Seminars in Cell & Developmental Biology 23 (2012) 230–236

Contents lists available at SciVerse ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

Understanding genetic variation and function- the applications of next generation sequencing

Richard J. Harrison*

East Malling Research, New Road, East Malling, Kent, ME19 6BJ, United Kingdom

ARTICLE INFO

Article history:
Available online 20 January 2012

Keywords:
Next generation sequencing
Genetic diversity
Population genomics
QTL mapping

ABSTRACT

Next generation sequencing (NGS) technology has had a transformatory effect upon population-level studies linking genetic variation to gene function. In this review, I briefly describe recent studies that have used top-down genome scanning and population genetic approaches to identify loci under recent selection, as well as some examples of how large NGS datasets can be deployed to detect the total level of deleterious, neutral and advantageous variation present in standing genetic variation. I then explore studies that have used some of these approaches to study gene function along with advances in sequencing populations under selection, QTL mapping techniques and emerging methodologies utilising targeted capture and NGS.

Contents

1. Introduction

Despite the significant advances in phenotyping and genotyping technologies in recent years, it is still challenging to link both phenotype to genotype (for example, the 'missing' heritability problem [1]) and genotype to function. This review will examine the impact of NGS technologies on studies of genetic variation with a specific emphasis on how studies of diversity within populations (through population genetics and quantitative genetics methods) have provided information about molecular function and the evolution of function in molecular systems. Studies of human disease and evolution will be omitted from this review, as will the majority of association studies, as many of the published examples do not directly use sequencing (the 1000 genome project [2] being a notable exception) and this work is adequately reviewed elsewhere [3,4]. Instead, this review will focus on both model and non-model systems in which a wide variety of NGS approaches have been deployed.

To understand the level of genetic variation in a species, an appreciation of the forces that affect the evolution of a species must be understood, a research field known as population genetics. A central aim of this research field is to ask what effects mutation, selection, genetic drift, demographic fluctuations, the environment and population subdivision have on observed levels of genetic diversity within a species? The qualitative and latterly quantitative description of molecular systems, has not, for the most part, drawn on population genetic approaches to understand gene or system function. This has probably been to the detriment of both fields,

* Tel.: +44 1732 523747.
E-mail address: richard.harrison@emr.ac.uk

1084-9521/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.semcdb.2012.01.006

however, over the last decade, large advances have been made in drawing together these research disciplines. The convergence of these two disciplines is, in part, driven by the desire to understand the genetic component of many common human diseases. The rush towards more personalised medicine through advances in mapping disease and lifestyle-affecting traits is also a major reason that NGS technologies have advanced so quickly, as benchtop sequencers offer the promise of faster and cheaper diagnostic tools, developed from (among other things) genome-wide association studies (GWAS).

The recent explosion in human GWAS have documented (to some extent) the genetic basis of a range of common human diseases and have also driven attempts to understand how population-level processes, such as genetic bottlenecks, population subdivision and environmental challenges (such as dietary shifts and diseases) have affected the total level and the type of variation segregating in the population. The underlying model of population structure is crucial to the accurate association between phenotype and genotype, as non-selective processes such as population subdivision can lead to spurious associations between genotype and phenotype. Furthermore, as has been widely documented, the 'missing heritability' – the disconnect between the amount of narrow sense heritable variation in traits explained by familial studies (high) and GWAS (low) [5] will increasingly draw on population genetics methods that offer the potential to predict how much variation in a particular trait is driven by deleterious, low-frequency variation, which is totally missed by current GWAS approaches [6,7]. Furthermore, all of these advances will be driven by developments in NGS technologies, which offer the ability to sequence to great depth (and hence accuracy) and latterly to target certain portions of the genome for 'fine-mapping' approaches to identify causal variants, which can be targeted for future study.

2. Population genomics

2.1. Genome-scale population genomics

The key areas in which NGS is aiding population genetics approaches to studies of diversity and function are the genome-wide scans for the action of natural selection (so-called top-down studies [8]) and the elucidation of the distribution of mutational effects, also known as the distribution of fitness effects (DFE) [10].

