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

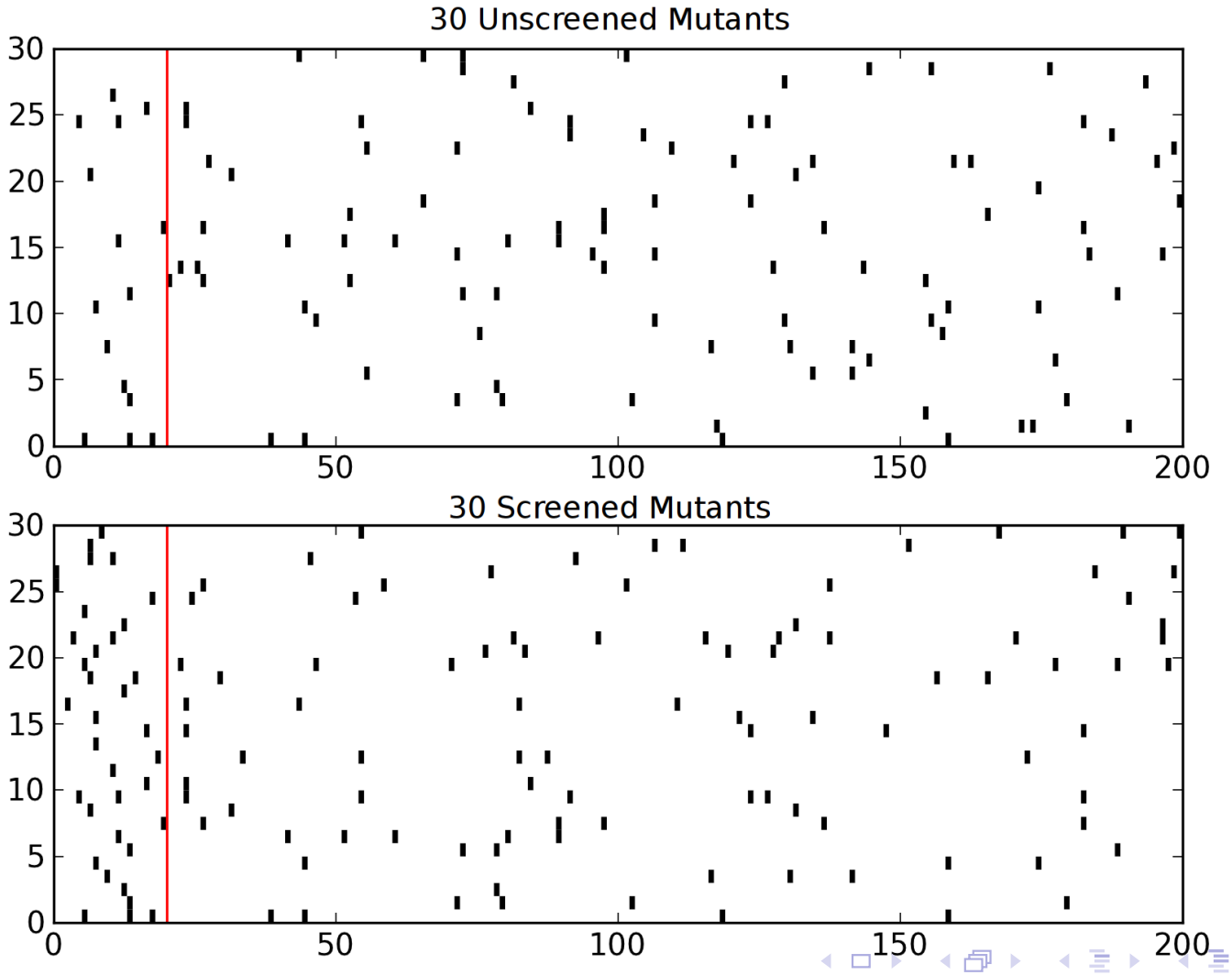
Experiment

- 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 $\rightarrow$ 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.

