High throughput genetics - summary

Largely reverse genetics
- Deletion libraries/collections
- RNAi and CRISPR libraries
- Missense variant libraries (VUS problem)

Genome scale functional insight
- # of essential genes
- # of nonessential genes
- # of haplo-insufficient genes
- Pervasive genetic redundancy (genetic equivalent of dark matter)

Generation of useful datasets (i.e., organism databases)
Generation of useful experimental/biological reagents

Also see Mike White's lectures in Genomics, in April
Genome Sequence Enables Systematic Approaches

Genome Sequence

High Throughput Genetic Reagents
(Intervention)

High Throughput Gene & Gene Product Analysis
(Observation)
Genome Sequence Enables Systematic Approaches

Genome Sequence

High Throughput Gene & Gene Product Analysis
- DNA Methylation
- Chromatin Modifications
- RNA (microarrays, RNA-Seq)
- Protein (modifications, localization, interactions, complexes)

Data collection

Comprehensive Data Sets
Genome Sequence Enables Systematic Approaches

Genome Sequence

High Throughput Genetic Reagents
Deletion Libraries
CRISPR-Cas9 Libraries
RNAi Libraries
Expression-based Libraries

High Throughput Gene & Gene Product Analysis
DNA Methylation
Chromatin Modifications
RNA (microarrays, RNA-Seq)
Protein (modifications, localization, interactions, complexes)

Experiments
Gene Function

Data collection

Comprehensive Data Sets
Comprehensive Data Set $\neq$ Comprehensive Understanding
High-throughput studies:

False Negative (Type II error): Failure to observe a phenotype, protein-protein interaction etc. that normally occurs in the cell/organism under study.

False Positive (Type I error): Observe a phenotype, protein-protein interaction etc. that does not occur in the cell/organism under study. (FDR - false discovery rate)

- Quality control of the genomic reagents, the experimental design and data analysis are important for reducing false negatives and positives.

- Assessment of sensitivity & specificity for extent of these types of errors.

- High-throughput data will always have False Negatives. Obviously best if method and resulting data has low level of False Negatives.

High-throughput data is most useful to the researcher and community if False Positives are low – decreased likelihood of researchers chasing after false result.
Genome Sequence Enables Systematic Approaches

Genome Sequence

High Throughput Genetic Reagents
Deletion Libraries
CRISPR-Cas9 Libraries
RNAi Libraries
Expression-based Libraries

Experiments
Gene Function

Comprehensive Data Sets

High Throughput Gene & Gene Product Analysis
DNA Methylation
Chromatin Modifications
RNA (microarrays, RNA-Seq)
Protein (modifications, localization, interactions, complexes)

Experiments

Follow-up experiments for validation and to gain a deeper understanding
Some goals of high throughput reverse genetic analysis

Provide information of predicative value/ hypothesis generation

Single gene assays
1) Assign function to genes
   (identify additional genes involved in a biological process)
2) Assess variant for affect on function (damaging)

Multi gene assays
1) Assign function to genes
   (identify additional genes involved in a biological process)
2) For a biological process, identify genes that act in parallel
   (redundantly) or act in the same pathway/complex
3) Relate molecular attributes of genes to function & phenotype
High-throughput / Systematic Genetic Analysis of Gene Function

- *S. cerevisiae*: deletion libraries, expression-based libraries
- *S. pombe* (fission yeast): deletion & expression libraries
- *C. elegans*: RNAi libraries
- *Drosophila* cell culture, fly lines: RNAi libraries
- Mouse and human cell culture:
  RNAi and CRISPR-Cas9 libraries
- Systematic analysis of mouse knockout mutant collections

Provides first pass functional information on most genes in the respective genomes.

Are Community Resources:

Data from lab to lab will be more comparable than *de novo* efforts by individual labs (e.g. in cases where there is isogenic strain background, same knockdown construct).
High-throughput genetic analysis of gene function

For first pass functional information, phenotypic analysis must be fast and simple.

- Visible Phenotypes (e.g. dead/ alive – extent of growth) [Static]
High-throughput genetic analysis of gene function

For first pass functional information, phenotypic analysis must be fast and simple.

