4

Genomics and expected benefits for vector entomology

Christos Louis[#]

Abstract

This paper summarizes the gains expected for vector entomology from the acquisition of the genome sequence of disease-transmitting arthropods. The results of this kind of high-throughput science, especially the direct consequences that could be summarily described as post-genomic activities, may lead to a better understanding of the biology of the vectors, including population studies, interactions with the disease agents and, finally, the direct development of tools, biological or bioinformatics-based, to be used in their control.
Keywords: genome sequencing; genome mining; comparative genomics

The facts and the prospects

Although the etymology of the term genomics is obvious, in contrast, its definition is fairly vague. Interestingly, unless one would encompass in it the notion of 'high-throughput research', Bridges' pioneering work on the cytogenetic mapping using the polytene chromosomes of *Drosophila* (Bridges 1935) would definitely qualify as genomic research, possibly marking the beginning of the discipline. The particularly tedious and time-consuming closing of gaps in the whole genome sequence (WGS) of the same organism, on the other hand, can hardly be described as 'high throughput', thus its inclusion in the category of '*-ics*' science could theoretically merely be done as a typical example of 'post-genomics'. In spite of these philological considerations genomics has not only found its place in biological research, even more so, it represents a constantly expanding field, also due to the continual development of new biotechnological, technological and informatics-based tools.

The first reported completion of the sequence of a large segment of eukaryotic DNA, that of chromosome XI of *Saccharomyces cerevisiae* (Dujon et al. 1994), is now about ten years old. Initiated in the late '80s, the WGS of brewers yeast, with a size of about 12 Mb was completed in a time frame of a little less than ten years (Goffeau et al. 1996). *D. melanogaster*'s genome, approximately 10 times larger, took about 5 years to be 'finished' (Adams et al. 2000), while the first draft of the complete WGS of *Anopheles gambiae*, with a genome size about twice as large as that of the fruit fly, took less than 18 months, counting from the date the programme was officially launched till its publication (Holt et al. 2002). This increase of speed also reflects the method chosen for determining the WGS, which was largely switched from clone-based (e.g. whole cosmids or BACs (Bacterial Artificial Chromosomes)) to a whole-genome shotgun approach during the fruit-fly project. Obviously, what

[#] Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology, Heraklion, Crete, Greece. E-mail: louis@imbb.forth.gr

genomics specialists call ‘complete’ is also a matter of definition. For example, four years after the publication of the full sequence of *D. melanogaster* (Adams et al. 2000) gaps in the sequence are still present, although these have now been reduced to 23 (see http://www.fruitfly.org/annot/release4.html), while the ‘finished’ mosquito genome is still made out of thousands of contigs and scaffolds, the longest of which is, nevertheless, several Mb long.

There is no question that the completion of the *Drosophila* genome gave a major thrust to genetic research. This is best exemplified by the number of papers that have been published in the three calendar years following the publication of the Adams et al. (2000) report. Searching Pubmed with the keyword “*Drosophila melanogaster*” yields 6,052 entries from this period or, in other words, one fourth of all papers that can be retrieved from the database with the same keyword and no restriction for the time of publication. For direct comparison, it should be stated that the corresponding figures for the three years preceding the ‘complete genome’ are 3,293 papers, or about 14% of the total. The more than 83% increase in scientific output can only be attributed to the availability of the WGS, information that is exploited not only by ‘fly labs’ but by researchers working with different experimental systems as well. It is easy to imagine what this wealth of information means for the understanding of the biology in general and that of the fruit fly in particular.

The publication of the complete *An. gambiae* genome sequence is much more recent and three-year statistics are not yet possible; yet, a similar trend, i.e. a significant increase of published papers dealing with the African malaria mosquito, is already apparent. Whether this increase is fully owed to the *Anopheles* WGS cannot be determined easily since an upward trend was already discernible before the Holt et al. (2002) paper: during the last five years some 600 papers described results dealing with the world’s most important malaria vector; strikingly, this number represents a little less than half of all of the Pubmed entries that are retrievable using *An. gambiae* as the sole search criterion. In other words, malaria entomology is experiencing a small boom that started in the 1990s. A few examples justify this statement. While the development of a genetic map of the fruit fly was initiated more than one hundred years ago, it was only in the previous decade, using genomic tools such as microsatellite markers, that a useful recombination map was worked out for the African malaria mosquito (Zheng et al. 1993; 1996). These microsatellites in turn helped give an impetus to population biology (see, for example, Lehmann et al. 2003; Tripet et al. 2003) since they could be translated into easily scored genetic markers. Furthermore, attempts to understand the molecular interactions between the mosquito vector and malaria parasites were, with a few exceptions, initiated only during the previous decade, 10-15 years after the advent of the recombinant-DNA era. Recently, these have been intensified, especially after the acquisition of the WGS. This becomes more apparent in the case of the study of the immune system (Levashina 2004), a physiological apparatus that could potentially be put in use for the development of antiparasitic strategies in the vector (Hemingway and Craig 2004). Finally, the WGS itself could be seen both as an example of this research boom and as a means of sustaining this increased research effort.

