Acknowledgements

- **Molecular Profiling**
  - Robert Phillips, Guided Solutions (now MP Lead, Oncology) and Alan Sachs
  - Greg Opiteck, Franchise Integrator
  - Pek Lum, Eric Schadt, Cliona Molony, Solly Sieberts, Genetics
  - Silvija Coulter and Michele Thornton, MP Program Management
  - Vijay Modur, Lori Roadcap, and Paul Deutsch, Clin MP
  - Chris Roberts, Spring Cheng, Hongyue Dai, Yudong He, and Richard Hu, Custom Analysis
  - George Tokiwa and Matt Marton, GEL R&D (developed protocol for preparing RNA from human skin in RNAlater)
  - GEL production (Jon Williams and Debbie Kessler)

- **Bone Biology**
  - Jim Ray, Shunichi Harada, Pascale Nantermet, Len Freedman

- **Clinical Pharmacology**
  - Aubrey Stoch,
  - Eugene Marcantonio
  - John Wagner

- **BARDS**
  - Xiang Yu
  - Peter Hu
  - Dan Holder
  - Yue Wang
Search for a Biomarker

1. Better Drug
2. Targeted population

1. Validated Targets (mechanistic proof)
2. New Targets

Molecular Profiling
- Genes
- Gene Expression
- Proteomics

Imaging

Lipids/Proteins

Histology

Discovery

Proof of Concept Clinical Trial

Clinical Trials

- 1. Better Drug
- 2. Targeted population
Outline

- Gene Expression in a Proof of Concept Clinical Study
- Genetics in the Clinical Study
- Summary
Use of Gene Expression – Androgen Biomarker Study Example

- Androgen receptors are widely expressed on different tissues and androgens physiologically exert diverse actions in women
- Androgen replacement therapy (ART) used to reduce menopausal symptoms and improve quality of life in postmenopausal women
- Dose and duration of ART ‘capped’ given concern of undesired mechanism-based consequences

Study Objectives
- Pilot pharmacodynamic measures of short-term androgen administration that may be used to predict long-term effects in postmenopausal women
- To qualify skin biomarkers suitable for assessing the androgen-mediated effects in postmenopausal women
History

- **Endpoints**
  - Hormone parameters (testosterone, estradiol, etc.)
  - Lipid parameters (cholesterol, HDL, LDL, TG)
  - Other (Ferriman-Gallwey scores, acne scores)
  - Bone biomarkers
    - Functional Measure: Sebum excretion rate
    - Structural: Sebaceous gland volume (skin biopsy)
    - Molecular: Gene expression changes (Taqman and microarray)

- Taqman panel started from preclinical model in Rhesus Monkeys
Androgen Biomarker Study Example: Pilot Study Design

Placebo (n=13)  Purell® Instant Hand Sanitizer

Androgel® 300 μg (n=13)

Androgel® 2.5 mg (n=10)

Baseline
Sebum and Skin Biopsy

Week 2
Skin Biopsy

Week 6
Sebum and Skin Biopsy

Week 10
Skin Biopsy

- Double-Blind, Randomized, Placebo-Controlled, Parallel-Group Design Study
- N=36 postmenopausal female volunteers
Link to Clinical Endpoints

Dose Dependent Increases in Free Testosterone (pg/mL)

Geometric Mean (pg/mL) ± SE

Study Week

Lipid Changes Observed at Week 6

Geometric Mean % Change ± SE

-20 -15 -10 -5 0 5 10 15

Total C  HDL-C  LDL-C

Placebo
300 µg AndroGel®
2.5 mg AndroGel®
A Dose-dependent Signature of Testosterone is Identified in Human skin at 2 weeks

Anova Between 2.5 mg and Placebo, p<0.05
Fold change > 1.5 in 2.5 mg treatment compared to Baseline pool
Fold change > 1.5 in 2.5 mg treatment compared to Placebo treatment