- Visible Phenotypes (e.g. alive or dead) [Static]

More complex phenotypic screening provides more functional information but is less amenable to high-throughput:

- Molecular marker phenotypes (presence/absence, temporal and spatial distribution of gene product or metabolite [static, e.g. antibody staining; real-time, e.g. GFP tagged protein]
- Quantitative phenotypes (e.g. growth rates, enzyme activities)
- Scalable expression analysis (e.g. NanoString, RNAseq?!$)
- Real-time visible phenotypes (e.g. microscopy)

Robotics/automation increases throughput and phenotypic depth
**S. cerevisiae Knockout Collection**


Deletions constructed in 5,916 genes (96.5% of total).

Deletion, from start codon to stop codon for each predicted gene are generated by homologous recombination, using oligo-nucleotide primers that target recombination to the specific gene, with a KanR marker to select for the integration event. Each deletion was verified by several PCR assays (quality control).

Strategy is feasible in yeast because of the highly efficient homologous recombination system and because yeast genes are small, most genes have no introns or when there are introns they are few and small, a very limited alternative splicing.
Summary of first pass functional data

- 1,105 (~19%) essential genes (growth in rich glucose media plates)
- 4811 (~81%) non-essential genes

- ~8% of non-essential genes have a closely related homolog (p < e-150) elsewhere in the genome while only ~1% of the essential genes have a closely related homolog.

- Of essential genes ~4% (15/356) were previously described as non-essential and ~0.2% (3/1620) of the non-essential genes were previously described as essential. (Different results likely a result of strain differences or in growth conditions.)

- Essential genes are more likely to have homologs in other organisms.

- Essential genes are more highly expressed than non-essential genes.
Why are only ~20% of genes needed for viability (function) in yeast?

Majority of genes are conserved and under purifying selection
Why are only ~20% of genes needed for viability (function) in yeast?

Majority of genes are conserved and under purifying selection

1a. Phenotypic assay is not assessed in the appropriate condition

1b. Phenotypic assay is insensitive – gene has a small quantitative requirement, that would be selected for on an evolutionary timescale

- Fitness assay - competitive growth, under different conditions (increased sensitivity phenotypic assay)
Why are only ~20% of genes needed for viability (function) in yeast?

Majority of genes are conserved and under purifying selection

1a. Phenotypic assay is not assessed in the appropriate condition

1b. Phenotypic assay is insensitive – gene has a small quantitative requirement, that would be selected for on an evolutionary timescale
   - Fitness assay - competitive growth, under different conditions (increased sensitivity phenotypic assay)

2. Redundant gene activities compensate (parallel pathways/buffering)
   - Synthetic lethal analysis
Parallel analysis (fitness assay) - Ron Davis

- Each deletion marked with oligo - barcode

- Testing haploid deletion strains in competitive growth experiment under various environmental conditions

Find that ~40% of deletions have a growth defect
Detecting molecular tags in yeast pools

Hybridize labeled tags to oligonucleotide array containing tag complements

PCR-amplify tags from pooled genomic DNA using fluorescently-labeled primers

Ron Davis et al., Stanford
Parallel Analysis
competitive growth experiment

Before
selection/treatment
red label

After
selection/treatment
green label

Ron Davis et al., Stanford
Parallel analysis of a large mutant library mutants in one experiment

Growth in minimal media
Yellow – equivalent
Red – Depletion (negative selection)
Green – Enrichment (positive selection)

What is the relationship between genes whose expression is induced by a particular environment and the function of those genes in that environment?

- Assume that if a gene is induced/expressed under a certain condition then the gene is important for growth in that condition.
- Therefore, deletion of an up-regulated gene would be expected to cause a growth defect in that condition.
Comparison of expression and fitness profiling data

Example - growth in 1.0 M NaCl

- Fitness determined by competitive growth experiment. Plot on Y-axis, from 0 (wild-type) to 1200 (>100 significant defect)

- Expression profiling in isogenic strain in media + or - 1.0 M NaCl. Plot on X-axis as Log ratio expression
  (<=-0.5 significant repression, >0.5 significant induction)
Comparison of Expression and Fitness Profile for the Deletion Collection
Little correlation between expression and fitness profiling data

For all conditions tested: 1.0 M NaCl, minimal media, galactose, 1.5 M sorbitol and alkali.

