Bioinformatics, Data Analysis and Troubleshooting
Topics

• Bioinformatics
• Understanding Ct, Efficiency and Performance
• Review of SDS Data Files (demo and/or customer files)
• Q&A
Can I just cut and paste a sequence from NCBI into an assay design software?
How do Custom Assays compare to Pre-Designed Assays?

**Pre-Designed Assays**
- Repeat Masking
- SNP Masking
- Assay Design Pipeline
- Genome QC

**Custom Assays**
- Repeat Masking
- SNP Masking
- Genome QC
- Assay Design Pipeline

The customer prepares the sequence.
The customer submits the sequence.
Bioinformatics Steps

1. Biological Significance
2. Sequence Length
3. Sequence Quality
4. Masking Sequence
5. Uniqueness of Sequence
Bioinformatics Steps

1. Biological Significance
   - Know which transcript(s) you want to interrogate
   - Regions of high homology should be masked with “Ns”
   - Multiple exon genes → note the exon-exon junctions
     > MGB Probes should be designed over junction
       — No gDNA detected
Bioinformatics Steps

1. Biological Significance
2. Sequence Length
3. Sequence Quality
4. Masking Sequence
5. Uniqueness of Sequence
Sequence Length

• Submit a sequence length of ~ 600 bases
  — Range (300-5000)
  — Fewer than 300 limit assay design possibilities

• Select a sequence so that the target site is toward the center of the submitted sequence
Bioinformatics Steps

1. Biological Significance
2. Sequence Length
3. Sequence Quality
4. Masking Sequence
5. Uniqueness of Sequence
Sequence Quality

• Inaccurate sequences can lead to failed assays
  — Poor binding of primers and/or probes

• Use public databases with curated sequences to determine quality
  — RefSeq or dbSNP

• Target must be >30 bases from the 3’ and 5’ ends
Bioinformatics Steps

1. Biological Significance  
2. Sequence Length  
3. Sequence Quality  
4. Masking Sequence  
5. Uniqueness of Sequence
Masking Sequence

- Mask sequence using RepeatMaker (www.repeatmasker.org)
  - Ambiguous sequences
  - Repetitive sequences
  - SNP sites

Ambiguous sequence
(R=A or G) (S=G or C)
Masking Sequence

- Mask sequence using RepeatMaker ([www.repeatmasker.org](http://www.repeatmasker.org))
  - Ambiguous sequences
  - Repetitive sequences
  - SNP sites
  - Caution: can over mask!

NNNNNNNNNNNNNCGGCATTNNNNNTCCTGTCCGCAATAGC

Ambiguous sequence (R=A or G) (S=G or C)

Repetitive sequence
Bioinformatics Steps

1. Biological Significance
2. Sequence Length
3. Sequence Quality
4. Masking Sequence
5. Uniqueness of Sequence
Uniqueness of Sequence

- BLAST sequence against public databases
  - Detect regions within sequence that match other sequences
  - Mask or find a new sequence

- BLAST site
AB’s website has a protocol for guiding you through the necessary bioinformatics checks of your sequence.

Go to www.appliedbiosystems.com

Click on “Support”

Click on “Tools, Tutorials, Maintenance, and Troubleshooting”

Real-Time PCR / TaqMan® Genomic Assays

Getting Started with Real-Time PCR
Real-Time PCR vs. Traditional PCR (262k PDF)
Essentials of Real Time PCR. (159k PDF)
Selection of Reagents for Real-Time PCR (630k PDF)

Applications
Performing Relative Quantitation of Gene Expression Using Real-Time Quantitative PCR (20 Mb PDF)

TaqMan® Genomic Assays
Bioinformatic Evaluation of a Sequence for Custom TaqMan® Gene Expression Assays (1.5 MB PDF)
Bioinformatic Evaluation of a Sequence for Custom TaqMan® SNP Genotyping Assays (4.31 MB PDF)
Questions?
Understanding $C_T$

- What causes a shift in Ct values?
  - Threshold setting
  - Master mix and sample composition
    - Increase in pH and salt can lead to lower emission of fluorophore
      - No real impact on sensitivity
Understanding $C_T$

- What can cause $C_t$ shifts?
Understanding $C_T$

**ROX™ Passive Reference Dye**

Greatly *improves precision* of replicates

\[ R_n = \text{Normalization} = \frac{\text{Reporter}}{\text{Reference}} \]
ROX™ dye = Precision

36 Replicates analyzed with ROX™ passive reference dye

36 Replicates analyzed without ROX™ passive reference dye
Understanding $C_T$

• Lower ROX™ = higher baseline $R_n = lower \Delta R_n = shift in Ct value (lower)$

• No bearing on sensitivity, but does have unintended consequences
Understanding $C_T$

- How does ROX™ affect precision?
Questions?
Too efficient PCR

• Why is the PCR efficiency 151.5%?
Performance of reaction

- Efficiency range from 70-168% when testing a dilution series of a single log, due to the standard deviation in one dilution
Too efficient PCR

- Why is the PCR efficiency 151.5%?
**Performance of reaction**

- $R^2$ is a statistical term that says how good one value is at predicting another.
- If $R^2$ is 1 then you can perfectly predict the value of $X$ (quantity) with the value of $Y$ (Ct).
Standard Curve

