台灣昆蟲 31: 149-156 (2011)
Formosan Entomol. 31: 149-156 (2011)

Behind the Scenes of an Ant Genome Project

Yannick Wurm*

Department of Ecology & Evolution, University of Lausanne, 1015 Lausanne, Switzerland
Vital-IT, Swiss Institute for Bioinformatics, University of Lausanne, 1015 Lausanne, Switzerland

ABSTRACT

Dramatic improvements in DNA sequencing technologies have led to a more than 1,000-fold reduction in sequencing costs over the past five years. Genome-wide research approaches can thus now be applied beyond medically relevant questions to examine the molecular-genetic basis of behavior, development and unique life histories in almost any organism. A first step for an emerging model organism is usually establishing a reference genome sequence. I offer insight gained from the fire ant genome project. First, I detail how the project came to be and how sequencing, assembly and annotation strategies were chosen. Subsequently, I describe some of the issues linked to working with data from recently sequenced genomes. Finally, I discuss an approach undertaken in a follow-up project based on the fire ant genome sequence.

Key words: genome assembly, how-to, genome annotation, fire ant, *Solenopsis invicta*

Introduction

Dramatic improvements in DNA sequencing technologies have led to an over 1,000-fold reduction in sequencing costs per basepair over the past five years (Stein, 2010). For example, the human genome published in 2001 cost an estimated 3 billion USD (Collins *et al*., 2003) and sequencing the honey bee genome completed in 2006 cost 8 million USD (Huang, 2007). More recently, several ant genomes were sequenced each for only tens to several hundred thousand USD (Chris R. Smith, private communication; Smith *et al*., 2011; Wurm *et al*., 2011) and prices continue to fall.

As this trend continues, there is increasing interest to obtain sequence information from "emerging model organisms" for which research communities are small in comparison with those that exist for established laboratory species such as *Drosophila melanogaster* or *Caenorhabditis elegans*. Many emerging plant and animal models have unique life histories and

*Corresponding email: yannick.wurm@unil.ch

studying them at molecular-genomic levels provides new opportunities to answer original ecological and evolutionary questions (Straalen & Roelofs, 2006; Hudson, 2008; Tautz *et al*., 2010).

The first steps in developing genomic tools for an emerging model organism today are sequencing its genome and cataloging all protein-coding genes. I was recently involved in such a project for the fire ant *Solenopsis invicta* (Wurm *et al*., 2011). I share the steps and approaches taken for the fire ant genome project. First, I detail how the project came to be and how sequencing, assembly and annotation strategies were chosen. Subsequently, I describe some of the issues linked to working with data from recently sequenced genomes. Finally, I describe a sequencing-based follow-up project that uses the fire ant genome sequence as a reference.

Results

Initiation & Consortium

The idea of sequencing the fire ant genome first materialized in 2005 with a white paper sequencing proposal (Ross *et al*., 2005) submitted to the United States Department of Energy Joint Genome Institute. But the proposal remained unfunded. The project resurfaced as sequencing prices plummeted and made it feasible to envision a genome project with "standard" laboratory funding. In summer 2008, University of Lausanne's Genomics Technologies Facility acquired an Illumina sequencer and ran some sequencing tests… including with fire ant DNA. Preliminary assemblies at the Swiss Institute of Bioinformatics' Vital-IT Center for High Performance Computing identified the full *S. invicta* mitochondrial sequence, suggesting that *de novo* genome assembly was feasible. Several months later, we decided to go ahead with the genome project in collaboration with DeWayne Shoemaker at USDA in Gainesville, Florida, John Wang, Laurent Keller and myself at University of Lausanne, Ioannis Xenarios at Vital-IT and Laurent Farinelli from sequencing provider Fasteris SA.

At the January 2009 Past-Present-Future of Ant Genomics meeting at Arizona State University (Smith *et al*., 2009), we announced the creation of the fire ant genome consortium in which each participating laboratory would contribute at least 20,000 USD for sequencing costs. The participation of more than ten laboratories provided sufficient funds to experiment with sequencing approaches.

