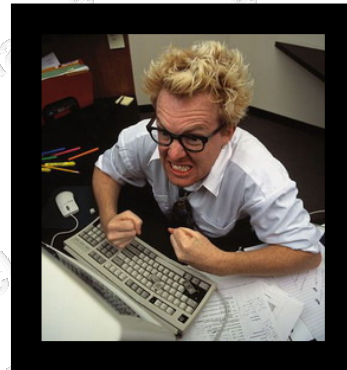
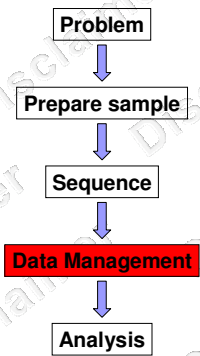


**ABRF GVRG 2009**

A Review of Heterozygous SNPs  
Detected in *S. cerevisiae* Using  
Multiple Next-Generation Sequencing  
Platforms

Disclaimer  
Disclaimer  
Disclaimer  
Disclaimer  
Disclaimer





## How it started

- 454 Life Sciences & Baylor College of Medicine published an analysis of Jim Watson's genome (Nature, 2008)
  - Identified 3.3 million SNPs
  - >600,000 were previously unknown
  - ~10,500 of these SNPs cause amino acid substitutions
  - >200,000 insertions/deletions and copy number variations
- Quality of next-generation sequence data is inherently lower than Sanger and other established genotyping methods
- Accuracy is crucial if a goal is personalized medicine


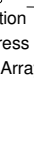


## GVRG proposal

- Compare genotyping data collected from current next-generation platforms to data collected from established genotyping methodologies.
  - Use wild-type yeast strain previously sequenced by the Stowers Institute (Li *et. al.*, Cell 2008)
  - Limit to exon regions by utilizing Nimblegen's Sequence Capture System
  - Look at 800+ SNP sites previously described in the Li paper



## Proposed Instrumentation

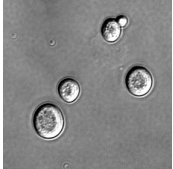
- |   |   |                                      |
|---|---|--------------------------------------|
| <ul style="list-style-type: none"> <li>□ Illumina GA</li> <li>□ Roche 454</li> <li>□ AB SOLiD</li> </ul>  |  | Generate and analyze data            |
| <ul style="list-style-type: none"> <li>□ Illumina Beadstation</li> <li>□ Illumina BeadXpress</li> <li>□ Sequenom MassArray</li> <li>□ AB 7900HT</li> <li>□ AB 3730</li> <li>□ Affymetrix</li> </ul> |  | Design assays based on next-gen data |



## GVRG "Reality"

- Illumina GAI – Li *et. al.*, Cell 2008
- Illumina GAIi
- AB SOLiD
- AB 3730 & 3730XL
- Sequenom MassArray

**Saccharomyces cerevisiae**



Kingdom: Fungi  
Phylum: Ascomycota  
Subphylum: Saccharomycotina  
Class: Saccharomycetes  
Order: Saccharomycetales  
Family: Saccharomycetaceae  
Genus: *Saccharomyces*  
Species: *S. cerevisiae*

- ~12 million base pairs
- 80% coding
- Diploid strain
- 16 chromosomes
- ~6,200 genes
- gDNA was extracted from a homogenous exponentially growing culture

## Next-Generation sequencing platforms

- **Illumina Genome Analyzer II**
  - Over 3 GB for single-read, 6 GB for paired-end (1 flow cell)
  - Up to 90 million single reads of 36 bases
  - Sequencing by synthesis chemistry utilizes fluorescently labeled, reversible nucleotide terminators
- **AB SOLiD 2.0™**
  - Up to 4 GB for single-read, 6 GB for mate-paired (2 slides)
  - Up to 120 million reads of 35 bases
  - 200-240 million reads of 2 x 25 bases
  - Sequencing by ligation chemistry utilizes fluorescently labeled hybridization probes

## Data Analysis Software

- **Maq**
  - Used for Illumina and SOLiD data
  - Maps to reference genomes
  - Uses quality scores to derive genotype calls
  - Makes full use of mate-pair & paired end data
  - Small indel detection
  - SNP detection
- **Corona Lite**
  - Used for SOLiD data
  - Supports mapping to reference genomes
  - Mapping against sequence databases
  - Pairing for mate-pairs
  - Small indel detection
  - SNP detection

## Illumina Genome Analyzer sequencing

- **Genome Analyzer I**  
(Li *et. al.*, *Cell* 2008)
  - 200 bp single-read genomic library
  - Sequenced by Stowers Institute
  - Analysis by Stowers Institute
- **Genome Analyzer II**
  - 400 bp single-read genomic library (U. of Missouri)
  - 4 lanes
  - Sequenced by University of Delaware and Illumina
  - Analysis by Stowers Institute



## AB SOLiD™ sequencing

- Genomic mate-pair library
- 3 quadrants
- Prepped and sequenced by University of Florida
- Analysis by AB and Stowers Institute