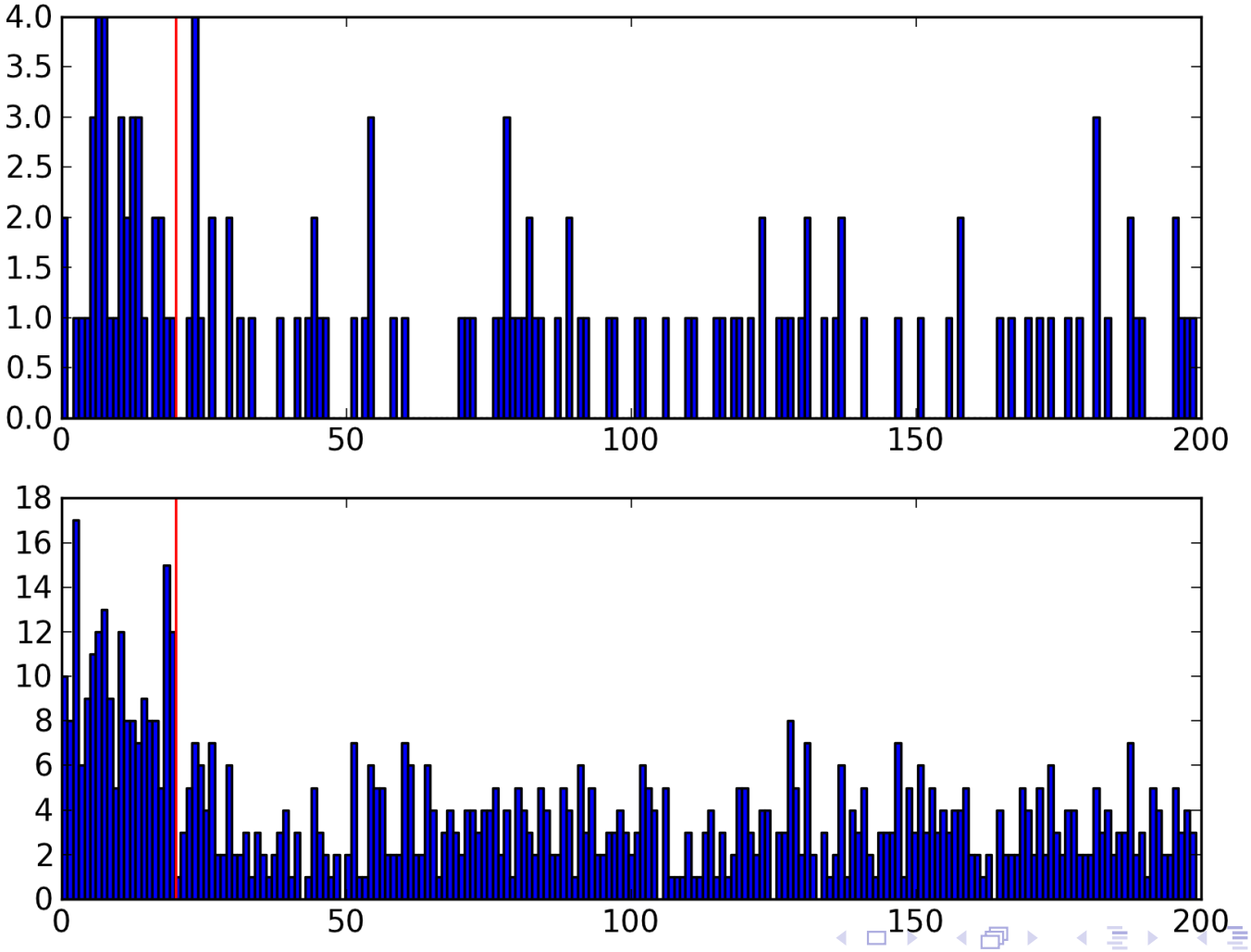
Simulation: 20 targets in 5000 genes, 30 unscreened genomes and 30 screened genomes.



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.



Experiment

- Once we have all the mutations, we basically count the number of times a particular gene is mutated

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.

