



Review

The revolution of whole genome sequencing to study parasites



Sarah Jayne Forrester, Neil Hall*

Department of Comparative and Functional Genomics, University of Liverpool, Crown Street, Liverpool L69 7ZB, United Kingdom

ARTICLE INFO

Article history:
Available online 8 August 2014

Keywords:
Genome sequencing
Genome structure
Antigenic variation
Comparative genomics
Drug target

ABSTRACT

Genome sequencing has revolutionized the way in which we approach biological research from fundamental molecular biology to ecology and epidemiology. In the last 10 years the field of genomics has changed enormously as technology has improved and the tools for genomic sequencing have moved out of a few dedicated centers and now can be performed on bench-top instruments. In this review we will cover some of the key discoveries that were catalyzed by some of the first genome projects and discuss how this field is developing, what the new challenges are and how this may impact on research in the near future.

© 2014 Published by Elsevier B.V.

Contents

1. Introduction	77
2. Genome structure	78
3. Antigenic variation	78
4. Comparative genomics and phylogenomics	79
5. Population genomics	79
6. Metabolic pathways and drug targets	79
7. Future prospects	80
References	80

1. Introduction

The first genome ever to be sequenced was that of a parasite, albeit a simple bacteriophage, phiX174 [46]. This pioneering work heralded a new era for biology, whereby scientists could start to unravel molecular biology of entire systems rather than focusing on individual genes and proteins. However, it was not until 1995 the first bacterial genome was sequenced and as technology accelerated it was only six years later that the human genome was published. By this time, a number of genome projects of human parasites had been initiated. The *Plasmodium falciparum* genome was first conceived in the mid 1990s, against a backdrop of increasing drug resistance high levels of transmission, the genome project was seen as a way to invigorate research for new drug targets and potential vaccine candidates. In late 2001, the first eukaryotic parasite genome was published, that of the microsporidian

parasite *Encephalitozoon cuniculi* [21], and the following year saw the publication of the *P. falciparum* genome [16]. By the end of 2005, numerous parasite genomes had been published, including veterinary parasites and model species. These resources were able to transform the way in which are able to study parasites and investigate how they interacted with their hosts.

In 2005, the first high-throughput genome sequencing methodology was published [34], heralding the advent of a raft of new technologies that would be later dubbed “next-generation sequencing” [36]. Next generation sequencing technologies differ from Sanger sequencing in that these methods are ‘massively parallel’, with read numbers generated that are several orders of magnitude higher than seen with capillary sequencing. Per-base costs are also significantly lower. The initial disadvantage was significantly lower read-length, with early technologies generating reads of as little as 21 bases, compared to up to 1 kilobase reads from Sanger sequencing. However, these technologies allowed genomic analysis to move out of large-factory genome centers and into the hands of bench scientists for hypothesis-driven research.

In this review we will overview how genomics has influenced the field of parasitology. While the term “genomics” can be used to

* Corresponding author. Tel.: +44 01517954516.

E-mail addresses: S.J.Forrester@student.liverpool.ac.uk (S.J. Forrester),
Neil.Hall@liverpool.ac.uk, neilhall@liv.ac.uk (N. Hall).

describe a multitude of approaches, this review will focus primarily on whole genome sequencing, specifically highlighting how some of the early genome projects have informed our understanding of parasite biology. We will also outline how the rapidly changing field of genomics is likely to transform our technical approaches and impact on our understanding of parasite biology in the future.

2. Genome structure

In 2005, the trypanosomatid genomes of *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi* were sequenced and reported a surprisingly conserved core gene set despite significant differences in their lifestyles and highly divergent subtelomeres, that contained many of the surface antigens [6,11,24]. As expected, the lifestyles of these parasites is reflected in their genome, with intracellular parasites *L. major* and *T. cruzi* displaying a greater degree of similarity in comparison to the extracellular parasite *T. brucei*. These genomes allowed us to look at how differences in immune evasion strategies are reflected in the accessory parts of the genome with the main difference being in the abundance of species-specific surface antigen proteins. This pattern of central conservation and telomeric diversity has now been observed in many other species, including the fungal parasite, *Pneumocystis carinii*, the apicomplexan parasite *Babesia bovis* and several helminth species [8,17,49].