Strong evidence that mRNA expression provides only part of the picture for a given biological process.
Why are ~80% of yeast genes non-essential under optimal lab conditions?

Homologous gene (paralog) may provide the same function (redundancy; likely for ~8% of the non-essential genes).

**Non-homologous gene or pathway may provide the same function (redundancy, parallel activity or pathway); this type of redundancy may provide buffering so that the process occurs with higher fidelity when the system is stressed environmentally or due to genetic variability that occurs in natural populations.
Redundant functions can be uncovered as synthetic genetic interactions – double loss-of-function mutants that enhance the original phenotype or gives a phenotype not observed by either single mutant.

High-throughput synthetic lethal analysis

- Generate tester strain (mating type \textit{MATalpha}) where the non-essential query gene has been deleted by homologous recombination using Nat$^R$ cassette.

- Mate tester haploid strain to an ordered array of \(\sim 4800\) yeast strains containing non-essential gene deletions (from Winzeler et al. and Giaever et al.; \textit{MATa}) and select for diploids that are Kan$^R$ and Nat$^R$.

- Diploids are sporulated and the haploid double mutant is obtained if viable, following selection for haploidy (\textit{MATa}) and for Kan$^R$, Nat$^R$.

Query gene \textit{BNII} – encodes a formin protein family member that functions in cortical actin assembly for polarized cell growth and spindle orientation.
Synthetic Genetic Array Methodology

Tester/query gene: $MAT^\alpha$ nat$^R$ bni1$\Delta$
Deletion library gene: kan$^R$

Mating

Sporulation

$MAT^\alpha$ Haplod Selection ($MFA1pr-HIS3$)

Double Mutant Selection: kan$^R$ nat$^R$
Double-Mutant Array and Tetrad Analysis

A
Select for nat$^R$ and kan$^R$

$bnr1\Delta$
$bnr1\Delta$

$bni1\Delta$

$bni1\Delta::natR$

B

T NPD PD

$bnr1\Delta$
$cla4\Delta$
$bud6\Delta$
Results from synthetic lethal/sick screen with \textit{BNI1} query gene

- 67 potential synthetic lethal/sick interactions found.
- 51 (~75\%) were confirmed by tetrad analysis (~25\% false positives)

- 51 synthetic lethal/sick interactors grouped based on “cellular roles” as defined by Yeast Proteome Database (YPD):
  - 20\% are “cell polarity” genes, e.g. bud emergence genes
  - 18\% are “cell wall maintenance” genes, e.g. chitin synthase genes
  - 16\% are “mitosis” genes, e.g. dynein/dynactin spindle orientation genes
  - 18\% are genes of unknown function

8/11 previously known synthetic lethal/sick genes found.
  {3 missing, 2 not among the 4700 tested, 1 not found (false negative)}

43 new synthetic lethal/sick interactions found.
But what could the synthetic lethality/sick interaction of two deletion null mutants be telling us?

- Two genes with redundant activities within a single pathway.

- Two genes, acting in separate pathways that are redundant for the same biological process.

- Two separate biological processes that when both impaired result in lethality (e.g. polarized cell growth and cell wall maintenance). Less or non-informative for understanding gene function - false positive.
Interaction Map of synthetic lethal/sick interactions

- Repeat the high-throughput synthetic lethal screen using
  1) gene of interest with unknown cellular roles
  2) genes with well characterized cellular roles

- Display genetic interactions as binary gene-gene relationships.

- A gene with an unknown cellular function(s) is expected to have greater connectivity and to be surrounded by genes of similar function.

Map/network obtained using the BIND package
(Bader et al., 2001, *NAR* 29:242)
Synthetic Lethal Genetic Interaction Network
Two gene products that act in the same pathway but are not redundant with each other will not show synthetic lethality.

\[ \text{gene-A} \rightarrow \text{gene-B} \rightarrow \text{gene-C} \rightarrow \text{Biological Process} \]

\text{gene-A null; gene-B null double will have the same phenotype as each single mutant.}
Two gene products that act in the same pathway but are not redundant with each other will not show synthetic lethality.