**Magenta:** up-regulated; **Cyan:** down-regulated transcripts

**2.5 mg Androgel**

**0.3 mg Androgel**

**Placebo**
Bone biomarker Data

Dose Dependent, Reversible Increases Sebaceous Gland Volume

Dose Dependent Increases in Sebum Excretion

* p-value < 0.05 for change from baseline vs. placebo

%Change from Predose ± SE

- n=13 Placebo
- n=12 300ug Androgel
- n=10 2.5mg Androgel

p=.002
Gene Regulation with Treatment at Weeks 2, 6 and 10
Difference from Placebo (Least Squares Mean ± 95%)

Note:
1. LSM difference estimated from ANCOVA model with terms treatment and 18S rRNA normalization factor
2. CT = Cycle Threshold. CT represents the number of cycles used to reach exponential amplification in quantitative PCR. One less cycle (ΔCT = -1) indicates roughly 2-fold more starting transcript.
Conclusions from the Pilot

- Androge® generally well tolerated
- 2.5-mg Androge® significantly increases pilosebaceous unit volume and function in postmenopausal women
  - Clinical efficacy confirmed
  - Gene expression changes seen by week 2
- Effects reversible 4 weeks post-discontinuation of therapy
- Overall, biomarkers can be used to test androgen-like compounds for mechanism-based skin effects in postmenopausal women
- Note that the qPCR had to be rebuilt from the animal models.
What Next?

- 6 gene signature qPCR
  - Validated with a new entity where none of the genes turned on but there was histological, clinical confirmation that gene signature was working.
  - Some genes up and some genes down – what does this mean?
  - Missing ‘housekeeping’ genes to confirm successful taqman experiment.
  - Look at them individually? Build a ‘composite’ or ‘predictive’ score?

<table>
<thead>
<tr>
<th>Gene</th>
<th>Clinical Confirmation</th>
<th>AEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Gene Linked to AE

Risk outweighed by clinical benefit?

CONTROL

Y
What does that Prediction Score Include?

Reduced Menopausal Symptoms

Gene/Pathway 1
Gene/Pathway 2
Gene/Pathway x

Gene/Pathway x+1
Gene/Pathway y

Weighted by species, and importance of pathway in leading to dx state

Known Clinical Co-morbidities

Younger Lean Non-smoker

Other Clinical or Biomarker Endpoints

Proteins, Lipids, Histology

Perhaps Genes or Pathway sufficient to predict alone

Older Obese Smoker

Proteins, Lipids, Histology
Prediction Score Gradient

For example, better QoL prediction score =
gene pathways \((1 + 2 + 3 + 4 + 7)^*\) *
(# known clinical factors)

- Not tied to primary event
- Difficult to build

- Lower Score indicates better Quality of Life
- Positive control
- Lead to Go decision?
- Other Marked Drugs
- Failed NCE
- Ambiguous Zone
- Higher Score more likely to lead to menopausal symptoms
- Not tied to primary event
- Difficult to build
How to set up next clinical study?

- When do we stop doing microarrays and move forward to qPCR and right into testing new clinical entities with the new biomarker without also doing the histology?
  - Ethical challenge to run positive control with new chemical entities, no gold standard for Rx
  - How predictive is enough in discovery versus utility in a phase III program?
    - Sensitivity, PPV, Specificity, NPV
    - Fit for Purpose
  - Need to change platforms from biopsy to blood – operational ease
Power Calculations for Gene Signature

- Conservative case – use highest variable gene
- Average Intensity across the genes?
- Assumptions about correlation among the genes
- Assumptions from animal or cell line to human subjects
- What effect size is necessary to link to endpoint?
Simplified Ballpark ‘Power’

Caveats:

1. Mouse to human data
2. Change from hypothalamus to salivary gland
3. Assumed variability is 0.05 SD based on mouse data
4. Assume average fold change for placebo is 1.3, low dose is 1.44 and high dose 2.0. Note that 1.3 is minimum detectable ratio above noise.
5. $\alpha=0.05$, 8 per treatment group

Assume that humans will have the same pattern, correlation and variability as mice

<table>
<thead>
<tr>
<th>SD</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>&gt;99%</td>
</tr>
<tr>
<td>0.42</td>
<td>~80%</td>
</tr>
<tr>
<td>0.47</td>
<td>~70%</td>
</tr>
</tbody>
</table>

