DOI: 10.1534/genetics.109.112938

Perspectives

Anecdotal, Historical and Critical Commentaries on Genetics

The Impact of Whole Genome Sequencing on Model System Genetics: Get Ready for the Ride

Oliver Hobert

Columbia University Medical Center, Howard Hughes Medical Institute, New York, New York 10032

ABSTRACT

Much of our understanding of how organisms develop and function is derived from the extraordinarily powerful, classic approach of screening for mutant organisms in which a specific biological process is disrupted. Reaping the fruits of such forward genetic screens in metazoan model systems like Drosophila, *Caenorhabditis elegans*, or zebrafish traditionally involves time-consuming positional cloning strategies that result in the identification of the mutant locus. Whole genome sequencing (WGS) has begun to provide an effective alternative to this approach through direct pinpointing of the molecular lesion in a mutated strain isolated from a genetic screen. Apart from significantly altering the pace and costs of genetic analysis, WGS also provides new perspectives on solving genetic problems that are difficult to tackle with conventional approaches, such as identifying the molecular basis of multigenic and complex traits.

GENETIC model systems, from bacteria, yeast, plants, worms, flies, and fish to mice allow the dissection of the genetic basis of virtually any biological process by isolating mutants obtained through random mutagenesis, in which the biological process under investigation is defective. Such forward genetic analysis is unbiased and free of assumptions. The rigor and conceptual simplicity of forward genetic analysis is striking, some may say, beautiful; and the unpredictability of what one finds—be that an unexpected phenotype popping out of a screen or the eventual molecular nature of the gene (take the discovery of miRNAs as an example; Lee *et al.* 1993)—appeals to the adventurous. Even though mutant phenotypic analysis alone can reveal the logic of underlying biological processes (take Ed Lewis' analysis of homeotic mutants as an example; Lewis 1978)—it is the identification of the molecular lesions in mutant animals that provides the key mechanistic and molecular details that propel our understanding of biological processes.

The identification of the molecular lesion in mutant organisms depends on how the mutation was introduced. Classically, two types of mutagens have been used in most model systems: biological agents such as plasmids, viruses, or transposons whose insertions disrupt functional DNA elements (either coding or regulatory elements) or chemical mutagens, such as ethyl methane sulfonate (EMS) or *N*-ethyl *N*-nitroso urea (ENU), that introduce point mutations or deletions. Point mutation-inducing chemical mutagens are in many ways a superior mutagenic agent because their mutational frequency is high and because the spectrum of their effects on a given locus—producing hypomorphs, hypermorphs, amorphs, neomorphs, etc.—is hard to match by biological mutagens. Moreover, chemical mutagens do not display the positional bias of many biological agents. In addition, point mutations in a gene are often crucial in dissecting the functionally relevant domains of the gene product. In spite of the advantages of chemical mutagens, model system geneticists often prefer biological mutagens simply because the molecular lesions induced by those agents are characterized by the easily locatable DNA footprint that these agents generate. In contrast, the location of a point mutation (or deletion) has to be identified through conventional mapping strategies, which tend to be tedious and time consuming. Even in model systems in which positional cloning is quite fast and straightforward (*e.g.*, *Caenorhabditis elegans*, which has a short generation time and a multitude of mapping tools available), it nevertheless is a significant effort that can occasionally present hurdles that are difficult to surmount (*e.g.*, if the gene maps into a region with few genetic markers that allow for mapping). These difficulties explain why RNAi-based "genetic screens" have gained significant popularity in *C. elegans*; they circum-

Address for correspondence: Columbia University Medical Center, Department of Biochemistry, 701 W. 168th St., HHSC 724, New York, NY 10032. E-mail: or38@columbia.edu

vent mapping and reveal molecular identities of genes involved in a given process straight away (Kamath and Ahringer 2003). However, genes and cells show differential susceptibility to RNAi; off-target effects and lack of reproducibility can be a problem, and the range of effects that RNAi has on gene activity is generally more limited compared to chemically induced gene mutations.