Sequencing, Assembly and Automated Gene Prediction Strategies

Sequencing. A sequencing project usually focuses on a single inbred strain or line because variations within a normal diploid genome are difficult to resolve during assembly. For example, assembly software may be unable to correctly merge sequence reads from two alleles of a single gene, and may be similarly confused when the two haplotypes differ by an insertion or deletion (Kelley & Salzberg, 2010). However, no strains or lines exist for fire ants. We were nevertheless able to circumvent this problem because the sex of an ant is determined by ploidy: Unfertilized haploid eggs become males (Hölldobler & Wilson, 1990). To minimize ambiguities, the bulk of our sequencing effort thus focused on a single haploid male. His DNA was sequenced using two methods: we obtained 20x genome coverage in single-ended reads from a *Roche 454* sequencer (average length: 314 basepairs (bp)), and 44x genome coverage from an *Illumina* sequencer (paired reads separated by 352 bp; each read 32 to 100 bp long). Clearly, these kinds of sequences cannot be used to resolve long repetitive regions. Because eukaryotes in general and *S. invicta* in particular harbors many repetitive regions (Li & Heinz, 2000), we additionally obtained 6x genome coverage in pairs of *Roche 454* sequences separated by 8,000 and 20,000 bp. Doing this required very large amounts of DNA

that we obtained from pools of brothers of the focal male. Together, these pools contained the diploid genome of their mother. *Roche 454* sequencing was performed at University of Florida's Interdisciplinary Center for Biotechnology Research, while Fasteris SA in Geneva generated *Illumina* sequence.

Assembly. When we obtained sequence data (mid to late 2009), there were still no reports of assemblies or assembly approaches for eukaryotic genomes. Our initial attempts using assembly tools including *Euler* (Pevzner *et al*., 2001), *Velvet* (Zerbino & Birney, 2008), *ABYSS* (Simpson *et al*., 2009), *cap3* (Huang & Madan, 1999) and *Roche 454 newbler* were unsuccessful: some required memory or processor resources impossible to meet, and others seemed like they might work but assembled only very few sequences. This shed doubt on the feasibility of the project. Subsequently hiring bioinformatics engineer Oksana Riba-Grognuz facilitated the systematic testing of many strategies for data filtering and assembly. Additionally, the performance of assembly algorithms improved dramatically during our few months of tests. Assemblies were compared using graphs that simultaneously indicated numbers of contigs or scaffolds, their N50 and maximal sizes, and the total amount of sequence in the resulting assemblies. In addition, assembly completeness and accuracy was estimated based on the presence of a repertoire of protein-coding genes that are conserved throughout eukaryotes (Parra *et al*., 2009) as well as using a specifically developed visual method based on alignment of assembled mRNA reads to the genome (Riba-Grognuz *et al*., in prep.). The retained approach creatively combined the different data types and two assembly tools:

1. Stringent quality filtering and trimming of *Illumina* sequence.
2. *SOAPdenovo* (Li, Zhu *et al*., 2010) which was used for the Panda genome (Li *et al.*, 2010) to iteratively assemble the *Illumina* sequence.
3. Cutting the resulting assembled *Illumina* data up into partially overlapping "imitation single-ended *Roche 454* sequences" 300 bp long.
4. Assembling the real and imitation single-ended *Roche 454* sequences with *Roche 454 newbler* software using extremely stringent parameters that helped resolve many repeats as well as the "-large" parameter required for assemblies of large genomes to converge.
5. Adding the pairs of sequences separated by 8,000 and 20,000 bp in turn with *Roche 454 newbler* using even more stringent parameters to bioinformatically select only the sequences matching the haplotypes of the focal haploid male.

Gene prediction. Automated gene prediction algorithms are based on similarity between the genome sequence and known protein or mRNA sequences, or discover genes *ab initio,* using models of the theoretical structure of genes, or both. Performance of gene identification software varies among algorithms and among gene and genomes. Genome projects thus often rely on predictions by several different programs that are may subsequently be merged (Elsik *et al*., 2007). To identify protein-coding genes in the assembled fire ant genome, we opted for the MAKER (Cantarel *et al*., 2008) software which is becoming the "BLAST of genome annotation" because it is easy to set up and transparently takes care of optimally running multiple gene prediction algorithms and creating consensus gene sets. Some *S. invicta* gene models were subsequently further refined using Genewise (Birney *et al*., 2004). Overall, gene identification greatly benefited from the use of protein sequence from other insects and of independently obtained *S. invicta* mRNA sequence.

Identifying unique features of a genome

It is challenging to identify what may be interesting or unique in the large fasta

files resulting from genome assembly and gene prediction. A first approach relies on candidate genes or pathways and comparing their presence/absence or evolution with what is known from related insects. For example, this approach helped determine that the *transformer*/*feminizer* sex-determination gene was independently duplicated in an ancestor of *S. invicta* and in an ancestor of the honey bee *Apis mellifera* (Wurm *et al*., 2011).