- Assume human variability 8 times or 9 times larger
- Know humans are more heterogeneous
- Note that by increasing the upper fold change to 2.1, a SD of 0.47 would give ~80% power
- Some tradeoff between SD and detectable fold change
QuickChip – Genetic Analyses

- Correlate gene expression to genetic DNA reduce SNPS [Zhu & Schadt, 2007]
- Reduce to ~6500 SNPS + drug metabolism genes and other candidate genes

**Goal: Differential Response to Treatment**
- Better Responder – high efficacy and/or
- More Prone to Adverse Experiences
Examples of Pharmacogenetic Interaction

Overall Rx difference from placebo = -5.5 kg in all cases
Analyses Diagram

**Full Data**

- Pre-selected 394 Markers + 3 Clinical Covariates
  - Criterion: (1). P-value for G X T interaction < 0.05 and (2). MAF > 0.25

**Random Forest**

- Top 394 Predictors + 3 clinical covariates

**Training Set**

- Top 30 Predictors

**Recursive Partitioning Regression Tree**

- Cross - validation Output

**Test Set**

- Cross - validation Output

**All 13698 markers + 3 clinical covariates**

- Top 30 Predictors
Next Steps

- Validate with 2\textsuperscript{nd} efficacy endpoint
- Move forward into a new population with composite biomarker to test for hyper-responders vs low responders
- Move to Genome Wide Association
  - Connect to PK/PD
- How to use Adaptive Design to enroll best number of samples.
Network/Pathway Analyses:
TGF-β Signaling Pathway

Schadt, Sachs, Friend. Sci. STKE, 2 August 2005
Network/Pathway Analyses

- Underlying premises - Bayesian Network (Zhang & Liu, Nature Genetics, 2007)

- How do we power this pathway?
  - From central node – weight for significance of overlap?
  - Simple tables to indicate overlap between animal and human?

- Calculation of n for POC study in humans?

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>Overlap between Human and Non-human Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Treatment Group</td>
<td></td>
</tr>
<tr>
<td>Treatment A: Treated</td>
<td>A</td>
</tr>
<tr>
<td>Treatment B: Placebo</td>
<td>C</td>
</tr>
</tbody>
</table>
References

Back-Up Slides
Candidate Gene Approach: Background

**Goal: Characterize differential response to treatment**

- Genetic covariates: 2453 SNPs located in 114 genes;
- Patients (~240 genotyped patients) on Placebo (N=93) and efficacious treatment pooled from two Phase II studies
  - P-value for treatment * genotype = 0.008, n=228
- Most significant hits were driven by rare alleles
- Additive Model Assumed
- Multiplicity Adjustment for Single SNP: False Discovery Rate
Candidate Gene Approach: Real Example

Minor Allele Frequency = 38.9%
Thus, Expected; actual on Drug
1/1 = 15%; 16%, n=21
1/2 = 48%; 49%, n=66
2/2 = 37%; 36%, n=48

Patient Treatment Response by Genotype

Non-responders

Typical Efficacy

Hyper-responders
Candidate Gene Approach: Treatment and Active Control

Genotype

Clinical Response (% Chg from Placebo)
Mean ±95% CI

Empty symbol = active control
Solid symbol = test treatment
Candidate Gene Approach: Conclusions/Summary

- Interactions
  - SNP analyses – most interesting significant result in example
  - Haplotype analyses – no significant interactions

- Exploratory

- Active Control Arm
  - Control treated patients only in one Phase II study, N=39
  - small sample size

- Needs Replication - Validating the significant gene by treatment interaction result in additional clinical trials.
How Do We Tackle the Challenges?

- Objective: To build a conditional parametric regression tree model $Y \sim X_1 + \cdots + X_k \mid \{Z_1, \ldots, Z_m\}$

- Advantages
  - Easy to interpret;
  - No need to model explicitly how markers interact with each other;

- Negatives:
  - Trees are prone to bias and over-fitting;
Why *randomForest*?

