Zebrafish Phenome Project

Introduction

In March, a meeting was held at the Wellcome Trust Sanger Institute in Hinxton to discuss the organization of a project to identify the function of every gene in the zebrafish genome. Within the next year, the zebrafish will join the exclusive ranks of a small number of vertebrate genomes with "finished" genomic sequences, *i.e.* nearly comprehensive coverage of the chromosomes with high quality, contiguous sequence. In contrast to other common vertebrate model organisms, zebrafish allow high-throughput analysis of phenotypes at relatively low cost. Thus it is feasible to obtain a basic phenotype analysis for every protein-coding gene. This is unprecedented for vertebrate biology and will provide a wealth of biological information for clinicians and research scientists alike.

Bioinformatic analysis and comprehensive RNA sequencing indicates that there are between 20k and 25k zebrafish genes. Approximately 80% of protein coding genes in the zebrafish genome can be assigned a putative function based on the presence of a Pfam domain. Similar to the human genome, approximately 78% of protein-coding genes possess a Pfam domain. However, most of these assigned functions are inferred and their *in vivo* roles in the organism have not been directly tested. As a general approach, genetics allows for a systematic testing of gene function based on the appearance of a phenotype when a gene has been disrupted. In model organisms ranging from yeast to mice, this approach has been proven to be an effective method for identifying genes critical in a wide variety of biological processes .

Traditional forward genetics is phenotype driven in that a measurable difference between mutant and wild-type animals is observed, then the mutated gene is identified. The main limitation of this approach is that a mutation is only identified if it generates a phenotype that would be observed during the screening process. Because of the large number of animals involved and no *a priori* knowledge of which genes are being screened, systematic, comprehensive phenotype screening of every animal from a random mutagenesis is not feasible.

Reverse genetics is a process whereby a researcher mutates a specific gene of interest and then characterizes the mutant in a systematic, and comprehensive fashion. Before the availability of whole genome sequences, it was not possible to use this approach to test every gene in a eukaryotic genome. With the release of the *Saccharomyces cerevisiae* genome in 1996, a new era of functional genomics via systematic genetics was born. Once the genome was fully sequenced, it became possible to inactivate every gene in the yeast genome. The Saccharomyces Genome Deletion Project has resulted in the most comprehensive datasets for gene function and genetic interactions of any organism. Similar systematic gene-knockout screens have been proven to be effective in both *C. elegans* and *D. melanogaster* .

In vertebrates, such systematic approaches are significantly more labor intensive and face additional logistical challenges. In 2003, an international effort to mutate all the protein coding genes in the mouse genome was launched . The knockout mouse project

(KOMP) had the primary goal of mutating each mouse gene, using gene targeting approaches and a secondary goal of phenotyping a relatively smaller number of those targeted genes. Within the next two years, there will be targeted alleles for the majority of mouse genes.

Until recently, despite the many advantages that zebrafish possessed for large-scale genetics, it did not have mutagenic approaches that could generate mutations in the majority of genes. But the expected completion of the zebrafish genome in combination with recent advances in next generation sequencing technologies and novel mutagenic approaches, have now made the possibility of a zebrafish gene knockout project possible. At the Hinxton meeting, plans were developed with two broad goals: 1) mutate every gene in the zebrafish genome and 2) perform high-throughput phenotypic analysis to functionally annotate every zebrafish gene. With the development of this resource, it would be a realistic goal to be the first completely functionally annotated vertebrate genome through phenotypic analysis.

Mutagenesis

The first phase of the zebrafish phenome project would be to identify genetic lesions in every predicted zebrafish gene, whether it is believed to be protein coding or not. There are three major approaches that can be used to generate lesions in a high throughput manner: 1) TILLING (*i.e.* re-sequencing of chemically mutagenized genomes to identify nonsense alleles), 2) mutation through insertion of DNA elements such as retroviruses or transposons, 3) targeted lesions via Zn-finger nuclease double stranded breaks and non-homologous end-joining. We will discuss the various approaches in detail.