However, if gene-A and gene-B act in the same pathway then expect them to show similar sets of genetic interactions with a deletion in another gene that functions redundantly in the same process.

\[\text{gene-A null; gene-Q null} \quad \text{Synthetic Lethal}\]

\[\text{gene-B null; gene-Q null} \quad \text{Synthetic Lethal}\]
High-throughput screen was performed with two genes as the query that function in actin assembly, *ARC40* and *ARP2*, which encode subunits of the Arp2/3 complex.

*ARC40* had 40 synthetic lethal/sick interactions.

*ARP2* had 44 synthetic lethal/sick interactions.

31/40 interactions with *ARC40* and 31/44 interactions with *ARP2* are shared.
False negatives and false positives make the method not fully comprehensive

a) Essential genes not included in the analysis.

b) Unintended genetic alterations (point mutations, aneuploidy) in the deletion strains. Tetrad analysis helps here.

c) Linked double deletions are recovered at lower frequency.

d) Haploid deletion strains that are slow growing, fail to mate or sporulate are often not present and confound the interpretation.

e) Negative regulatory interactions will not be recovered in a synthetic lethal screen.
The genetic landscape of a cell. Costanzo et al.  
*Science* 327:425-431
- Automated imaging platform for quantitative scoring of growth phenotype.

- 1712 query genes, including 334 conditional or hypomorphic alleles of essential genes, >>>~ 5.4 million gene pairs assayed

- ~170,000 interactions identified (3% of dbls). Most genes have few interactions, some genes have numerous interactions (hubs, e.g. ARC40 and ARP2).

- Most hub genes evolve more slowly ($dN/dS$) than genes with few interactions.

- There is a lack of intersection between genes that were identified as interacting genetically and physical interactions (IP-mass spec, yeast two-hybrid) between the corresponding gene products.

Interpretation: Vast majority of synthetic lethal interactions are between genes in separate pathways/complexes that converge to promote a biological process while most physically interacting gene products function in the same pathway/complex.
- Nodes (dots) are genes
- Lines (edges) connect gene pairs that share common sets of interactions (synthetic lethality; e.g. ARC40 and ARP2)
Systematic Gene Overexpression (gf)  
Sopko et al., 2006, Mol Cell 21:319-330

5280 genes (85\%) conditionally overexpressed on a multicopy plasmid from a galactose inducible promoter

No bias to essential genes (~20\%)
184/769 show obvious cytological defect
Their classification of gene overexpression
Genetic basis of overexpression gain-of-function

Hypermorphic - increased wt activity

Antimorphic poisoning
 i) Poisoning “itself” (dominant negative) - acts like loss-of-function
 ii) Poisoning other gene products that it normal associates with
   - predicted to show loss-of-function-like-phenotypes of the
     other gene products (if known)

Neomorphic - novel unregulated effect
Overexpression cytological phenotype of 42/184 genes resembled the corresponding gene deletion - thus analogous to antimorphic poisoning of itself.

Antimorphic (dominant negative) poisoning of “itself”
3/184 show hypermorphic behavior (binary switch)

far1p \[\rightarrow\] cdc28p \[\rightarrow\] G1/S transition

OFF \hspace{2cm} ON \hspace{2cm} G1/S transition

ON \hspace{2cm} OFF \hspace{2cm} G1 arrest

null \hspace{2cm} G1 arrest

null \hspace{2cm} Premature G1/S transition

far1p \hspace{2cm} Overexpression \hspace{2cm} G1 arrest
Antimorphic (dominant negative) poisoning of “itself”

Hypermorphic
139 displayed a phenotype not obviously related to the gene deletion.

**Antimorphic (dominant negative) poisoning of “itself”**

**Hypermorphic**

**Antimorphic poisoning of a physically interacting gene product**
RNAi libraries allows high-throughput investigation of gene function in *C. elegans* and other systems

- Rapid gene specific method for inducing depletion and sometimes elimination of gene function.

  - Induces degradation of the endogenous mRNA.