The recent application of next generation, deep sequencing technology (see Bentley 2006; Morozova and Marra 2008 for technology reviews) is beginning to significantly alter the landscape of genetic analysis as it allows the use of chemical mutagens without having to deal with its disadvantages. Deep sequencing technology incorporated into platforms such as Illumina's Genome Analyzer or ABI's SOLiD, allows one-shot sequencing of the entire model system's genome, resulting in the detection of mutagen-induced sequence alterations compared to a nonmutagenized reference genome. Proof-of-concept studies have so far been conducted in bacteria, yeast, plant, worms, and flies, all published within the last year (Sarin *et al.* 2008; Smith *et al.* 2008; Srivatsan *et al.* 2008; Blumenstiel *et al.* 2009; Irvine *et al.* 2009; Rigola *et al.* 2009). Many more studies are under way; for example, since our first proof-of-principle study (Sarin *et al.* 2008), my laboratory has identified the molecular basis of >10 *C. elegans* strains defective in neuronal development and homeostasis (V. Bertrand, unpublished data; M. Doitsidou, unpublished data; E. Flowers, unpublished data; S. Sarin, unpublished data).

The advantages of whole genome sequencing (WGS) are obvious. The process is extraordinarily fast with the sequencing taking only ~5 days and the subsequent sequence data analysis only a few hours, particularly if the end user employs bioinformatic tools customized for mutant detection (Bigelow *et al.* 2009). The process is also remarkably cost effective. For example, a *C. elegans* genome can be sequenced with a required sequence coverage of ~10 times for <$2,000 in reagent and machine operating costs. The capacity of deep sequencing machines—and hence the costs associated with sequencing a genome—apparently follow Moore's law of doubling its capacity about every 2 years, like many technological innovations do (Pettersson *et al.* 2009). That is, the <$1,000 genome for *C. elegans* (~100-Mb genome) and Drosophila (~123-Mb genome) is just around the corner and other models will sooner or later follow suit. The cost effectiveness becomes particularly apparent if one compares the cost of WGS to the personnel and reagent costs associated with multiple-month to multiple-year mapping-based cloning efforts.

WGS identifies sequence variants between a mutated genome and a premutagenesis reference genome. Chemical mutagens randomly introduce many mutations in the genome and, therefore, the phenotype-causing sequence variant needs to be identified as such out of a large pool of sequence variants. Sequence variants that have no impact on the phenotype can be outcrossed before sequencing or eliminated through some rough mapping of the mutation, which allows the experimenter to focus only on those variants contained in a specific sequence interval. Ensuing functional tests such as transformation rescue or phenocopy by RNAi and the availability of other alleles of the same locus are critical means to validate a phenotype-causing sequence variant (Sarin *et al.* 2008). The latter approach—the availability of multiple alleles of the same locus—is in many ways the most powerful one to sift through a number of candidate variants revealed by WGS. In this approach, candidate loci revealed by WGS are resequenced by conventional Sanger sequencing in allelic strains and only those that are indeed phenotype causing will show up mutated in all allelic variants of the locus (Sarin *et al.* 2008). The availability of multiple alleles of a locus is highly desirable for many aspects of genetic analysis anyway and therefore does not represent an additional and specific burden for undertaking a WGS project.

The importance of WGS on model system genetics will be substantial and wide ranging. Speed and cost effectiveness means that the wastelands of genetic mapping can be trespassed fast enough to allow an experimenter to multitask a whole mutant collection in parallel, thereby closing in on the "holy grail" of genetic analysis—the as-complete-as-possible mutational saturation of a biological process and the resulting deciphering of complete genetic pathways and networks. What will become limiting steps are not any more the tediousness of mapping, but rather the effectiveness with which mutant collections can be built. Novel technologies that involve machine-based, semiautomated selection of mutant animals have been developed over the past few years to study a variety of distinct biological processes in several metazoan model systems, *e.g.*, *gfp*-based morphology or cell fate screens in worms (Crane *et al.* 2009; Doitsidou *et al.* 2008) or behavioral screens in flies (Dankert *et al.* 2009) and are important steps in this direction. Such an "industrial revolution" of genetic screening (*i.e.*, the mutant selection part, followed by WGS) moves us geneticists away from, not into the trenches of factory life and frees us up to do what we should like to enjoy most—thinking of designing interesting screens, seeing how genes interact, and interpreting it all.

Another important impact of WGS is that it will allow tackling problems that were previously hard to deal with. For example, the tediousness of following subtle phenotypes, low penetrance phenotypes, or phenotypes that are cumbersome to score often hampers positional cloning approaches that rely on identifying rare recombinants in a large sibling pool. Moreover, many genetic traits such as behavioral genetic traits are very sensitive to genetic background and are therefore also

often hard to map in the conventional way. WGS hones in on candidate genes straight away. Taking this notion a step further, WGS will also be able to get at the molecular basis of multigenic traits and quantitative trait loci, which again are hard to molecularly identify through conventional mapping strategies; a proof-of-principle study has made this point already in bacteria (Srivatsan *et al.* 2008). In principle, such multigenic traits may have been mutationally induced or could be present in natural variants of a species, which provides intriguing perspective for the population geneticist.