TILLING

In 2002, it was shown that a mutation could be identified in a specific zebrafish gene by sequencing exons of the target gene from thousands of samples of mutagenized genomes . Several independent efforts have resulted in large "libraries" of mutagenized fish, either as living fish (currently approximately 16,000 fish) or frozen sperm samples with matching genomic DNA samples (currently approximately 25,000 fish). The efficacy of this approach is directly proportional to the mutation rate and the number of F_1 individuals screened. With capillary or slab-gel based sequencing or genotyping strategies, this approach is labor and consumable intensive. The advent of the next-generation sequencing technologies has significantly changed the cost structure of this approach. Using the sequencing power available to major sequencing centers such as the Sanger Institute, an estimated 184,320 100bp amplicons can be tested weekly (480 384-well plates). Based on measured mutation rates, the expected mutant discovery rates for nonsense alleles would be 250 new alleles per week, or more than 10,000 nonsense alleles per year. A five-year effort would give sufficient coverage to identify at least two nonsense alleles for the majority of protein coding genes. The first two years would deliver at least one allele for a majority of the genes. The cost effectiveness of this approach relies on the economy of scale that comes from sequence production facilities. The total cost of this effort with recovery of mutant clutches of embryos and first-pass morphological

phenotyping is ~$31M or $1,269 per gene for two nonsense allele in each of 25,000 genes. Without first-pass phenotyping the cost would be ~$19M or $750 per gene.

Insertional mutagenesis

The first demonstration of a DNA element that could be efficiently used as a mutagen in zebrafish was in 1996 by Gaiano et al . Since that time, MLV-based retroviruses have been used in a large-scale classical genetics screen and a proof-of-principle screen has been published demonstrating that retroviruses can also be used as a reverse genetics approach when coupled with high-throughput sequencing . Through an undetermined mechanism, when the provirus integrates into the first intron of a gene, the mRNA is destabilized and mRNA levels are typically <10% of normal expression levels. Approximately 20% of the proviral integrations are mutagenic events and F_1 fish carry 10 integrations on average or 2 mutagenic events per F_1 fish. Mapping retroviral integrations in 10,000 F_1 fish would generate approximately 12,000 unique mutations (given some redundancy). Overall costs of this effort would be approximately $3,000,000 or a cost of $260 per mutation.

Recently two transposable elements, Sleeping Beauty and Tol2 have been shown to efficiently transpose when injected into zebrafish embryos. Tol2 in particular has been shown to be a highly efficient transposition reaction and recent modifications have adapted Tol2 into an efficient mutagenic element through gene trap technologies . The most important aspect of the transposon constructs lie in the possibility of creating conditional alleles that are convertible through recombinase reactions through the cre/lox (or similar) systems . The integration rates are slightly lower than retroviral rates, but are still scalable at rates that can efficiently generate tens of thousands of mutageneic events. By using a combination of reporter gene expression and high throughput sequencing, it is possible to generate 10-15,000 conditional gene inactivation events in 5 years at an approximate cost of $360 per mutation, or a project total cost of $5,400,000.

Zn-finger nucleases

Two groups have demonstrated that injecting the FokI nuclease linked to zinc finger sequences designed to target specific DNA sequences into zebrafish can result in lesions at the targeted sequences . These constructs are known as zinc finger nucleases (ZFN). This approach can be highly efficient and will transmit through the germline, making genuinely targeted lesions in the zebrafish genome a reality for the first time. More recently, groups in Boston were able to show that an approach known as Oligomerized Pool ENgineering (OPEN) of ZFN's could streamline the workflow for this approach and still yield promising results in zebrafish . While a relatively new technology, the approach has tremendous promise and integrates well with the other approaches mentioned. As part of the overall strategy to create a genetic lesion in every zebrafish gene, several groups in collaboration propose to build an archive of validated "4 finger" Zn finger constructs in conjunction with computationally predicting target sequences in the 1st to 3rd coding introns of all protein coding zebrafish genes. The archive will take approximately 2 years to construct at a cost of approximately $2,000,000. The resource would then cost approximately $500 per targeted gene. Because this approach has not yet been tested for how scalable it will be

and it will take approximately 2 years to build the resource, the most likely integration of ZFN's in the Zebrafish Phenome Project would be to first use them to specifically target genes that have not been mutated with another approach. Assuming 2,000 genes are targeted by this approach, the total cost of the project would be $3,000,000.

A combination of all three major approaches would allow us to achieve complete mutagenesis of the heterozygous viable zebrafish protein-coding genome.