- One of the most robust, accurate and fast algorithms in data-mining literature ([http://www.clopinet.com/isabelle/Projects/NIPS2003/analysis.html](http://www.clopinet.com/isabelle/Projects/NIPS2003/analysis.html))
- Mechanism proved to be unbiased and no over-fitting;
- Therefore it’s an ideal platform to test the influence of sparse data due to low MAF on variable selection.
What Do We Suspect?

- Accuracy of `randomForest` depends on correlation between inputs and the strength of individual tree
  - The weaker the correlation and the stronger the individual trees, the more accurate;

- What can be said about the ~14000 markers?
  - Low power in global adjustment for FDR during evaluation of individual markers
    - 99% of the markers are probably just weak classifiers individually, if cannot serve as classifiers at all;
  - Many markers with low MAF, causing individual trees unstable;
  - Low MAF drives between-marker correlation towards high.
### Composite Genetic Marker Signature

<table>
<thead>
<tr>
<th>Marker ID</th>
<th>Candidate Gene</th>
<th>Rank of Importance@</th>
<th>P-Value*</th>
<th>MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Known</td>
<td># 1</td>
<td>0.0328</td>
<td>0.32</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>#3</td>
<td>0.0233</td>
<td>0.30</td>
</tr>
<tr>
<td>3</td>
<td>Known</td>
<td># 4</td>
<td>0.0039</td>
<td>0.28</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td># 6</td>
<td>0.0044</td>
<td>0.40</td>
</tr>
<tr>
<td>5</td>
<td>?</td>
<td># 15</td>
<td>0.0144</td>
<td>0.29</td>
</tr>
<tr>
<td>6</td>
<td>?</td>
<td># 22</td>
<td>0.0177</td>
<td>0.45</td>
</tr>
<tr>
<td>7</td>
<td>?</td>
<td># 26</td>
<td>0.0100</td>
<td>0.29</td>
</tr>
</tbody>
</table>

@ After the 3 clinical covariates are removed

* For genotype-by-treatment interaction in individual marker evaluation

? Not annotated to the genomic regions of any known candidate genes but linked to gene expression profiles in liver
Cross-validation

- Strategy:
  - Multiple training:test ratios;
  - At each fixed rate of training:test, multiple versions training samples were drawn.

- A Useful Decomposition:

\[
PE(Y_{test}, \hat{Y}) = Var(Y_{test}) + MSE(\hat{Y}, Y_{test}) \equiv \begin{cases} 
Var(\hat{Y}) \hfill \\
irrducible \\
\end{cases} + \\
(E\hat{Y} - EY_{test})^2 \\
bias^2 \\
\]

where \( \hat{Y} \) is the estimator, i.e., the tree model
Cross-validation for the Tree Model Derived from the 394 Pre-chosen Markers

1.05 1.10 1.15 1.20 1.25 1.30
Prediction Error and Test Data Variance
1:2 1:1 2:1 3:1 4:1

Model Variance
Absolute Bias

Training:Test Ratio in Sample Sizes

Prediction Error
Test Data Variance

Cross-validation for the Tree Model Derived from the 394 Pre-chosen Markers

Model Variance
Absolute Bias

Training:Test Ratio in Sample Sizes
Conclusion & Future Direction

Conclusions:

- Even after adjusting for bias, the hyper-responders characterized by the composite genetic marker continue to have strong efficacy compared to the overall efficacy.

- A balanced point between bias and model variation is found with a training to test ratio of 3:1.

- Markers with low MAF are more likely bring in more noises than informative signals during tree construction.

Future Direction:

- Need to validate using independent data.
- Try other powerful tools such as boosting trees.
References

- J. A. Sparano and S. Paik
  Development of the 21-Gene Assay and Its Application in Clinical Practice and Clinical Trials

- K. K. Dobbin, Y. Zhao, and R. M. Simon
  How Large a Training Set is Needed to Develop a Classifier for Microarray Data?

- K. K. Dobbin and R. M. Simon
  Sample size planning for developing classifiers using high-dimensional DNA microarray data

- P. Liu and J. T. G. Hwang
  Quick calculation for sample size while controlling false discovery rate with application to microarray analysis