  - Depletes both maternal and zygotic mRNA (worms). (often advantageous for looking at embryonic phenotypes that are not accessible in zygotic mutants due to maternal rescue.)

- ~19,500 predicted genes in *C. elegans*.

- RNAi of 16,757 genes (86%)

- Bacterial feeding of wild-type (Bristol N2 strain) with gene specific RNAi, then scoring the F1 generation for visible (static) phenotypes in a hermaphrodite population.

- 1,722 (~10%) gave one or more visible phenotypes. (198 have close homologs that were likely also cross-inactivated.)
Scored 21 dissecting microscope visible phenotypes that were grouped into 3 mutually exclusive phenotypic classes:

- nonviable (Nonv) - embryonic or larval lethality or sterility
- growth defects (Gro) - slow or arrested postembryonic growth
- viable postembryonic (Vpep) – defects in postembryonic development (e.g. movement or body shape)

Genes with homologs in other species are much more likely to display an RNAi phenotype (21% vs 6%).

X-chromosome has a strong under-representation for Nonv genes, while having a slight over-representation of Vpep genes.

Within individual chromosomes there is a non-random distribution of genes with RNAi phenotype.
Low level of False Positives

>0.5%

High level of False Negatives

RNAi phenotypes missed for ~30% of known essential genes and ~60% of known genes required for postembryonic development.

Resistant cells/tissue

Inefficient RNAi

Experimental variation
Why do only ~15% of *C. elegans* genes show a phenotype in genome wide RNAi screens?

Genome wide screens are Po or F1 generation and are gross morphological/ low resolution phenotypes (dissecting microscope)
Why do only ~15% of *C. elegans* genes show a phenotype in genome wide RNAi screens?

Genome wide screens are Po or F1 generation and are gross morphological/low resolution phenotypes (dissecting microscope)

1. **Fitness assay** (multi-generation)
   - small affects on growth and fertility are detectable

2. **High content phenotypic screens**
   - phenotype examined at the cellular or subcellular level
   - phenotype related to specific molecular readouts
The Majority of Animal Genes Are Required for Wild-Type Fitness


- Competitive growth experiment
- Found that 60 – 70% of genes have a loss of function fitness defect
The Majority of Animal Genes Are Required for Wild-Type Fitness

Raman et al, 2012, Cell, 148:792-802

- Competitive growth experiment
- Found that 60 – 70% of genes have a loss of function fitness defect

Higher frequency of genes require for fitness than budding yeast

Propose this is due to:
* Multicellularity – requirement for function in one or more cell types
* Less “buffering” from parallel pathways
- Used vital dye Nile Red to detect stored fat

- RNAi identified 305 genes that promote fat storage
- RNAi identified 112 genes that inhibit fat storage

- Genes known to control fat storage in humans identified (tubby ortholog, serotonin biosynthesis, etc)

- Identifies candidate genes for further study as >50% of genes found have mammalian homologs not previously implicated in fat storage.
Systematic screens in mammalian cell culture

- RNAi screens (shRNA)

- CRISPR-Cas9 screens
  - nuclease active
  - nuclease inactive (dCas9)

Two general platforms

- Arrayed screens

- Pooled screens
The most efficient method for delivery of shRNAi and CRISPR-Cas9 reagents to cultured cells is via lentiviral vectors.

The lentiviruses is integrated into the genome at less than one copy per cell (MOI = 0.3) that are recovered by puromycin selection, also present in the virus vector.

Shalem et al, 2015
shRNA and sgRNA sequences are computationally identified, made by oligonucleotide synthesis, flanked by universal primers, relevant bar codes, then cloned and packaged into lentiviruses.

Arrayed screen

Reagents (singly or small pools) of known identity are placed in each well and then processed through all subsequent steps.

*Plus:* Allows high content microscopy screening (e.g. fluorescent or luminescent markers) for complex phenotypes.

*Minus:* Expensive & time consuming as reagents are separately prepared; requires robotic automation and thus specialized facilities.

Shalem et al, 2015
**Pooled screen**
(parallel analysis)

Manipulations done in batch. In silico designed shRNA and sgRNA are synthesized as high complexity pools.

*Plus*: Less expensive and labor intensive than arrayed screens.