Phenotype Analysis

A key aspect of a systematic approach in zebrafish is the ability to perform high-throughput phenotyping on an unprecedented scale. The ultimate goal of an international zebrafish gene knockout effort should be to provide a comprehensive and searchable phenotype dataset for every gene mutation. The dataset would consist of four parts:

1) A first-pass description of morphological phenotypes during the first five days of development including images of appropriate stages where phenotypes are observed.

2) Transcriptional profiling of all mutants at an appropriate embryonic and/or larval stage to capture molecular fingerprints of individual genes and pathways.

3) A quantitative assessment of key behaviors during larval stages for all mutants.

4) Annotated 3D imaging at cellular resolution at an embryonic and a larval stage.

A strong emphasis is the requirement for the dataset to be readily mined by tools that might be developed in the future such as automatic, qualitative and quantitative image analysis. Importantly, the data will need to be presented in a format that is accessible to non-zebrafish and medical researchers. Also, phenotype data should be searchable and comparable across species.

Live assays such as morphological and behavioral phenotyping should to be carried out at the places where mutants are generated in order to reduce husbandry costs. Assays on fixed embryos can be performed elsewhere. The utility of examining adult and maternal/zygotic phenotypes is clear and would be expected to comprise the 70-80% of nonsense alleles that do not cause a morphological phenotype during the first five days of development. The value of adult and maternal phenotypes cannot be overstated, but the cost of analysis would be large and should be considered as a second phase proposal.

Because of the large scope of the phenotyping aspect of this project would most likely require coordinated efforts from several screening centers. We propose a minimum of three such centers with likely locations in the US, Europe, and possibly Asia. All phenotypic screening would, by necessity, have to be performed in parallel at each site, requiring redundant equipment. All protocols, screening "forms" and annotation would have to be standardized across all centers. Continual quality checks and annual meetings of the screening centers would facilitate common standards across all centers. One possible approach is to have at least 2 centers evaluate each mutant screened either by redundancy in the actual screening, or by "virtual screening" confirmation of annotation. Pipelines for

ensuring the correct gene is characterized in each screened family will have to be developed. Estimates of cost for screening all mutants using all the proposed approaches would be $45,000,000. Prioritized approaches are possible to reduce costs.

Proposal Summary

Here are what we consider the key points of the Phase I Zebrafish Phenome Project:

Mutagenesis:

Mutagenesis should be done in such a way to take advantage of each of the separate schemes, ultimately to obtain at least two nonsense alleles per gene:

1. A two tiered TILLING approach with relevant-exome resequencing to cover the 1st half, of the protein coding genome followed by targeted PCR product resequencing for 2nd.
2. Simultaneous recovery of insertional alleles
3. ZFN targeting for difficult genes

Phenotyping:

1. Systematic recovery of mutant animals for 1st pass phenotyping and expression analysis for those that show an embryonic phenotype.
2. Several screening centers will need to be established; this should be an international effort.
3. Deeper characterizations include behavioral screening (up to 4 behaviors), and high-resolution imaging. These will incur significant additional cost.
4. Adult and maternal mutation characterization is a desired goal but comes with significant additional cost. This would likely be a Phase II goal.

Literature Cited

Approach	Cost/mutant	Status	Pro/Con
TILLING	$750	TILLING is a well-established approach using traditional sequencing. The next-gen sequencing aspects are in the testing phase, but the principles are identical to those used in other resequencing projects, so there is a deep knowledge base to build on.	**Pro**: Highest throughput, multiple alleles **Con**: Multiple, undefined background mutations, most expensive approach
Retroviral Mutagenesis	$260 ($1,800 exons only)	Mutagenesis and mapping underway. All necessary technologies in place.	**Pro**: Inexpensive, established approaches, low background mutation rate. **Con**: Majority of mutations are hypomorphic, not null. Non-random integrations.
Transposons	$360	Pilot screens performed, multiple constructs available.	**Pro**: Easily adopted technologies. Rapid changes in constructs possible. Potential for conditional alleles. Low costs. **Con**: Only pilot-scale tests, transgenic rates currently lower than retrovirus. Genomic site preferences unknown.
Zinc-finger nucleases	$500	Proof-of-principle published. Active development of resources, proposal submitted	**Pro**: Only true targeted approach. Multiple alleles. **Con**: Significant development of approach still needed. Throughput untested.