It is naturally very difficult, or perhaps even impossible, to translate the impact of the number of scientific publications and correlating it to the importance of the results obtained. This generally true fact may be even more critical in the case of applied or semi-applied sciences such as entomology in general, and more specifically malaria entomology. If one were to describe in only a few words the benefits that whole-genome sequencing offers to the advantage of biology, initially this could be

summarized as the discovery of genes, something that could ultimately lead to the better understanding of any given organism. In the case of disease vectors it is recognized that gene discovery could ultimately lead to the development of potentially novel insect-based intervention mechanisms that are based on molecular mechanisms elucidated by genomic and, especially, post-genomic dissection. This would also be helped by the comprehension of interactions between the two and potentially even three organisms (i.e. the vector and both the vertebrate and invertebrate hosts), again something that will become easier through the availability of WGS of all "partners". Finally, the use of the novel genomics-derived tools in the study of populations and, ultimately, the epidemiology of disease could also contribute greatly towards the development of novel insect control approaches.

What are the concrete effects that one can expect from the availability of the genome sequence of *An. gambiae*? A golden bullet should not be expected as an outcome, but a series of silver ones may be a real possibility. A relatively long list of applications is led by the already mentioned understanding of the molecular interactions between *Plasmodium* parasites and the mosquito in the latter's key tissues, i.e. midgut and salivary gland (Alavi et al. 2003; Siden-Kiamos and Louis 2004). Although their existence is assumed, *bona fide* receptors for the parasites have not yet been identified in either of them. It is clear that their potential recognition and detailed study would help devise molecular interventions in order to stop the transmission of *Plasmodium* by anophelines. More or less along the same path, the better understanding of the mosquito's immune system and the way that this could be enhanced in order to attack invading parasites is also one of the goals of post-genomic research. The list obviously cannot stop here, and conceivably the genome can be mined in order to find metabolic pathways that can be used to stop the development of *Plasmodium* in the insect host. In *Plasmodium*, pathways have already been identified and antimalarial drugs are already being developed based on the genomic information (see for example Wiesner, Borrmann and Jomaa 2003). In analogy, in mosquitoes one could potentially think of either novel targets for insecticidal chemicals or for molecules that could block directly the development of the parasite in the insect (Craig et al. 2003).

A further field in which genomic research can find immediate use is that of population biology and genetics. The microsatellite markers preceded the WGS of *An. gambiae* and, as already mentioned, they have helped substantially increase our understanding of the malaria mosquitoes' evolution, ecology and population structure (see Barker 2002). Through the use of single-nucleotide polymorphisms (SNP markers) that the WGS offers as a 'by-product', the availability of genetic markers is now enhanced manifold (Marth et al. 1999; Brumfield et al. 2003).

The potential ease of the SNP analysis brings into discussion a different aspect, namely that of the expansion of genomic analysis into other disease-vector species. The increased research output cited in the beginning of this paper predominantly concerned *An. gambiae* and, to a lesser degree, the non-malaria vector *Aedes aegypti*. These two mosquitoes were established, in a sense, as the model systems for vector biology although, of course, important findings were also described for other vectors such as, for example, the development of germ-line transformation. This latter technology, by the way, although not to be discussed further here, is to be considered equally important as genomics for the advancement of vector biology (Jacobs-Lorena 2003). In addition to *An. gambiae*, prominent among other African malaria vectors are *An. arabiensis* and *An. funestus*, for which genomic data have started accumulating, even though no genome project in the proper sense of the work has been launched.

