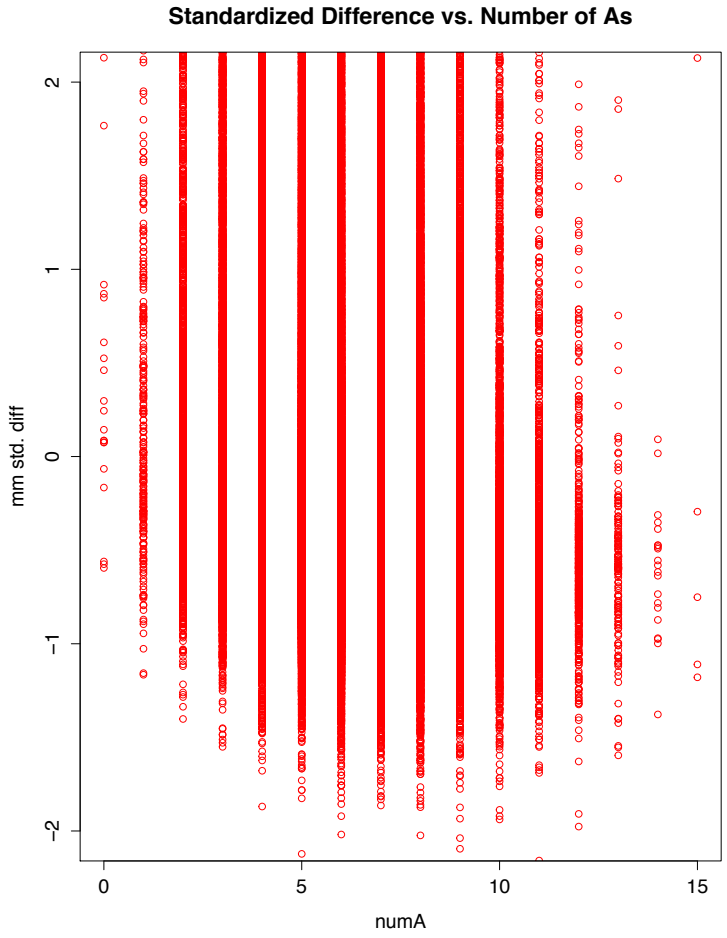


Graphical Results (cont.):

$Y\downarrow i - Y / SD(Y)$



Graphical Results (cont.):



Model:

- $y = \beta 0 + \beta 1 * x$
  - y = standardized difference
  - $\beta 0$ = y-intercept
  - $\beta 1$ = slope
  - x = number of A's, C's, T's, and G's

Implications:

- The p-values for the linear models of the perfect match probe sequence data and the mismatch probe sequence data were both less than .0001 in relation to an alpha value of 0.05
- This shows that there is enough evidence to suggest that the number of A's, C's, T's and G's in our sequence affect the probe intensity of our probe sequences.

Shortcomings:

- The size of our data (2.8GB) did not permit us to see the full extent of the gene expressions. We could only use a subset of the data or else our computers would crash.

Future Direction:

- Writing software that improves on current approaches to preprocessing microarray data.
- Current approaches to preprocessing microarray data include:
- **MAS5**: Microarray Analysis Suite 5 (MAS5) is an algorithm which uses aggregation and normalization on both perfect match and mismatch probes. It does really well in head to head tests and has a fair popularity.
- **RMA**: Robust Multi-array Average (RMA) is a normalization approach that does not take advantage of these mismatch spots, but still must summarize the perfect matches through median polish. The median polish algorithm, although robust, behaves differently depending on the number of samples analyzed.
  - **Median polish**: Is an exploratory data analysis procedure proposed by the statistician John Turkey. It finds an additively-fit model for data in a two-way layout table of the form row effect + column effect + overall median.
- There are also other methods that are used to preprocess microarray data.

Future Direction (cont.):

- Also, the statistical methodology could be extended to other technologies such as RNA-Seq.
- RNA-seq (RNA Sequencing), also called Whole Transcriptome Shotgun Sequencing (WTSS), is a technology that uses the capabilities of next-generation sequencing to reveal a snapshot of RNA presence and quantity from a genome at a given moment in time
- Last but not least, we could use as covariates not only a certain letter (A,C,T or G), but also the position of that certain letter in the sequence.

Thanks:

- We would like to say thank you to Dr. Rafael Irizarry, Ryan Sun, Tonia Smith, Dr. Rebecca Betensky, Heather Mattie, Eleanor Murray, Joshua Barback and all of the biostatistics faculty and students for this opportunity and experience.

References:

- http://www.phschool.com/science/biology_place/biocoach/images/transcription/startrans.gif
- http://www.myvmc.com/uploads/VMC/TreatmentImages/2437_dna_450_v2.jpg
- National Institutes of Health (NIH). (n.d.). Retrieved July 20, 2015.
- Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., & Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 909-917.