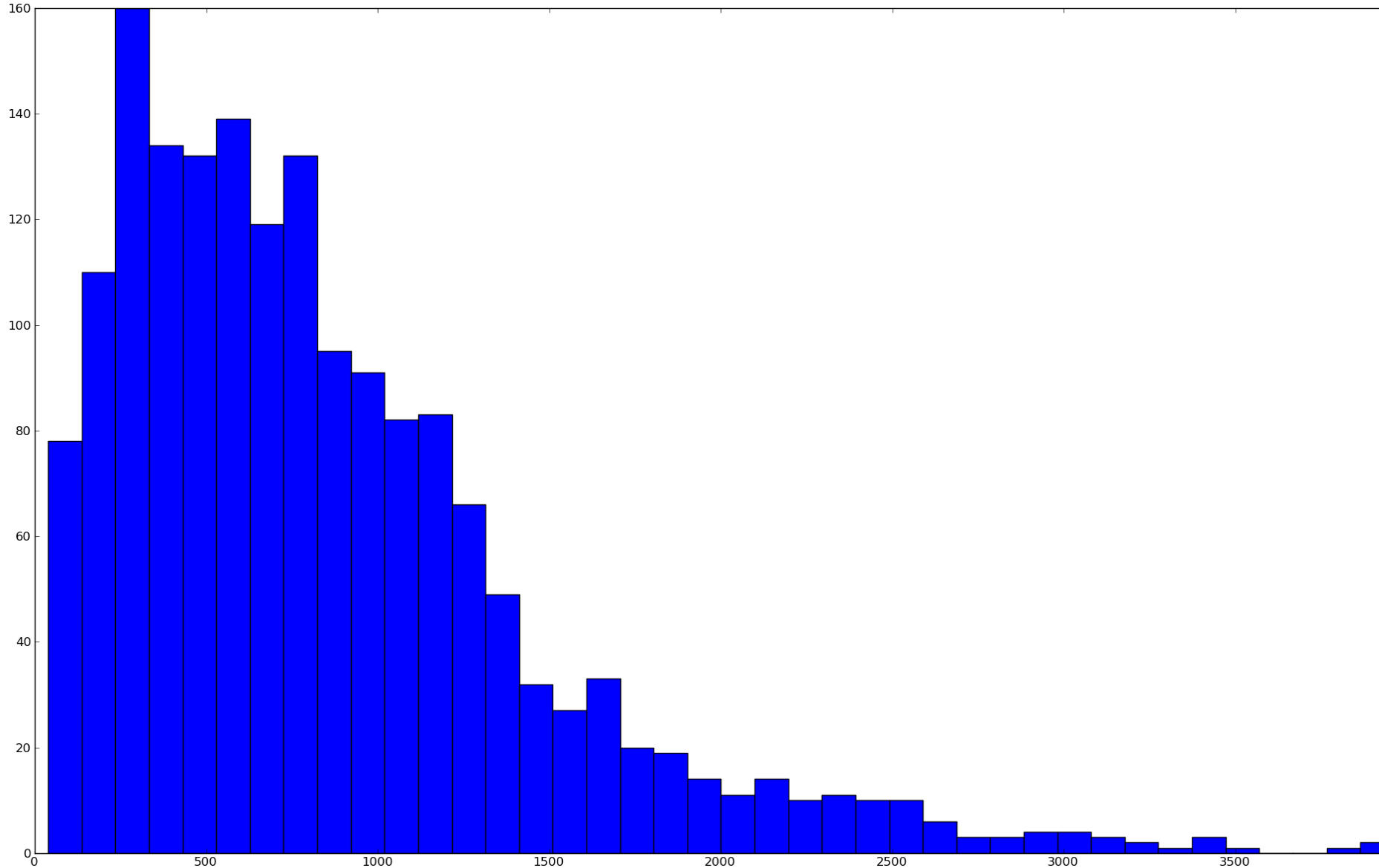
Experiment

- 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)

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)
- Correct for multiple hypothesis testings

E. coli Gene Length Distribution



Mutagenesis Bias

Mutation Spectra: Comparison

Organism	**Mutagenesis**	**AT → GC**	**GC → AT**	**AT → TA**	**GC → TA**	**AT → CG**	**GC → CG**
E. coli	NTG	2.17%	96.6%	0.07%	0.07%	0.46%	0.61%
T. reesei	UV then NTG	30%	26%	15%	13%	10%	6%
E. coli	Spontaneous	13.0%	46.8%	12.0%	7.85%	16.4%	4.1%

Mutagenesis Bias

Mutation Spectra: Comparison

Organism	Mutagenesis	AT → GC	GC → AT	AT → TA	GC → TA	AT → CG	GC → CG
E. coli	NTG	2.17%	96.6%	0.07%	0.07%	0.46%	0.61%
T. reesei	UV then NTG	30%	26%	15%	13%	10%	6%
E. coli	Spontaneous	13.0%	46.8%	12.0%	7.85%	16.4%	4.1%