Asian and American malaria mosquitoes, in contrast to the ones mentioned, lag behind. Looking beyond malaria, genome projects have been initiated for *Ae. aegypti* and the tsetse fly, while *Culex pipiens* is also being discussed, being a vector of the emerging West Nile Virus infection. It is hoped that this accumulation of data will also help bring forward the scientific knowledge pertaining to the other tropical diseases that cost disabilities worldwide. It should be noted here that the relative speed of data acquisition for these 'new' research objects is expected to be even higher than was the case for *Anopheles*. This is not only because WGS, as stated above, is helped by new technological developments, but also by the fact that the available WGS data for insects have a direct effect on the strategies to analyse new related organisms, making their pertinent study much easier.

A last item dealing directly with the sequencing of whole genomes was not addressed so far. This refers to biological databases in general and genome databases in particular. It is a fact that the large amount of information that is obtained by WGS projects cannot be handled by end-users unless sophisticated databases are put in place. This fact was demonstrated early on by FlyBase (The FlyBase Consortium 2003), a database that, since its inception in the late 1980s, has compiled and stored all information dealing with *Drosophila* (http://flybase.bio.indiana.edu/). By now 'life without FlyBase' is no longer possible for the fruit-fly researcher, since all genetic, cytogenetic and biological facts from as early as the 17th century (Metzel 1684) are included in it, as well as every information that has come out of the finished genome, incorporating the annotation of the so-called release 3.0 (Celniker et al. 2002).

Databases should no longer be viewed as simple storage devices using minimal search facilities. This is true not only for the classical sequence databases such as EMBL and Genbank, but also it is even more so the case for refined databases that directly handle genome data. The *Anopheles* genome, for example, has been 'adopted' by ENSEMBL, a joint project of the Sanger Centre and the European Bioinformatics Institute (EBI). The mosquito database at ENSEMBL, thus, contains all sequences, annotations and additional tools that can be used by the end-user to access these data (Mongin et al. 2004). An additional bonus for the mosquito genome is the fact that, in addition to providing the database, ENSEMBL is also responsible for the automatic annotation and re-annotation of the genome, which happens at regular intervals. Finally, ENSEMBL also handles input of hand annotation by members of the research community, using these data in its own automatic annotation pipeline. Thus, information available at http://www.ensembl.org/Anopheles_gambiae/ is updated frequently.

Having mentioned earlier the fact that additional insect vectors have now entered the genomic era, it should also be stated that the genome and biological databases for these species are now to be combined in a single one that will be called Vectorbase, and which will be initiated soon as a novel project. This database is planned to contain the genome information of at least five arthropod vectors (*An. gambiae*, *Cx. pipiens*, *Ae. aegypti*, *Glossina* spp. and the tick *Ixodes scapularis*), while additional vectors may be added at later stages. In addition to the genome data, the plans call for the inclusion of general biological and genetic data similar to what is already stored in AnoBase (http://www.anobase.org), the *Anopheles* database. Finally, new sections are to be developed that will contain data on population biology and data on post-genomics such as information on cDNAs (EST), images, expression profiles, etc.

Conclusions

This brief report presents the situation in vector genomics and, especially, post-genomics in the summer of 2004. As far as the arthropod disease vectors are concerned, it is obvious that the discipline is now almost monopolized by post-genomic research originating in the publication of the WGS of *An. gambiae*. It is, however, expected that in the near future the genomes of additional vectors will have been sequenced. This wealth of information that can only be managed with up-to-date informatics tools is expected to yield results that may soon be useful for the design of alternative strategies aimed at controlling the diseases that are transmitted through these vectors. Recent advances in molecular-biological techniques, especially the possibility to knock down genes through the RNAi technology (Fjose et al. 2001) have now opened up ways to use surrogate genetics for the manipulation of insects (Wimmer 2003). The efficient and sustained regulation of transcription of effector genes, natural and 'artificial' that will interfere with the transmission of disease agents is theoretically achievable, and most of these advantages are the results of genomic and postgenomic research on the fruit fly and the malaria mosquito. The future, thus, although not foreseeable, should definitely be viewed with an optimistic eye.

References

Adams, M.D., Celniker, S.E., Holt, R.A., et al., 2000. The genome sequence of *Drosophila melanogaster*. *Science,* 287 (5461), 2185-2195.