With Outliers

Slope = -3.199
$R^2 = 0.949$

Without Outliers

Slope = -3.277
$R^2 = 0.999$
Low efficient PCR
Slope of standard curve is – 4.03

Reason 1:
- Long PCR products (>150 bp)
- Insufficient primer and/or probe concentrations
- Low primer hybridization and extension efficiency

Solution:
- Shortening PCR amplicon
- Re-optimize assay (MM Protocol)
- Redesign primers, shift for a couple of bases

Reason 2:
- Error in standard curve generation
- Range of template input too small
- Pipetting errors

Solution:
- Use broader range of template input
- Calibrate the pipettes
- Pipette large volumes, use dedicated pipettes
Too efficient PCR
Slope of standard curve is $-2.7$

**Reasons:**
- Outliers
- Pipetting errors
- Some dilutions did not amplify (too little material)
- Some dilutions show inhibition (too much template)

**Solutions:**
- Remove outliers and/or negative reactions
- Perform dilution series doing enough dilutions in replicates
- Omit dilutions showing inhibition to get correct slope
- Optimize nucleic acid extraction and RT reaction
- Pipette larger volumes, use dedicated pipettes
Performance of reaction

• Precision
  - The greater the standard deviation of the replicates
    > The decreased ability to discriminate small fold changes
    > If a PCR is 100% efficient, there is one Ct between the mean of a 2-fold dilution
    > To be able to quantify a 2-fold dilution in more than 99.6% of cases, the standard deviation has to be $\leq 0.167$ ($1 \div 6 = 0.167$)
Questions?
Standards and 1 Sample

- What is causing varied melt curves for the replicates?
Variability Analysis

- Two Variability Analysis Methods determine the Error Bars for RQ:
  - Standard Deviation Based Method (SD)
  - Confidence Interval Based Method (CI)

- **SD Method:** Assigns variability estimate for computed RQ values according to precision of the experiment. Additional data points do not necessarily correspond to an improvement in the variability estimate.

- **CI Method:** Assigns variability estimate for computed RQ values according to the accuracy of the computations. Additional data points theoretically correspond to an improvement in the variability estimate.

- CI Method statistically accounts for typically small sample sizes found in RQ Assays and is found in all software versions. SD simplifies the estimates and is found only in StepOne and 7500 v2.x versions.
Selecting a Statistical Test

Parametric tests (assumptions)

- How to compare a group of samples with a mean?
  - Simple T-test
- How to compare two group of samples?
  - Unpaired T-test
- How to compare a group of samples before and after treatment?
  - Paired T-test
- How to compare three or more groups?
  - 1 & 2 way ANOVAs
- How to study the effect of treatment?
  - Pearson correlation
  - Simple linear regression
Selecting a Statistical Test

Non-parametric tests (no assumptions)

Wilcoxon rank sum test for paired data: compares a group of samples before and after treatment

Mann-Whitney test: compare two groups of impaired samples
You measure the gene expression changes in liver after treatment of 40 rats for 2 weeks with one known carcinogen and 40 rats for 2 weeks treated with one non-carcinogen.

**Independent variable:** known carcinogen and non-carcinogen.

**Dependent variable:** gene expression changes in liver

Statistical test you would use: **Unpaired t-test**

Use this test to compare the mean values (averages) of two sets of data.
You look for a relationship between the TP73 gene expression of 40 oligodendrogliomas patients before and after chemotherapy, to evaluate the effectiveness of this treatment.

**Independent variable:** chemotherapy

**Dependent variable:** TP73 gene expression

Statistical test you would use: **Correlation (statistics: r² and r)**
You measure the gene expression of 40 onion plant genes given 3 types of pesticides and 2 types of fertilizers. Estimate if there are significant differences on the expression of these genes depending of these factors.

**Independent variable:** pesticide and fertilizer

**Dependent variable:** plant gene expression

Statistical test you would use: **2-way ANOVA** to evaluate the individual influence of pesticide, fertilizer and the interaction of both factors.
You read a scientific paper that claims that the average gene expression level for PLA2G2A is 32 times higher in gastric tumor compare to non-tumor tissues. Now, you want to check whether or not this is true in your samples:

- **Predicted value of the variable variable**: 32 times higher
- **Variable under study**: gene expression level for PLA2G2A
- **Suitable statistical test**: Simple t-test

Use this test to compare the mean values (averages) of one set of data to a theoretical mean value.
You want to estimate the prognosis of colon cancer patients using 70 samples’ data for SLC5A8 gene expression and survival time.

**Independent variable:** SLC5A8 gene expression  
**Dependent variable:** survival time  
Statistical test you would use: **Linear regression (statistics: r² and r)**  
Fit a line to data having only one independent variable and one dependent variable.
You measure gene expression in 50 monkey kidney cells before and after treatment with the mycotoxin fumonisin B for the identification of differentially expressed genes following treatment.

**Independent variable:** mycotoxin fumonisin B (before and after treatment)

**Dependent variable:** differentially expressed genes following treatment

Statistical test you would use: **Paired t-test**

Use this test to compare data from the same subjects under two different conditions.