## Sequencing Statistics

| Sequencing technology | Total number of reads (millions) | Total sequence | Average aligned sequence coverage |
|-----------------------|----------------------------------|----------------|-----------------------------------|
| Illumina GAI          | 31.1                             | 1.12 GB        | 73x                               |
| Illumina GAI          | 18.1                             | 651 MB         | 43x                               |
| AB SOLiD™             | 122.3                            | 3.01 GB        | 147x                              |



## Depth-coverage statistics

| Depth-coverage | Illumina GAI | Illumina GAI | AB SOLiD™ |
|----------------|--------------|--------------|-----------|
| % with ≥ 1     | 97.8%        | 99.9706%     | 99.9780%  |
| % with ≥ 4     | 97.6%        | 99.9318%     | 99.9585%  |

| AB                 | Positions 0 | Positions < 4 | % with > 0 | % with > 3 |
|--------------------|-------------|---------------|------------|------------|
| <b>Quacrania</b>   |             |               |            |            |
| bc3,06,211         | 3673        | 3045          | 99.9780%   | 99.9585%   |
| bc3,08             | 3368        | 3442          | 99.9723%   | 99.9585%   |
| 113                | 2942        | 3202          | 99.9694%   | 99.9605%   |
| 51                 | 4394        | 3221          | 99.9585%   | 99.9605%   |
| 96                 | 4049        | 3227          | 99.9687%   | 99.9488%   |
| <b>Illumina</b>    |             |               |            |            |
| Leads 4,6,7        | 3372        | 3292          | 99.9706%   | 99.9718%   |
| <b>Gen</b>         | 899         | 3265          |            |            |
| <b>Genome Size</b> | 12154674    |               |            |            |

Maq generated SNP list was post-filtered according to criteria related to depth and repetitiveness by means of custom scripts written at Stowers



The Data





## SNPs Identified

|                    | Total | Heterozygous |
|--------------------|-------|--------------|
| Illumina GAI / Maq | 1035  | 73           |
| Illumina GAI / Maq | 1115  | 166          |
| AB SOLiD / Corona  | 1271  | 253          |
| AB SOLiD / Maq     | 906   | 97           |

809 SNPs are in agreement among the systems, of these  
23 are heterozygous  
(this does not include the SOLiD / Maq data set)

## Validation of the SNP data

Sequenom MASSarray to assay all the 1035 SNPs identified by the Illumina GA1

Sanger sequencing to go after SNP's near or in repetitive sequence

Allelic Discrimination Assays to cover any SNP's that are still left ambiguous

## Sanger Sites

| Primer | SNP         | REP | Ref | Expected | Sanger Call |
|--------|-------------|-----|-----|----------|-------------|
| 1      | I:70874     | NR  | A   | G        | R           |
| 2      | II:38800    | NR  | C   | A        | A           |
| 3      | II:241359   | NR  | G   | A        | A           |
|        | II:241396   | NR  | G   | A        | A           |
| 4      | III:358496  | IR  | G   | G or T   | N/A         |
|        | IV:1525459  | IR  | T   | G or T   | N/A         |
| 5      | V:308984    | NR  | G   | T        | T           |
|        | V:309047    | NR  | G   | C        | C           |
| 6      | V:434284    | NR  | C   | T        | T           |
| 7      | VI:58052    | NR  | G   | A        | A           |
|        | VI:58035    | NR  | G   | A        | A           |
| 8      | VII:203957  | NR  | C   | A        | A           |
| 9      | X:99469     | NR  | C   | G        | G           |
| 10     | X:687986    | NR  | A   | G        | G           |
| 11     | X:715089    | IR  | G   | A        | A           |
| 12     | XI:457775   | NR  | C   | T        | T           |
| 13     | XII:699663  | NR  | G   | A        | A           |
| 14     | XIII:808976 | NR  | A   | T        | T           |
| 15     | XIII:809197 | IR  | A   | G        | G           |
| 16     | XIII:837514 | IR  | A   | G        | G           |
| 17     | XIII:837771 | NR  | C   | A        | A           |
|        | XIII:837797 | NR  | T   | G        | G           |
| 18     | XIII:851734 | IR  | G   | A        | A           |
|        | XIII:851740 | IR  | G   | A        | A           |
| 19     | XIV:359024  | IR  | C   | T        | T           |
| 20     | XV:60240    | NR  | C   | T        | T           |
| 21     | XVI:338827  | NR  | T   | C        | C           |
| 22     | I:172042    | NR  | G   | T        | ?           |
| 23     | VII:530038  | IR  | A   | A or C   | C           |
| 24     | XV:954166   | NR  | G   | A or G   | G           |

## Sanger sequencing

- 30 chosen SNP sites were near, or in, a repeat. Based on Stowers yeast sequence.
- 3 SNP sites could not be sequenced
- Of remaining 27 SNP sites:
  - Illumina GAI – 2 not called, 25 agree
  - AB SOLiD – 1 not called, 2 called as heterozygous, 24 agree

