Phenotype Sequencing

Marc Harper

UCLA Bioinformatics, Genomics and Proteomics

March 4th, 2013
Collaborators

- Statistical analysis, simulations: Chris Lee (UCLA Bioinformatics, Genomics and Proteomics, Computer Science)
- Sequencing: Stan Nelson, Zugen Chen (UCLA Sequencing Center)
- E. coli mutants, screening: James Liao, Luisa Gronenberg (UCLA Chemical and Biomolecular Engineering)
The Basic Biological Problem

Relating Genotype and Phenotype
How can we determine which genetic variations are responsible (i.e. causally-connected) to particular traits (phenotypes)?
The Basic Biological Problem

Relating Genotype and Phenotype
How can we determine which genetic variations are responsible (i.e. causally-connected) to particular traits (phenotypes)?

Experiment Design
More generally, how can we design experiments to efficiently and confidently determine such genes given a set of (independently generated) individuals with a particular phenotype?
What is Phenotype Sequencing?

- A method for the discovery of genetic causes of a phenotype
What is Phenotype Sequencing?

- A method for the discovery of genetic causes of a phenotype
- Statistical model ranks genes most likely to be causal
What is Phenotype Sequencing?

- A method for the discovery of genetic causes of a phenotype
- Statistical model ranks genes most likely to be causal
- Takes advantage of high-throughput sequencing and pooling to dramatically reduce cost
What is Phenotype Sequencing?

- A method for the discovery of genetic causes of a phenotype
- Statistical model ranks genes most likely to be causal
- Takes advantage of high-throughput sequencing and pooling to dramatically reduce cost
- Can take advantage of known gene and mutation databases
What is unique/beneficial about Phenotype Sequencing?

- Comprehensive discovery of all genetic causes of a phenotype
What is unique/beneficial about Phenotype Sequencing?

- Comprehensive discovery of **all** genetic causes of a phenotype
- Cheap and Efficient
What is unique/beneficial about Phenotype Sequencing?

- Comprehensive discovery of all genetic causes of a phenotype
- Cheap and Efficient
- Open source simulation and computation pipeline
What is unique/beneficial about Phenotype Sequencing?

- Comprehensive discovery of all genetic causes of a phenotype
- Cheap and Efficient
- Open source simulation and computation pipeline
- Easy to extend and combine experimental results
Experiment

- Starting with a parent organism, create many mutants using random mutagenesis (e.g. UV, NTG)
Experiment

- Starting with a parent organism, create many mutants using random mutagenesis (e.g. UV, NTG)
- Screen mutants for phenotype (e.g. chemical tolerance, growth on particular medium)
Experiment

- Starting with a parent organism, create many mutants using random mutagenesis (e.g. UV, NTG)
- Screen mutants for phenotype (e.g. chemical tolerance, growth on particular medium)
- Sequence screened mutants and look for genes that are most commonly mutated: demultiplex, align, call SNPs/Indels
Experiment

- Starting with a parent organism, create many mutants using random mutagenesis (e.g. UV, NTG)
- Screen mutants for phenotype (e.g. chemical tolerance, growth on particular medium)
- Sequence screened mutants and look for genes that are most commonly mutated: demultiplex, align, call SNPs/Indels
- Since we only care where the mutations are, combining genomes into pools and tagging prior to sequencing can decrease sequencing cost 5-10 fold without losing any information
Experiment

- Starting with a parent organism, create many mutants using random mutagenesis (e.g. UV, NTG)
- Screen mutants for phenotype (e.g. chemical tolerance, growth on particular medium)
- Sequence screened mutants and look for genes that are most commonly mutated: demultiplex, align, call SNPs/Indels
- Since we only care where the mutations are, combining genomes into pools and tagging prior to sequencing can decrease sequencing cost 5-10 fold without losing any information
- Lower mean sequencing error → more pooling, typically 3-5 genomes into up to 12 tags (depending on genome size)
Effects of Screening
Screening boosts the mutation count signal in target genes.
Simulation: 20 targets in 5000 genes, 30 unscreened genomes and 30 screened genomes.
Effects of Screening

Screening boosts the mutation count signal in target genes.
Simulation: 20 targets in 5000 genes, 30 unscreened genomes and 30 screened genomes.
Once we have all the mutations, we basically count the number of times a particular gene is mutated.
Once we have all the mutations, we basically count the number of times a particular gene is mutated.