Effective Gene Size

Define the effective gene size as:

$$\lambda = N_{GC}\mu_{GC} + N_{AT}\mu_{AT}$$

Mutagenesis Bias

Mutation Spectra: Comparison

Organism	Mutagenesis	AT → GC	GC → AT	AT → TA	GC → TA	AT → CG	GC → CG
E. coli	NTG	2.17%	96.6%	0.07%	0.07%	0.46%	0.61%
T. reesei	UV then NTG	30%	26%	15%	13%	10%	6%
E. coli	Spontaneous	13.0%	46.8%	12.0%	7.85%	16.4%	4.1%

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

Organism	Mutagenesis	AT → GC	GC → AT	AT → TA	GC → TA	AT → CG	GC → CG
E. coli	NTG	2.17%	96.6%	0.07%	0.07%	0.46%	0.61%
T. reesei	UV then NTG	30%	26%	15%	13%	10%	6%
E. coli	Spontaneous	13.0%	46.8%	12.0%	7.85%	16.4%	4.1%

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 λ and observed mutations k_{obs}, for the null hypothesis that the gene is not a target:

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

Scoring

P-values

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

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

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

Scoring

P-values

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

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

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.

Results

- 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

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

Gene	p-value
iclR	1.39×10^{-25}
aceK	8.43×10^{-14}
malT	4.81×10^{-4}
malE	0.045
yjbH	0.088