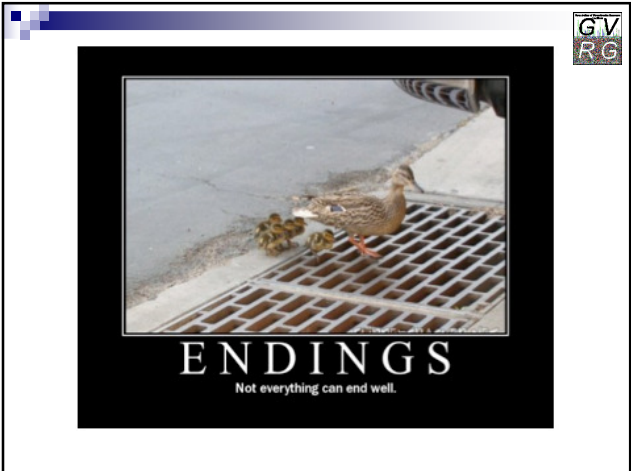
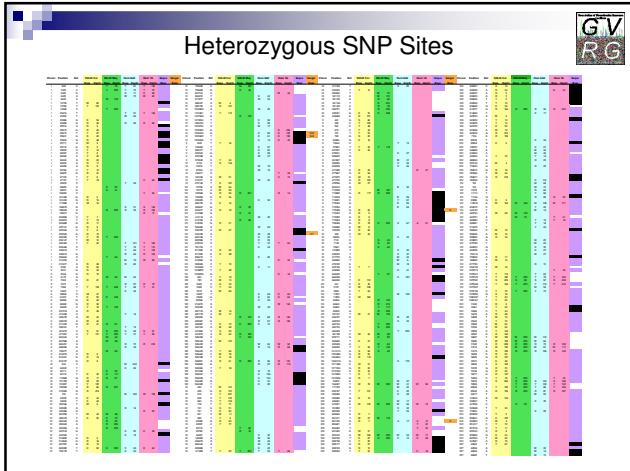
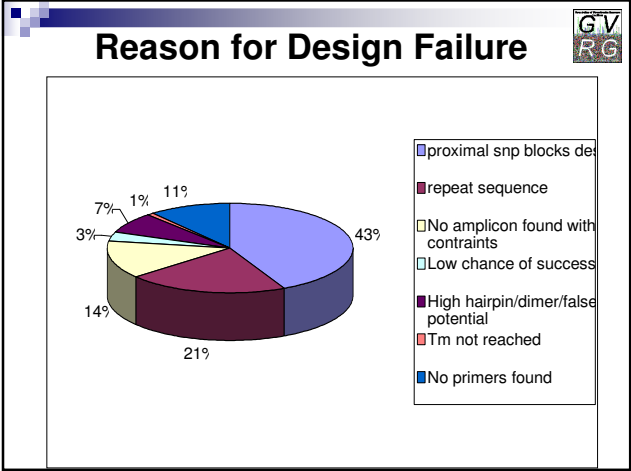
**GV**  
**RG**

## Sequenom Sites Selected

273 SNP's from the Illumina GA2 data that were not detected on the SOLiD

429 SNP's from the SOLiD data that were not detected by the Illumina

200 SNP Sites the were both picked up by the Illumina and the SOLiD as controls



## What we are considering for the future

- Paired-end sequencing for Illumina, or even mate-paired
- More even yield and/or depth-coverage
  - Reduce for SOLiD
  - Increase for Illumina
- Adjust analysis standards and parameters
  - Vendor analysis software
  - 3<sup>rd</sup> party analysis software
- Calibrate analysis for yeast
- Reanalysis with updated software
- Look for and at Indels
- Add data from a Roche 454

## We Would Like to Acknowledge the following for their contributions

### Data Analysis

Jingwei Ni  
Staff Scientist Bioinformatics  
Applied Biosystems

Aaron Noll  
Stowers Institute for Medical  
Research  
Kansas City, MO

Charles Cochran  
Applied Biosystems  
SOLiD Systems Specialist

SOLiD  
Bill Farmerie and his staff  
ICBR,  
University of Florida

Thanks to our donors for their generous support of this study.

#### Applied Biosystems

For the SOLiD reagents and disposables.

#### Stowers Institute

For giving us access to the Yeast cell line and SNP the data

#### IDT

For the discounted oligos used in the Sequenom assays

#### ABRF

For their continual support additional funding.

## Bob Welch's GVRG Legacy



Photo from [http://dceg.cancer.gov/fellowships/information/welch\\_fellowship](http://dceg.cancer.gov/fellowships/information/welch_fellowship)



## Genomic Variation Research Group

**Christian Lytle**

Dartmouth College, Hanover, NH

**Bruce Kingham**

University of Delaware, Newark, DE

**Alison Brown**

Harvard Partners Center for Genetics and Genomics, Cambridge, MA

**Joe Forrester**

University of Missouri, Columbia, MO

**Nathan Bivens**

University of Missouri, Columbia, MO

**Brian Sanderson**

Stowers Institute for Medical Research, Kansas City, MO

**Amy Hutchinson**

CGF, NCI, SAIC-Frederick, Gaithersburg, MD

**Helmen Escobar**

University of Utah, Salt Lake City, UT

**Michelle Detwiler**

Roswell Park Cancer Institute, Buffalo, NY



# Thank you

Our Poster can be found next to V51