Have to control for many sources of variation, including mutagenesis bias, gene size, etc.
Experiment

- Once we have all the mutations, we basically count the number of times a particular gene is mutated.
- Have to control for many sources of variation, including mutagenesis bias, gene size, etc.
- Filter out synonymous, non-functional mutations (if possible).
Once we have all the mutations, we basically count the number of times a particular gene is mutated.

Have to control for many sources of variation, including mutagenesis bias, gene size, etc.

Filter out synonymous, non-functional mutations (if possible).

Correct for multiple hypothesis testings.
E. coli Gene Length Distribution
## Mutagenesis Bias

### Mutation Spectra: Comparison

<table>
<thead>
<tr>
<th>Organism</th>
<th>Mutagenesis</th>
<th>AT $\rightarrow$ GC</th>
<th>GC $\rightarrow$ AT</th>
<th>AT $\rightarrow$ TA</th>
<th>GC $\rightarrow$ TA</th>
<th>AT $\rightarrow$ CG</th>
<th>GC $\rightarrow$ CG</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>NTG</td>
<td>2.17%</td>
<td>96.6%</td>
<td>0.07%</td>
<td>0.07%</td>
<td>0.46%</td>
<td>0.61%</td>
</tr>
<tr>
<td><em>T. reesei</em></td>
<td>UV then NTG</td>
<td>30%</td>
<td>26%</td>
<td>15%</td>
<td>13%</td>
<td>10%</td>
<td>6%</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>Spontaneous</td>
<td>13.0%</td>
<td>46.8%</td>
<td>12.0%</td>
<td>7.85%</td>
<td>16.4%</td>
<td>4.1%</td>
</tr>
</tbody>
</table>
Mutagenesis Bias

Mutation Spectra: Comparison

<table>
<thead>
<tr>
<th>Organism</th>
<th>Mutagenesis</th>
<th>AT → GC</th>
<th>GC → AT</th>
<th>AT → TA</th>
<th>GC → TA</th>
<th>AT → CG</th>
<th>GC → CG</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>NTG</td>
<td>2.17%</td>
<td>96.6%</td>
<td>0.07%</td>
<td>0.07%</td>
<td>0.46%</td>
<td>0.61%</td>
</tr>
<tr>
<td><em>T. reesei</em></td>
<td>UV then NTG</td>
<td>30%</td>
<td>26%</td>
<td>15%</td>
<td>13%</td>
<td>10%</td>
<td>6%</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>Spontaneous</td>
<td>13.0%</td>
<td>46.8%</td>
<td>12.0%</td>
<td>7.85%</td>
<td>16.4%</td>
<td>4.1%</td>
</tr>
</tbody>
</table>

Effective Gene Size
Define the effective gene size as:

\[
\lambda = N_{GC}\mu_{GC} + N_{AT}\mu_{AT}
\]
Mutagenesis Bias

Mutation Spectra: Comparison

<table>
<thead>
<tr>
<th>Organism</th>
<th>Mutagenesis</th>
<th>AT → GC</th>
<th>GC → AT</th>
<th>AT → TA</th>
<th>GC → TA</th>
<th>AT → CG</th>
<th>GC → CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>NTG</td>
<td>2.17%</td>
<td>96.6%</td>
<td>0.07%</td>
<td>0.07%</td>
<td>0.46%</td>
<td>0.61%</td>
</tr>
<tr>
<td>T. reesei</td>
<td>UV then NTG</td>
<td>30%</td>
<td>26%</td>
<td>15%</td>
<td>13%</td>
<td>10%</td>
<td>6%</td>
</tr>
<tr>
<td>E. coli</td>
<td>Spontaneous</td>
<td>13.0%</td>
<td>46.8%</td>
<td>12.0%</td>
<td>7.85%</td>
<td>16.4%</td>
<td>4.1%</td>
</tr>
</tbody>
</table>

Effective Gene Size
Define the effective gene size as:

\[ \lambda = N_{GC} \mu_{GC} + N_{AT} \mu_{AT} \]

Can further account for other errors in a similar manner (e.g. gene length by normalizing)
Mutagenesis Bias