As with the first eukaryotic genome sequencing project [9], the first whole-genome resequencing project (a portion of which was NGS data) was carried out using *Saccharomyces cerevisiae* as a model organism [10]. This served as a pilot study to the larger 1000 human genomes project [2]. It revealed in fine detail the population structure of both *S. cerevisiae* and a wild sister species, *Saccharomyces paradoxus* and the large differences in population structure observed between the two species that has primarily been driven by human-aided dispersal of *S. cerevisiae* and secondary contact between allopatrically isolated populations [10]. Much of this was already known from isozyme studies and multilocus genotyping approaches for *S. paradoxus* [11], however the fine-detail of ancestral populations of *S. cerevisiae* were revealed in a greater resolution.

Genome-wide McDonald Kreitman tests [12] on the *S. cerevisiae* and *S. paradoxus* populations yielded little evidence for historical positive selection (i.e. selective events that have gone to fixation), indicating that at the protein-level positive selection is either rare or difficult to detect. McDonald Kreitman approaches, while useful are a fairly blunt instrument, as the presence of deleterious mutations can cause the test to be overly conservative. Indeed, if there is a large proportion of slightly deleterious mutations segregating in the population, then evidence for positive selection will be missed [13]. Correction factors can be applied, the simplest of which is to remove low frequency non-synonymous variants from polymorphism data [14], however, this itself alters estimates of the fraction of substitutions fixed by positive selection [13]. However, the fact that there was little evidence for positive selection at the amino acid level is indicative of the fact that the majority of historical fixations in both species are either neutral or slightly deleterious. Indeed, follow-on studies revealed that the strength of purifying selection, in some populations of *S. cerevisiae* was markedly lower than in others, indicating that within some populations there is a larger fraction of deleterious variation segregating in the population than in others, due to the smaller effective population size [15]. It should be noted that this later work excluded the NGS portion of the data due to issues with hybrid assemblies and potential platform-specific biases introduced by two different sequencing technologies.

2.2. NGS and the distribution of mutational effects

Although this is well covered in other reviews [16], the problem with accuracy in NGS approaches and population genetic inferences is key to the utilisation of this technology in population genetic approaches to accurately interprete observed levels of genetic diversity at the population level [17]. Much effort has been devoted to the analysis of NGS data, using statistical modelling in order to maximise the accuracy of allele frequency data. Two recent studies in particular deal with this problem and present similar statistical methods to infer allele frequency data directly from population samples, using maximum likelihood methods to estimate the allele frequency at any given site [18,19]. The authors of one of these studies, Keightley and Halligan deployed their maximum likelihood (M.L.) method on low coverage NGS data of 57 humans to determine the site frequency spectra (SFS) which is needed to infer the shape of the distribution of mutational effects. They found, as with earlier studies, that the majority new mutations are strongly deleterious, with only about 24% evolving effectively neutrally, broadly in agreement with other, earlier studies [18,20]. More interesting, was the discovery that about 30% of synonymous sites were under purifying selection (about 11% under strong purifying selection). This is interesting, as some forms of weak selection on synonymous sites (such as selection for biased codon usage) have been shown to be ineffective in humans [21] suggesting that other forms of selection on synonymous sites are operating. Most concerning about this study was the finding that the sequencing error-rate parameter in the M.L. model was higher than the observed value of the population-scaled mutation rate, $\boldsymbol{\theta_\pi}$, at zero-fold sites. In other words, as the levels of genetic diversity in humans are comparatively low due to a historic population bottleneck, sequencing error may be making a relatively larger contribution to estimates of sequence diversity than in other species, making estimates of key population genetic parameters less accurate. If estimates of the distribution of mutational effects are flawed due to sequencing error, then there will be a knock-on effect on estimates of the proportion of adaptive mutations fixed by natural selection. Thus, improvements in NGS technologies, or different strategies (such as longer reads and much deeper sequencing) are needed to alleviate these problems in species, such as humans, where levels of diversity are low.

A very recent study in *Arabidopsis thaliana*, a precursor to the 1001 genome project has recently documented a large range of high-confidence mutations across 80 genomes of *A. thaliana*, drawn from a pan-European sample of accessions including a vast number of structural variations (small indels and deletions) [22]. As well as this, in line with previous studies in other species [15], examination of the site-frequency spectrum across a range of subpopulation sizes highlighted a clear relationship between the estimated