Model organisms of biological interest that were previously relatively intractable for classic genetic mutant analysis due to the absence of genetic markers or other practical problems such as prohibitive generation times, may also now be movable into the arena of genetic model systems, through the WGS-mediated molecular analysis of mutagen-induced variants or through the study of natural variants.

The sequencing of human cancer genomes has already begun to illustrate the impact of WGS on human genetics (Campbell *et al.* 2008; Ley *et al.* 2008). However, those human WGS studies illustrate why model systems will continue to be extremely important—their experimental accessibility allows us to address which of the many variants detected by WGS is indeed the phenotype-causing one.

The message to model system geneticists is clear: Get access to a deep sequencer, buckle up, and get ready for the ride.

The author thanks members of his lab, particularly S. Sarin, H. Bigelow, E. Flowers, and M. Doitsidou, for establishing the WGS approach in the laboratory and several colleagues for comments on this manuscript. Work in the author's laboratory is funded by the Howard Hughes Medical Institute and the Natonal Institutes of Health.

LITERATURE CITED

Bentley, D. R., 2006 Whole-genome re-sequencing. Curr. Opin. Genet. Dev. **16:** 545–552.

Bigelow, H., M. Doitsidou, S. Sarin and O. Hobert, 2009 MAQGene: Software to facilitate C. elegans mutant genome sequence analysis. Nat. Methods **6:** 549.

Blumenstiel, J. P., A. C. Noll, J. A. Griffiths, A. G. Perera, K. N. Walton *et al.*, 2009 Identification of EMS-induced mutations in *Drosophila melanogaster* by whole-genome sequencing. Genetics **182:** 25–32.

Campbell, P. J., P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li *et al.*, 2008 Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. **40:** 722–729.

Crane, M. M., K. Chung and H. Lu, 2009 Computer-enhanced high-throughput genetic screens of C. elegans in a microfluidic system. Lab Chip **9:** 38–40.

Dankert, H., L. Wang, E. D. Hoopfer, D. J. Anderson and P. Perona, 2009 Automated monitoring and analysis of social behavior in Drosophila. Nat. Methods **6:** 297–303.

Doitsidou, M., N. Flames, A. C. Lee, A. Boyanov and O. Hobert, 2008 Automated screening for mutants affecting dopaminergic-neuron specification in C. elegans. Nat. Methods **5:** 869–872.

Irvine, D. V., D. B. Goto, M. W. Vaughn, Y. Nakaseko, W. R. McCombie *et al.*, 2009 Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing. Genome Res. **19:** 1077–1083.

Kamath, R. S., and J. Ahringer, 2003 Genome-wide RNAi screening in Caenorhabditis elegans. Methods **30:** 313–321.

Lee, R. C., R. L. Feinbaum and V. Ambros, 1993 The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell **75:** 843–854.

Lewis, E. B., 1978 A gene complex controlling segmentation in Drosophila. Nature **276:** 565–570.

Ley, T. J., E. R. Mardis, L. Ding, B. Fulton, M. D. McLellan *et al.*, 2008 DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature **456:** 66–72.

Morozova, O., and M. A. Marra, 2008 Applications of next-generation sequencing technologies in functional genomics. Genomics **92:** 255–264.

Pettersson, E., J. Lundeberg and A. Ahmadian, 2009 Generations of sequencing technologies. Genomics **93:** 105–111.

Rigola, D., J. van Oeveren, A. Janssen, A. Bonne, H. Schneiders *et al.*, 2009 High-throughput detection of induced mutations and natural variation using KeyPoint technology. PLoS One **4:** e4761.

Sarin, S., S. Prabhu, M. M. O'Meara, I. Pe'er and O. Hobert, 2008 Caenorhabditis elegans mutant allele identification by whole-genome sequencing. Nat. Methods **5:** 865–867.

Smith, D. R., A. R. Quinlan, H. E. Peckham, K. Makowsky, W. Tao *et al.*, 2008 Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. **18:** 1638–1642.

Srivatsan, A., Y. Han, J. Peng, A. K. Tehranchi, R. Gibbs *et al.*, 2008 High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. PLoS Genet. **4:** e1000139.