Mutation Spectra: Comparison

<table>
<thead>
<tr>
<th>Organism</th>
<th>Mutagenesis</th>
<th>AT → GC</th>
<th>GC → AT</th>
<th>AT → TA</th>
<th>GC → TA</th>
<th>AT → CG</th>
<th>GC → CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>NTG</td>
<td>2.17%</td>
<td>96.6%</td>
<td>0.07%</td>
<td>0.07%</td>
<td>0.46%</td>
<td>0.61%</td>
</tr>
<tr>
<td>T. reesei</td>
<td>UV then NTG</td>
<td>30%</td>
<td>26%</td>
<td>15%</td>
<td>13%</td>
<td>10%</td>
<td>6%</td>
</tr>
<tr>
<td>E. coli</td>
<td>Spontaneous</td>
<td>13.0%</td>
<td>46.8%</td>
<td>12.0%</td>
<td>7.85%</td>
<td>16.4%</td>
<td>4.1%</td>
</tr>
</tbody>
</table>

Effective Gene Size
Define the effective gene size as:

\[ \lambda = N_{GC} \mu_{GC} + N_{AT} \mu_{AT} \]

Can further account for other errors in a similar manner (e.g. gene length by normalizing)
Scoring

**P-values**

P-values are computed from a Poisson model for the target size $\lambda$ and observed mutations $k_{obs}$, for the null hypothesis that the gene is not a target:

$$p(k > k_{obs} | \text{non-target}, \lambda) = \sum_{k=k_{obs}}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!}$$

In other words, what is the probability of observing $x$ mutations in a normalized gene via random chance?

**Multiple Hypothesis Testing: Bonferroni Correction**

Finally we apply a Bonferroni correction to the p-values to reduce false positives due to chance in multiple hypothesis tests. In this case that means multiplying the resultant p-values by the total number of genes or pathways being tested.
Scoring

P-values
P-values are computed from a Poisson model for the target size $\lambda$ and observed mutations $k_{obs}$, for the null hypothesis that the gene is not a target:

$$p(k > k_{obs}|\text{non-target}, \lambda) = \sum_{k=k_{obs}}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!}$$

In other words, what is the probability of observing $x$ mutations in a normalized gene via random chance?

Multiple Hypothesis Testing: Bonferroni Correction
Finally, we apply a Bonferroni correction to the p-values to reduce false positives due to chance in multiple hypothesis tests. In this case that means multiplying the resultant p-values by the total number of genes or pathways being tested.
Scoring

P-values

P-values are computed from a Poisson model for the target size $\lambda$ and observed mutations $k_{obs}$, for the null hypothesis that the gene is not a target:

$$p(k > k_{obs}|\text{non-target, } \lambda) = \sum_{k=k_{obs}}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!}$$

In other words, what is the probability of observing $x$ mutations in a normalized gene via random chance?

Multiple Hypothesis Testing: Bonferroni Correction

Finally we apply a Bonferroni correction to the p-values to reduce false positives due to chance in multiple hypothesis tests. In this case that means multiplying the resultant p-values by the total number of genes or pathways being tested.
We identified three causal genes from 32 E. coli mutants selected for isobutanol tolerance (for biofuel production)
Results

- We identified three causal genes from 32 E. coli mutants selected for isobutanol tolerance (for biofuel production)
- Verified by multiple independent experiments (by our group and another)
Results

- We identified three causal genes from 32 E. coli mutants selected for isobutanol tolerance (for biofuel production)
- Verified by multiple independent experiments (by our group and another)
- We found many genes in several metabolic pathways from 24 E. coli mutants able to grow on glucose medium as the only carbon source

Each experiment cost approx $2400 ($1200 for sequencer lane + $1200 in reagents and labor for pooling)
Results

- We identified three causal genes from 32 E. coli mutants selected for isobutanol tolerance (for biofuel production)
- Verified by multiple independent experiments (by our group and another)
- We found many genes in several metabolic pathways from 24 E. coli mutants able to grow on glucose medium as the only carbon source
Results

- We identified three causal genes from 32 E. coli mutants selected for isobutanol tolerance (for biofuel production).
- Verified by multiple independent experiments (by our group and another).
- We found many genes in several metabolic pathways from 24 E. coli mutants able to grow on glucose medium as the only carbon source.