A common feature of parasitism is the ability to obtain nutrients from the host and therefore lose genes for processes that are no longer required. This is often reflected by a reduction in genome size, and many parasites have smaller genomes than their free-living relatives [45]. An extreme example of this would be the genome of the microsporidian parasites *Encephalitozoon caninuli* and *Trichomonas hominis* [21,27]. An exception to the genome reduction rule is the diplomonad *Trichomonas vaginalis*, a sexually transmitted parasite, and currently the largest protozoan pathogen to be sequenced, with a 160 Mb genome [10]. This genome has an unprecedented number of repeats and transposable elements, which account for approximately two thirds of its genome. It has expanded gene families and a number of suspected prokaryote to eukaryote lateral gene transfers. Despite its large genome size, it still has the reduced metabolic capabilities seen in other parasites. The reason for *T. vaginalis* having such a large genome is not clear but it could be an artifact of weak selection allowing the expansion of selfish DNA elements.

Despite the insight gained into genome structure from the multitude of genome sequences generated, there are still relatively few genomes that have been fully sequenced to generate a high-quality reference. For example, the *T. brucei* genome core was published in 30 contigs [6] while the genome of the related cattle-infective *Trypanosoma congolense* was published in 3181 contigs [25]. This pattern is becoming increasingly common as the cost for generating whole genome sequence shotgun data decreases while the costs for closing gaps in raw assemblies are still very high. It is likely that, had the *P. falciparum* or *T. brucei* genomes been sequenced using short read shotgun approaches and had not been manually finished by teams of scientists who meticulously closed gaps, key insights into the genome architecture and their antigen repertoires would have been missed. More recently, the use of single molecule real time sequencing of bacterial genomes has presented a possible solution to this, as complete closed genomes have been obtained directly from shotgun sequencing [29]. While bacterial genomes are simpler than many parasite genomes, if this technology continues to develop, then there may be good prospects for complete parasite genomes to be generated at a much reduced cost.

3. Antigenic variation

Antigenic variation is utilized by a wide range of pathogens in order to evade the host's immune system. Due to the frequency of switching of antigens, the mechanisms used to generate this variation are often hard to study using traditional genetic techniques. However, genome sequencing has revealed a lot about the diversity and organization of the antigen families and how they have evolved. The ability to exploit genome sequence data has also provided insights into the mechanisms by which switching occurs.

A common feature of pathogen genomes is that the surface antigen genes often reside in subtelomeric locations. This may be adaptation to enable greater non-homologous recombination between members of gene families. Genes within the interstitial regions are very restricted in their ability to recombine, whereas subtelomeric genes are capable of ectopic recombination, allowing the antigen genes to generate a high level of diversity [3].

Sequencing of the *T. brucei* genome led to the discovery that only 5% of variable surface glycoproteins (VSGs) were encoded in the genomes as complete genes; the majority being pseudogenes, which probably generate new variants through recombination [33]. This has also been seen in several genera including bacteria from the *Ehrlichia* genus and protozoans such as *Babesia* spp [8].

The telomeric location of the bloodstream expression sites was known prior to the *T. brucei* genome project and in 2001, Pays et al. initially described the structure of the bloodstream expression sites (BES) and the arrangement of Expression Site Associated Genes (ESAGs) within it [41]. However, subsequent systematic genomic studies determined the number of the BESs and inter-BES differences [5,22]. These studies demonstrated that despite the mosaic nature of both the ESAGs, and the BESs as a whole, the structure of the BES is relatively conserved, painting a complex picture of highly dynamic sites as products of extensive recombination but with selective forces imposing some rules on a minimal gene set.

In *P. falciparum*, antigenic variation is generated, in part, by the VAR family of genes, which consists of ~60 members per strain, and diversity is increased through mechanisms such as mosaicism, similar to those seen in *T. brucei* [20,47]. When the complete *P. falciparum* genome sequence was published [16], an inventory was made of the entire VAR gene complement of the genome, demonstrating that the VAR genes with similar domain structures were not randomly distributed but certain types were more likely to be telomere proximal. It also demonstrated that the VAR genes were associated with one of three conserved upstream sequences (UPS). Later studies have demonstrated that these sequences are integral to the regulation of antigenic variation [30].

Antigenic variation has evolved many times in different systems. Key mechanisms underlying this process are the generation of variation, silencing of non-expressed copies and switching of active copies. Genome sequencing has uncovered a number of convergent mechanisms that enable the generation and maintenance of antigenic diversity as well as antigen switching. However the detailed molecular processes regulating expression are likely to be highly species specific. As sequencing has matured as an assay, new methods such as RNAseq, ChIPseq, and CAGE [23,26] can be employed to understand these processes in detail. More recently genomic methods have been developed for studying long-range interactions between chromosomal loci that are potentially responsible for regulating activity [31]. Hence we can move from descriptive observations about organization and repertoire of antigen families to hypothesis driven research that can give a mechanistic view of the process.

Download English Version:

<https://daneshyari.com/en/article/5915426>

Download Persian Version:

<https://daneshyari.com/article/5915426>

[Daneshyari.com](https://daneshyari.com)