The second approach relies on unbiased automated comparisons. Within a species, mRNA sequence can be used to identify genes with extreme numbers of splice-forms. Between genomes of related species, comparison of sizes of gene families or protein domain families can identify differences between species. The latter approach using PFAM (Finn *et al*., 2010) and Prosite (Sigrist *et al*., 2010) domain analyses highlighted several unique aspects of the fire ant genome that are likely linked to the complex social behavior of this species. For example, its more than 400 putative olfactory receptors constitute the largest repertoire found in insects, and the genome also harbors an expansion of lipid-processing genes that may be used in the production of pheromones. These gene family expansions likely reflect the importance of chemical communication in ants (Wurm *et al*., 2011).

Discussion

Issues with Young Genome Projects

While recent technological progress has made genome sequencing, assembly and annotation widely accessible, caution must be taken when working with the obtained data. First, the new sequencing technologies remain young: they provide only short sequence reads with relatively high error rates. Similarly, algorithms and bioinformatics tools to deal with the new data types are still improving. Furthermore, *de novo* assemblies of eukaryotic genomes consist in many pieces of contiguous sequence (contigs) arranged in thousands of scaffolds instead of small numbers of chromosomes (Salzberg & Yorke, 2005). Labor-intensive mapping and finishing steps to reduce fragmentation were normally part of Sanger-based genome projects but are now generally foregone. The exons of a gene may thus be split across multiple contigs or scaffolds. Finally, while gene prediction tools have matured, adjacent genes may be incorrectly joined into a single gene model and single genes may be incorrectly truncated or split into multiple gene models. Studying the molecular evolution of specific gene families thus requires case-by-case inspection and often editing of the relevant gene models (Lumi Viljakainen, personal communication).

A follow-up study based on high-resolution genotyping

Explaining how interactions between genes and the environment influence social behavior is a fundamental question, yet there is limited relevant information for species exhibiting natural variation in social organization. *S. invicta* is characterrized by a remarkable social polymerphism: The presence of one or multiple reproductive queens within a colony as well as other phenotypic and behavioral differences are completely associated with allelic variation at a single Mendelian factor marked by the gene *Gp-9* (Ross & Keller, 1998). Furthermore, the *b* allele at *Gp-9* is a rare example of a "green beard gene" because workers carrying the *b* allele favor the reproduction of queens carrying *b* by executing queens that do not carry *b*. This selfish allele has not reached fixation because of balancing selection: The phenotypes associated with *b* are counter-selected in certain environments, and *bb* homozygotes are lethal (Keller & Ross, 1998; Ross & Keller, 1998).

When it was identified, *Gp-9* was one of only fifteen examined allozyme markers, leaving open the question of whether other genes may be tightly linked to and thus

co-segregating with *Gp-9*. John Wang, DeWayne Shoemaker, Laurent Keller and I are now following up on this question using the fire ant genome sequence and RADseq genotyping (Baird *et al*., 2008; Davey & Blaxter, 2011). The RADseq approach consists in extracting the same 0.01% of the genome of many individuals and sequencing the extracted sequence from all individuals within a single *Illumina* sequencing reaction while keeping track of which sequence comes from which individual. Individual genotypes at several thousand polymorphic sites should thus be found (Davey & Blaxter, 2011). By analyzing the resulting data, we should be able to determine whether any genes are completely linked to *Gp-9*. Such genes could be responsible for the phenotypes associated with *Gp-9*.

Conclusion

Recent improvements in sequencing technology have drastically increased accessibility of molecular and genomic investigations. However, the technologies remain young and continue to improve: Between July 2009 and September 2010, the length of useable *Illumina* sequence reads more than doubled, and the quality and the quantity of sequence obtained from a single lane increased more than 10-fold (my unpublished data). Similarly, software for assembling, mapping and analyzing the new data types is maturing and performs far better than before. Given the quick improvements of both sequencing and analysis tools, one continues to wonder whether new sequencing projects should be initiated immediately, or whether a few months delay may lead to better data and easier analysis to the extent of obtaining results more rapidly.

Acknowledgments

I thank John Wang and Chin Cheng Yang for comments on this manuscript; Oksana Riba-Grognuz, Laurent Keller, DeWayne Shoemaker, Laurent Farinelli, John Wang, Ioannis Xenarios, Krista Ingram and the many members of the fire ant genome consortium for their help and support. I am funded by a Swiss National Science foundation grant and an ERC grant to Laurent Keller.

References

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376.

Birney E, Clamp M, Durbin R. 2004. Gene wise and genomewise. Genome Res 14: 988-995.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. 2008. Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18: 188-196.

Collins FS, Morgan M, Patrinos A. 2003. The human genome project: lessons from large-scale biology. Science 300: 286-290.

Davey JW, Blaxter ML. 2011. RADSeq: next-generation population genetics. Brief Funct Genom 2: 416-423.