*Minus*: Analysis is limited to growth phenotypes (effects on proliferation and survival) or enrichment methods such as cell sorting (eg. FACS). Readout is from high throughput sequencing of barcode tags.

Shalem et al, 2015
Two types of selections used with the pooling platform

Positive selection:
Culture conditions has strong selective pressure – viability (resistance) following treatment with drug, toxin or pathogen.
For example
-*siRNA knockdown results in host resistance to HIV infection (Zhou et al, 2008, Cell Host & Microbe, 4:495-504.

Negative selection:
CRISPR-Cas9 mutation or shRNA knockdown results in depletion of cells.
This is modeled after the yeast knockout competitive growth experiments.
For example
Negative selection (sequence depleted)

Positive selection (sequence enriched)
High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities
Hart et al, 2015, Cell 163:1515–1526

~50% of all human genes are expressed in a cell type. How many of these genes are “essential” for cell viability across most or all cell types? (Essential = fitness growth assay)

Distinguish between
- Core essential genes – essential across multiple cell types and
- Context dependent essential genes – essential only in specific contexts such as environment or genetic background (eg. synthetic lethality)
(1) Second generation high complexity CRISPR-Cas9 sgRNA library

~10 sgRNAs per gene for ~17,500 genes to give a library with a complexity of ~175,000 sgRNAs.

(2) Test of 5 cell lines. Comparison with prior GeCKo-A375 cell line CRISPR-Cas9 screen (Shalen et al, 2015, Science 343:84-87) and shRNA screen (Hart et al, 2014, Mol Syst Biol, 10:73)
Identified ~2000 genes that are essential in each cell line, based, on their fitness criteria.
Identified a set of 1580 core essential genes
Majority of Essential genes found in protein complexes

shRNA knock down most effective for high abundance gene products with longer mRNA half-life.
The cell culture approach vastly under-represents genes that function in organismal fitness, development and general cell non-autonomous function.
Considerations with nuclease active CRISPR/Cas9 screens
(NHEJ, no repair template)

- Generated stable gene mutations (although mutation maybe selected for or against)

- Were variants generated on both homologs? (assuming diploid cell lines)

- Will generate a variant continuum – from complete elimination of function to essentially wild type
Catalytically inactive dCas9 Screens
(usually pooled screens)

- CRISPRi (inactivation)     loss of function (necessity)

- CRISPRa (activation)       gain of function (sufficiency)

Catalytically inactive dCas9 Screens

- CRISPRi (inactivation) loss of function (necessity)

- CRISPRa (activation) gain of function (sufficiency)

Similarities to pharmacology
CRISPR/dCas9 must remain active for sufficient time to yield a change in gene expression and thus phenotype

Catalytically inactive dCas9 Screens

- **CRISPRi**
  - sgRNAs recruits dCas9 tagged with transcriptional repressor (e.g., KRAB)
  - sgRNAs target 50 to 100bp downstream of TSS

- **CRISPRa**
  - sgRNAs recruits dCas9 tagged with transcriptional activator (e.g., VP64)
  - sgRNAs target -400 to -50bp upstream of TSS

Nucleosomes block access to CRISPR/Cas9
sgRNA design incorporates nucleosome position info

Catalytically inactive dCas9 Screens

- CRISPRi
  Reduced expressed genes normally found in cell type analyzed
  Up to ~100 fold reduction in expression observed

- CRISPRa
  Increased expression of genes normally found in cell type analyzed
  Ectopic expression of genes not normally found in cell type analyzed
  Can uncover convergent activities/pathways that function in process
  Up ~10x increase in expression observed

Catalytically inactive dCas9 Screens

CRISPRi screens identified ~2100 essential genes (~10%) in K562 cells

In toxin resistance/sensitivity screens, can identify genes where CRISPRi and CRISPRa give opposite phenotypes

Systematic genetic analysis in the mouse

International Knockout Mouse Consortium (IKMC)
International Mouse Phenotyping Consortium (IMPC)
< http://www.mousephenotype.org/ >

- Uniform design of conditional floxed allele (with LacZ reporter)
- Uniform strain background, C57BL/6
- Uniform phenotypic analysis pipeline
- Mutants available to the community