Alavi, Y., Arai, M., Mendoza, J., et al., 2003. The dynamics of interactions between *Plasmodium* and the mosquito: a study of the infectivity of *Plasmodium berghei* and *Plasmodium gallinaceum*, and their transmission by *Anopheles stephensi*, *Anopheles gambiae* and *Aedes aegypti*. *International Journal for Parasitology*, 33 (9), 933-943. Erratum in: *International Journal for Parasitology*, 2004, 34 (2), 245-247.

Barker, G.C., 2002. Microsatellite DNA: a tool for population genetic analysis. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 96 (Suppl. 1), S21-24.

Bridges, C.B., 1935. Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. *Journal of Heredity,* 26, 60-64.

Brumfield, R.T., Beerli, P., Nickerson, D.A., et al., 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, 18 (5), 249-256.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., et al., 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biology,* 3 (12), research0079.1–0079.14. [http://genomebiology.com/2002/3/12/research/0079]

Craig, A., Kyes, S., Ranson, H., et al., 2003. Malaria parasite and vector genomes: partners in crime. *Trends in Parasitology*, 19 (8), 356-362.

Dujon, B., Alexandraki, D., Andre, B., et al., 1994. Complete DNA sequence of yeast chromosome XI. *Nature,* 369 (6479), 371-378.

Fjose, A., Ellingsen, S., Wargelius, A., et al., 2001. RNA interference: mechanisms and applications. *Biotechnology Annual Review,* 7, 31-57.

Goffeau, A., Barrell, B.G., Bussey, H., et al., 1996. Life with 6000 genes. *Science,* 274 (5287), 546, 563-567.

Hemingway, J. and Craig, A., 2004. Parasitology: new ways to control malaria. *Science*, 303 (5666), 1984-1985.

Holt, R.A., Subramanian, G.M., Halpern, A., et al., 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298 (5591), 129-130,141-149.

Jacobs-Lorena, M., 2003. Interrupting malaria transmission by genetic manipulation of anopheline mosquitoes. *Journal of Vector Borne Diseases*, 40 (3/4), 73-77.

Lehmann, T., Licht, M., Elissa, N., et al., 2003. Population structure of *Anopheles gambiae* in Africa. *Journal of Heredity*, 94 (2), 133-147.

Levashina, E.A., 2004. Immune responses in *Anopheles gambiae*. *Insect Biochemistry and Molecular Biology*, 34 (7), 673-678.

Marth, G.T., Korf, I., Yandell, M.D., et al., 1999. A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, 23 (4), 452-456.

Metzel, C., 1684. De musca vini vel cerevisiae ascentis. *Miscellanea Curiosa sive ephemeridum medico-physicarum Academiae Caesareo-Leopoldinae naturae curiosum*, 2, 96-98.

Mongin, E., Louis, C., Holt, R.A., et al., 2004. The *Anopheles gambiae* genome: an update. *Trends in Parasitology*, 20 (2), 49-52.

Siden-Kiamos, I. and Louis, C., 2004. Interactions between malaria parasites and their mosquito hosts in the midgut. *Insect Biochemistry and Molecular Biology*, 34 (7), 679-685.

The FlyBase Consortium, 2003. The FlyBase database of the *Drosophila* Genome projects and community literature. *Nucleic Acids Research*, 31 (1), 172–175.

Tripet, F., Touré, Y.T., Dolo, G., et al., 2003. Frequency of multiple inseminations in field-collected *Anopheles gambiae* females revealed by DNA analysis of transferred sperm. *American Journal of Tropical Medicine and Hygiene*, 68 (1), 1-5.

Wiesner, J., Borrmann, S. and Jomaa, H., 2003. Fosmidomycin for the treatment of malaria. *Parasitology Research*, 90 (Suppl. 2), S71-S76.

Wimmer, E.A., 2003. Innovations: applications of insect transgenesis. *Nature Reviews Genetics*, 4 (3), 225-232.

Zheng, L., Benedict, M.Q., Cornel, A.J., et al., 1996. An integrated genetic map of the African human malaria vector mosquito, *Anopheles gambiae*. *Genetics*, 143 (2), 941-952.

Zheng, L., Collins, F.H., Kumar, V., et al., 1993. A detailed genetic map for the X chromosome of the malaria vector, *Anopheles gambiae*. *Science*, 261 (5121), 605-608.