Each experiment cost approx $2400 ($1200 for sequencer lane + $1200 in reagents and labor for pooling).
## Results – 24 E. coli mutants

### Top hits

<table>
<thead>
<tr>
<th>Gene</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>iclR</td>
<td>$1.39 \times 10^{-25}$</td>
</tr>
<tr>
<td>aceK</td>
<td>$8.43 \times 10^{-14}$</td>
</tr>
<tr>
<td>malT</td>
<td>$4.81 \times 10^{-4}$</td>
</tr>
<tr>
<td>malE</td>
<td>0.045</td>
</tr>
<tr>
<td>yjbH</td>
<td>0.088</td>
</tr>
</tbody>
</table>
Using EcoCyc

- For phenotypes dependent on altering or shutting down particular metabolic pathways, the positive signal is split over the genes in the pathway.
Using EcoCyc

- For phenotypes dependent on altering or shutting down particular metabolic pathways, the positive signal is split over the genes in the pathway.
- EcoCyc pathways and functional groups allow the concentrating of the signal.
Using EcoCyc

- For phenotypes dependent on altering or shutting down particular metabolic pathways, the positive signal is split over the genes in the pathway
- EcoCyc pathways and functional groups allow the concentrating of the signal
- Finds many more genes than single-gene level analysis
Effects of Screening
Screening boosts the mutation count signal in target genes.
Simulation: 20 targets in 5000 genes, 30 unscreened genomes and 30 screened genomes.
Metabolic Pathways
## Results

**Table:** Top 10 gene groups ranked by pathway-phenoseq p-value (Bonferroni corrected for 536 tests)

<table>
<thead>
<tr>
<th>Group</th>
<th>Genes</th>
<th>p-value (phenoseq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD04099</td>
<td>aceK iclR</td>
<td>$2.01 \times 10^{-39}$</td>
</tr>
<tr>
<td>CPLX0-2101</td>
<td>malE malF malG malK lamB</td>
<td>$2.84 \times 10^{-9}$</td>
</tr>
<tr>
<td>ABC-16-CPLX</td>
<td>malF malE malG malK</td>
<td>$7.17 \times 10^{-8}$</td>
</tr>
<tr>
<td>PD00237</td>
<td>malS malT</td>
<td>$4.29 \times 10^{-4}$</td>
</tr>
<tr>
<td>GLYCOGENSYNTH-PWY</td>
<td>glgA glgB glgC</td>
<td>$4.25 \times 10^{-3}$</td>
</tr>
<tr>
<td>CPLX-155</td>
<td>chbA chbB chbC ptsH ptsI</td>
<td>0.145</td>
</tr>
<tr>
<td>PWY0-321</td>
<td>paaZ paaA paaB paaC paaD paaE paaF paaG paaH paaJ paaK</td>
<td>0.146</td>
</tr>
<tr>
<td>RNAP54-CPLX</td>
<td>rpoA rpoB rpoC rpoN</td>
<td>0.53</td>
</tr>
<tr>
<td>APORNAP-CPLX</td>
<td>rpoA rpoB rpoC</td>
<td>0.62</td>
</tr>
<tr>
<td>APORNAP-CPLX</td>
<td>rpoA rpoB rpoC rpoD</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Other and Ongoing Experiments

- Identified the cause of a rare disease in eight unrelated Korean individuals
Other and Ongoing Experiments

- Identified the cause of a rare disease in eight unrelated Korean individuals
- 8 mutants in mice looking for benzo(a)pyrene tolerance, identified several isoforms now being tested

Looking for collaborators for two larger-scale projects.
Other and Ongoing Experiments

- Identified the cause of a rare disease in eight unrelated Korean individuals
- 8 mutants in mice looking for benzo(a)prene tolerance, identified several isoforms now being tested
- 21 MRSA mutants, using binary pooling that allows for mutant identification

Looking for collaborators for two larger-scale projects.
Other and Ongoing Experiments

- Identified the cause of a rare disease in eight unrelated Korean individuals
- 8 mutants in mice looking for benzo(a)prene tolerance, identified several isoforms now being tested
- 21 MRSA mutants, using binary pooling that allows for mutant identification
- 21 Bacillus mutants, using binary pooling
Other and Ongoing Experiments

- Identified the cause of a rare disease in eight unrelated Korean individuals
- 8 mutants in mice looking for benzo(a)prene tolerance, identified several isoforms now being tested
- 21 MRSA mutants, using binary pooling that allows for mutant identification
- 21 Bacillus mutants, using binary pooling
Other and Ongoing Experiments

- Identified the cause of a rare disease in eight unrelated Korean individuals
- 8 mutants in mice looking for benzo(a)prene tolerance, identified several isoforms now being tested
- 21 MRSA mutants, using binary pooling that allows for mutant identification
- 21 Bacillus mutants, using binary pooling

Looking for collaborators for two larger-scale projects.
References

Software
Open source package phenoseq available at github: https://github.com/cjlee112/phenoseq

Contact
Marc Harper: marcharper@ucla.edu