Workflow and phenotype analysis pipeline

A

Chimaera → GLT → Het +/- → Intercross → Hom viability P14 → Yes → Hom +/- → Hom fertility

No → Het +/- → Hom viability E14.5 → MGP Pipelines

LacZ reporter gene expression
- Adult
- E14.5

B

Terminal Bleed and Necropsy

Hair Morphology
- Hair Follicle Cycling
- Open Field
- Modified Grouping
- Grip Strength
- Hot Plate
- Dysmorphology
- Indirect Calorimetry (TMR)
- Glucose Tolerance (ip)
- Auditory Brainstem Response (n = 4)
- Body Composition (DEXA)
- Stress Induced Hyperthermia
- Slit Lamp
- Ophthalmoscope
- Haematology (CBC)
- Peripheral Blood Lymphocytes
- Erythrocyte Micronuclei (TMR)
- Tissue Blotting (2M + 2F)
- Heart Weight
- Heart Histology (3M + 3F)

Age (wks):
- 4
- 6
- 9
- 10
- 12
- 13
- 14
- 15
- 16

Body weight

White et al, 2013 Cell

509 phenotypes examined
3,328 germline transmitting het lines have been examined

- **23% are lethal** (largest class is lethality before E9.5)
  Much less likely to have a paralog

- **11% are subviable**
  More likely to have a paralog

- **66% viable genes**
  More likely to have a paralog

- **5% of genes required for fertility**
3,328 germline transmitting het lines have been examined

- 65% of gene mutants have at least one phenotype
- 42% of mutant heterozygotes show haploinsufficiency

* Inability to predict phenotype based on gene sequence or expression pattern
* For genes previously examined, phenotypes missed mostly because not screened for
* Incomplete penetrance & variable expressivity are common even with null alleles and uniform genetic background
What is the population frequency of homozygous loss of function mutations in the human genome?

Identification of a large set of rare complete human knockouts

Sulem et al., 2015, Nature Genetics 47:448-453

- Icelandic population (bottle neck ~1000 years ago during colonization, moderate level of inbreeding) from deCODE project.

- Sequencing of trios (parents + offspring or families).
  :. Phase is known - make interpretations from homozygotes
- For 1,171 genes (~6%), individuals that are homozygous or compound heterozygous for loss of function rare variants (minor allele frequency (MAF), <2%) were identified.

Suggests these genes are not essential in humans.

However, homozygous loss of function offspring from heterozygous parents occurred less frequent than expected, indicating partial deleterious effect.

Is incomplete penetrance or modifier (suppressor) loci the explanation for lack of lethality?
The Variants of Uncertain Significance problem in human genetics
The Variants of Uncertain Significance problem

Massively parallel single gene variant analysis (aka deep mutational scan) to generate sequence-function (genotype-phenotype) maps

- Provide insight into protein structure/function

- For known disease genes (e.g., BRCA1) there are numerous variants of unknown functional significance (VUS). Functional assays can provide prospective information on deleteriousness of variants, allowing “look-up tables” for interpreting new variants to be generated that can be used clinically in patient management.

Generalized workflow for massively parallel single gene variant analysis

1) Construct variant library (programmed oligonucleotide synthesis) for gene of interest.

2) Deliver variant library to cells, through lentivirus or via CRISPR and homologous recombination repair.

3) Evaluate all variants in functional (phenotypic) assays [stratification] (design & validation of phenotype assay is the most challenging step of workflow)

4) Associate genotype with phenotype - usually sequence to quantify the representation of each variant in the context of the assay.

5) Ranking of findings, calibrated relative to benchmarks of known null and benign variants.

Issues

- Not yet feasible in organisms, only in single cells (yeast, tissue culture).
  What is the appropriate cell type? Ploidy?

- What phenotype (wild type function) is assayed?
  Many proteins have multiple functions;
  Which function is assayed?
  Distinct functions may require different levels of gene activity.
  Functional assay needs to be scalable, $>10^4$ assays.
  Trade-off of simple/scalable assay with sensitivity, specificity and biological validity of the assay.