Elsik C, Mackey A, Reese J, Milshina N, Roos D, Weinstock G. 2007. Creating a honey bee consensus gene set. Genome Biol 8: R13.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A. 2010. The Pfam protein families database. Nucleic Acids Res 38: D211-D222.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res 9: 868-877.

Huang Z. 2007. The honey bee genome: the untold stories. Chinese Bull Entomol 44:

5-9.

Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8: 3-17.

Hölldobler B, Wilson EO. 1990. The ants. The Belknap Press of Harvard University Press.

Keller L, Ross KG. 1998. Selfish genes: a green beard in the red fire ant. Nature 394: 573-575.

Kelley DR, Salzberg SL. 2010. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol 11: R28.

Li J, Heinz KM. 2000. Genome complexity and organization in the red imported fire ant *Solenopsis invicta* Buren. Genet Res 75: 129-135.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam T, Yiu S, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. 2010. The sequence and *de novo* assembly of the giant panda genome. Nature 463: 311-317.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res 20: 265-272.

Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res 37: 289-297.

Pevzner PA, Tang H, Waterman MS. 2001. An eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA 98: 9748-9753.

Riba-Grognuz O, Keller L, Falquet L, Xenarios I, Wurm Y. in prep. Visualization and quality assessment of *de novo* genome assemblies of emerging model organisms.

Ross KG, Keller L. 1998. Genetic control of social organization in an ant. Proc Natl Acad Sci USA 95: 14232-14237.

Ross KG, Robinson GE, Abouheif E, Crozier RH, Gadau J, Johnston JS, Keller L, Shoemaker DD, Suarez AV, Meer RKV, Vargo EL, Vinson SB, Wheeler DE. 2005. Proposal for the sequencing of a new target genome. White Paper for a Fire Ant Genome Project.

Salzberg SL, Yorke JA. 2005. Beware of mis-assembled genomes. Bioinformatics 21: 4320-4321.

Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161-D166.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117-1123.

Smith CD, Smith CR, Mueller U, Gadau J. 2009. Ant genomics: strength and diversity in numbers. Mol Ecol 19: 31-35.

Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R,

Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Favé M, Fernandes V, Gibson JD, Graur D, Gronenberg W, Grubbs KJ, Hagen DE, Viniegra ASI, Johnson BR, Johnson RM, Khila A, Kim JW, Mathis KA, Munoz-Torres MC, Murphy MC, Mustard JA, Nakamura R, Niehuis O, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Suen G, Tao S, Torres CW, Tsutsui ND, Viljakainen L, Wolschin F, Gadau J. 2011. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc Natl Acad Sci USA 108: 5667-5672.

Stein LD. 2010. The case for cloud computing in genome informatics. Genome Biol 11: 207.

van Straalen NM, Roelofs D. 2006. An introduction to ecological genomics. Oxford: Oxford University Press.

Tautz D, Ellegren H, Weigel D. 2010. Next generation molecular ecology. Mol Ecol 19 Suppl 1: 1-3.

Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, Dijkstra MB, Oettler J, Comtesse F, Shih C, Wu W, Yang C, Thomas J, Beaudoing E, Pradervand S, Flegel V, Cook ED, Fabbretti R, Stockinger H, Long L, Farmerie WG, Oakey J, Boomsma JJ, Pamilo P, Yi SV, Heinze J, Goodisman MAD, Farinelli L, Harshman K, Hulo N, Cerutti L, Xenarios I, Shoemaker D, Keller L. 2011. The genome of the fire ant *Solenopsis invicta*. Proc Natl Acad Sci USA 108: 5679-5684.

Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

Received: April 28, 2011
Accepted: April 30, 2011

火蟻基因組計畫的幕後側寫

Yannick Wurm*

瑞士洛桑大學　生態與演化系

摘　　要

近五年 DNA 定序技術的進展已使得定序成本比過去減少 1,000 倍以上。因此全基因體研究的方法不僅應用於解決醫學相關問題，也能進一步探討大部分生物的基礎生物學，包括以分子遺傳為主之行為模式、發育及獨特之活史。以新興模式生物為例，首要步驟即為建立參考基因組序列。本文將提供以火蟻基因體定序計畫所得之啟發。首先，詳述此計畫的緣由，並介紹如何選擇基因體解序、組裝和基因體註解的策略。接著說明如何將火蟻基因體資料與近來被解序之物種作結合。最後，討論以火蟻基因體序列為基礎，進行後續研究計畫的方法。

關鍵詞：基因體組裝、指引、基因體註解、火蟻、入侵紅火蟻。

*論文聯繫人
Corresponding email: yannick.wurm@unil.ch