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

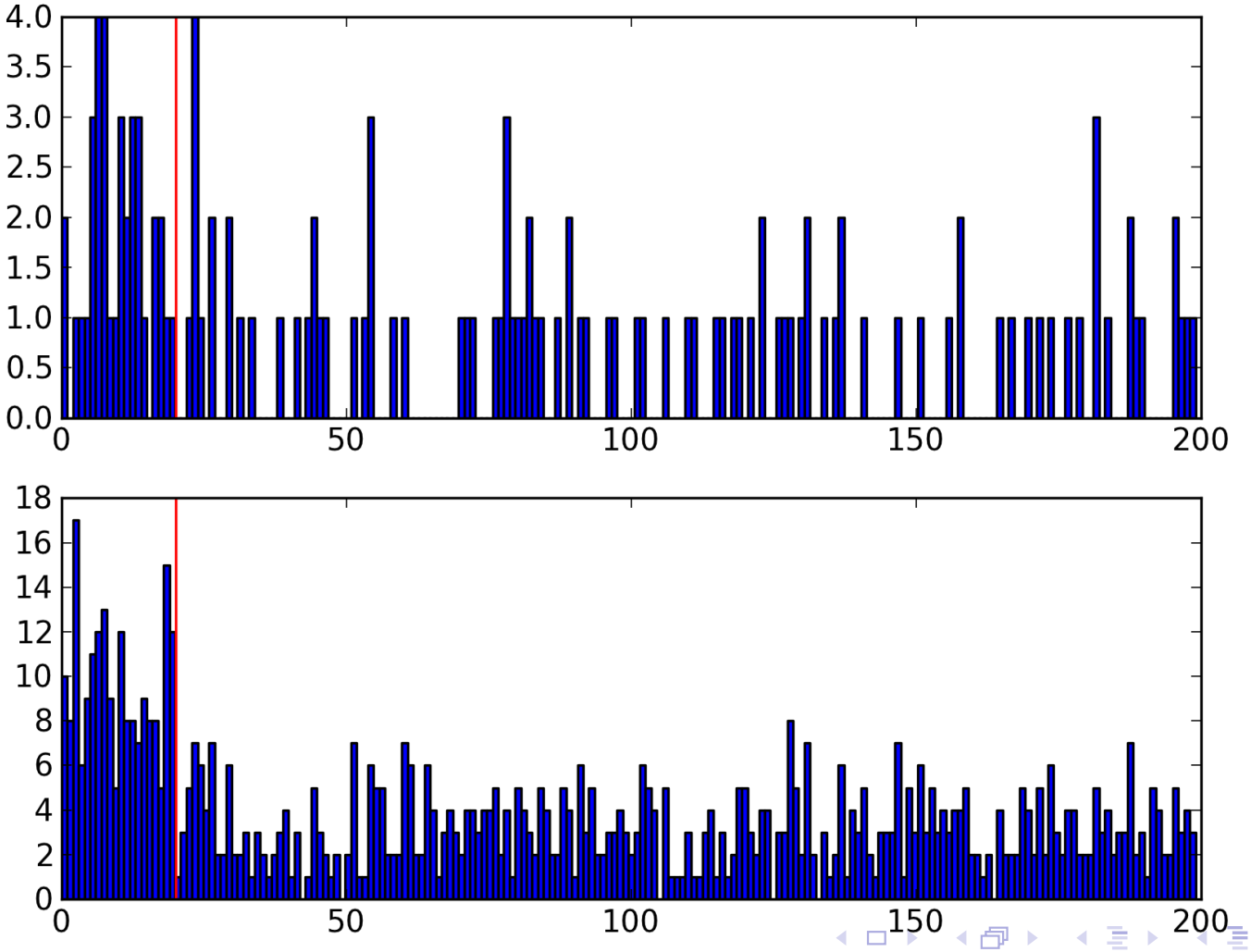
Using EcoCyc

- 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

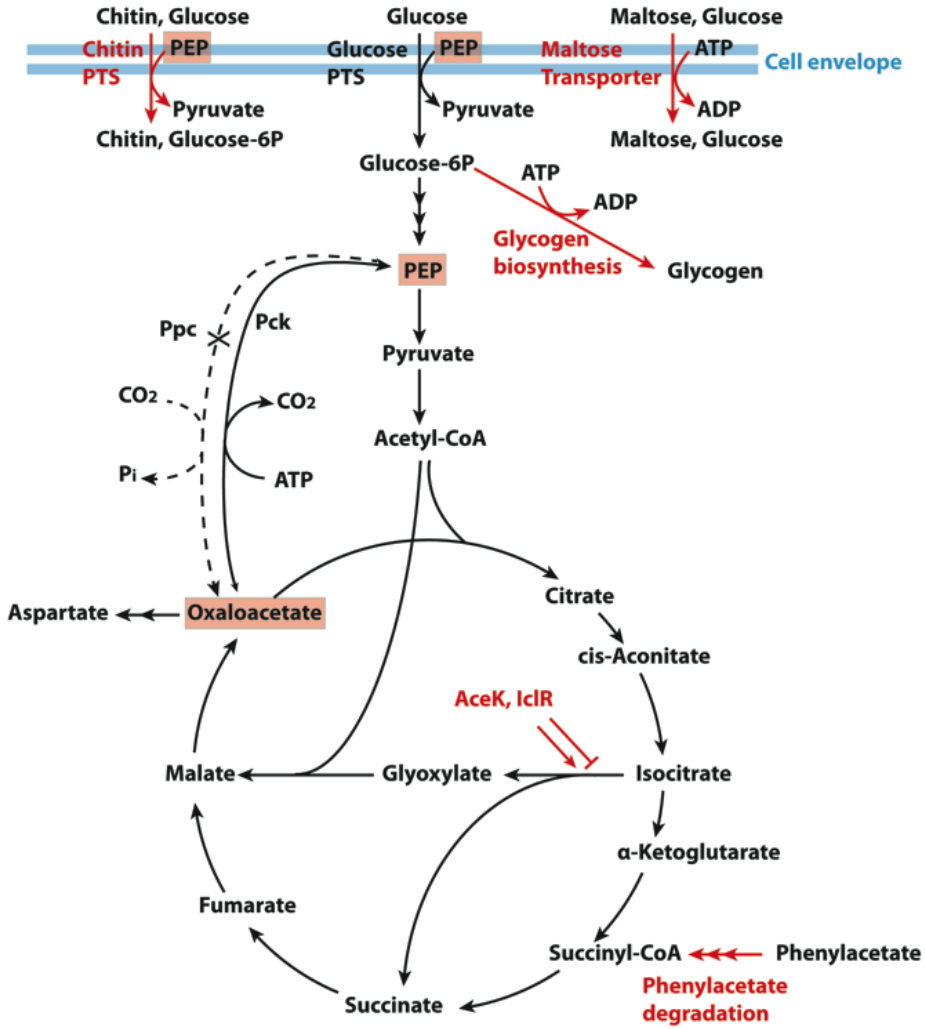
Effects of Screening

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)

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

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)prene tolerance, identified several isoforms now being tested

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

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

(1) Phenotype Sequencing, PLoS ONE, Feb 2011. Marc Harper, Zugen Chen, Traci Toy, Iara M. P. Machado, Stanley F. Nelson, James C. Liao, Chris Lee (http://www.plosone.org/article/info:doi/10.1371/journal.pone.0016517)
(2) ArXiv: "Comprehensive Discovery of Genes Causing a Phenotype using Phenotype Sequencing and Pathway Analysis", Marc Harper, Luisa Gronenberg, James Liao, Chris Lee

Software

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

Contact

Marc Harper: marcharper@ucla.edu