- Cell non-autonomous effects are problematic to model.

- Dominant presentation in humans is the rule.
  Assay in heterozygous state (endogenous gene product level)
TP53 (p53) - most frequently mutated tumor suppressor gene in cancer (somatic).

Of TP53 variants analyzed functionally, many appear to be antimorphic/dominant negative (functions as a homo-tetramer)

What somatic TP53 variants found in tumors result in disrupted TP53 function and contribute to cancer, and which function as dominant negatives?

- TP53 promotes cell cycle arrest, apoptosis and senescence in response to DNA damage, through positive regulators ATM, ATR, CHEK2 & CHEK1.

- TP53 is negatively regulated by MDM2, an E3 ubiquitin ligase. Cancer cell lines that are wild type for TP53 require MDM2 to grow.

```
DNA damage

ATM  ATR  CHEK2

MDM2  TP53  →  Cell arrest

MDM2  TP53  →  Cell Growth

MDM2  TP53  →  Cell Arrest

mutant

mutant

MDM2  TP53  →  Cell Growth
```
Nutlin-3 is an chemical inhibitor of MDM2 repression of TP53, which can be used to assess function in a library of TP53 amino acid substitution variants.

Nutlin-3

MDM2 ┌───┐ TP53  →  Cell Arrest (depleted)
   Wild type

Nutlin-3

MDM2 ┌───┐ TP53  →  Cell Growth (enriched)
   mutant
- Generate a library of TP53 variants containing all 20 aa substitutions and stop codons.

- Introduce into A549 cells via lentivirus (puromycin selection), with promoter expressing TP53 at ~wild type levels.

- In A549 TP53 null cells, conduct competitive growth experiment, in the presence of Nutlin-3, where fold enrichment is related to extent of TP53 loss of function.

- In A549 cells with a single copy of wild type genomic TP53, assess dominant negative activity of variants.

Giacomelli et al.
Deleterious (red), benign (blue) variants
- Vast majority of Stop variants are strong loss of function
- For a number of residues, many amino acid substitutions result in loss of function, while for others, only certain variants disrupt function.
- Proline substitution more likely to disrupt function.
Dominant negative

Deleterious (red), benign (blue) variants
- Vast majority of Stop variants are not strongly dominant negative
- Dominant negative variants are more restricted in distribution than loss of function.
Of 3,819 DBD variants, 1,481 are loss of function. 1,219 (82%) are dominant negative.
Loss of function

<table>
<thead>
<tr>
<th>Variant</th>
<th>p53^NULL + nutlin-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA Substitution</td>
<td>ACDEFGH I KLMNPQRSTVWY</td>
</tr>
<tr>
<td>Average Z-score</td>
<td>-30.3</td>
</tr>
</tbody>
</table>

Dominant negative

<table>
<thead>
<tr>
<th>Variant</th>
<th>p53^WT + nutlin-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA Substitution</td>
<td>ACDEFGH I KLMNPQRSTVWY</td>
</tr>
<tr>
<td>Average Z-score</td>
<td>-30.3</td>
</tr>
</tbody>
</table>

Protein domains, tumor variant freq

- Variants in cancer patients - almost all are dominant negative based on functional analysis.

- Hot spots, C>T at mCpG, signature 1 among 30 mutational signatures in cancer (TCGA).
- Have strong loss of function and dominant negative.

- Also observed are variants V157I, R290H & R282N that are also C>T at mCpG, but have wild type activity. Conclude these are passenger mutations in hot spot residues.

Giacomelli et al.
High-throughput discovery/hypotheses generating approaches provide a rich source of candidate genes and interactions to be examined in follow-up “reductionalist” experiment.
Biology is complex

Many gene products act at multiple times or continuously.
- Cdc2 kinase acts at multiple points in the cell cycle.
- Notch receptor signaling acts in multiple developmental decisions (sequentially and contemporaneously).

Some (many?) gene products have diverse biochemical functions.
- Yeast and mammalian mitochondrial transcription factor B is both a transcription factor and an adenine methyltransferase.

Such complexities are currently beyond high-throughput genetic/functional methods, requiring “reductionalist” gene, pathway or process specific approaches.