OPINION ARTICLE

# The life cycle of a genome project: perspectives and guidelines inspired by insect genome projects [version 1; referees: 2 approved, 1 approved with reservations]

Alexie Papanicolaou

Hawkesbury Institute for the Environment, University of Western Sydney, Richmond, NSW 2753, Australia

**v1** **First published:** 05 Jan 2016, **5**:18 (doi: 10.12688/f1000research.7559.1)
**Latest published:** 05 Jan 2016, **5**:18 (doi: 10.12688/f1000research.7559.1)

## Abstract

Many research programs on non-model species biology have been empowered by genomics. In turn, genomics is underpinned by a reference sequence and ancillary information created by so-called “genome projects”. The most reliable genome projects are the ones created as part of an active research program and designed to address specific questions but their life extends past publication. In this opinion paper I outline four key insights that have facilitated maintaining genomic communities: the key role of computational capability, the iterative process of building genomic resources, the value of community participation and the importance of manual curation. Taken together, these ideas can and do ensure the longevity of genome projects and the growing non-model species community can use them to focus a discussion with regards to its future genomic infrastructure.

**Open Peer Review**

**Referee Status:** ✓ ? ✓

| | Invited Referees 1 | 2 | 3 |
|---|---|---|---|
| **version 1** published 05 Jan 2016 | ✓ report | ? report | ✓ report |

1 **Stephen Richards**, Baylor College of Medicine USA

2 **John W. Davey**, University of Cambridge UK

3 **Ian Holmes**, University of California Berkeley USA

**Discuss this article**

Comments (0)

**Corresponding author:** Alexie Papanicolaou (alpapan@gmail.com)

**How to cite this article:** Papanicolaou A. **The life cycle of a genome project: perspectives and guidelines inspired by insect genome projects [version 1; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2016, **5**:18 (doi: 10.12688/f1000research.7559.1)



**Grant information:** The author is supported by the Hawkesbury Institute for the Environment (Western Sydney University); no grants were involved in supporting this work.

**Competing interests:** No competing interests were disclosed.

**First published:** 05 Jan 2016, **5**:18 (doi: 10.12688/f1000research.7559.1)

## Introduction

In this perspectives and opinion article, created from the viewpoint and experience of an insect genome informatician, I seek to explain how the generation, maintenance and publication of genome projects has now reached a stage that requires the community to pause for thought. Each genomics community is at a crossroads while it decides on what is the best strategy for generating and using high quality genome projects. There are undoubtedly leaders who have been pursuing specific strategies but the community is not necessarily on the same path. For over a decade, dozens of genome projects have been completed and this number is increasing exponentially. Due to my experience around insect genomics and participation in the i5k activities (an international consortium providing leadership and resources for insect genome projects[1]), I shall present my views centred around the insect community but hopefully these views are broadly applicable. Previously, a co-author and I explained how the new, cheaper technologies changed the landscape by which genome projects are conceived and organised in the hope we can guide a burgeoning insect genomics community[2]. We also argued that individual researchers can produce a genome reference sequence for their favourite species and the large consortia are no longer needed. Since then, we have seen first-hand how an incredible worldwide effort has managed to shed light on why and how we can initiate hundreds (if not thousands) of insect genome projects[1]. The i5k's aim is to educate and support individual scientists as they seek to acquire genomics skills. It has also produced the first tranche of key insect genomes, mainly picked due to their place in the phylogenetic tree, an achievement which may be currently underutilised but whose importance cannot be understated. After involvement with multiple genome projects, a diverse array of transcriptomic projects and at least one widely used genomics software, I have come to conclude that in our rush to initiate the genomes projects of a larger part the tree of life we have neglected some important issues. First, the only reason our throughput is so high is because all published genome papers present merely just a draft. It is a useful draft but still only an initial effort. Second, this draft does indeed contain many of the instructions of how to generate an organism, but a genome sequence alone does not decipher it. It merely transcribes so we can conduct experiments with it. Deciphering will require both good experimental design and the capability to integrate such experiments. Third, the research community is not just the end-user but also part of the project team; we have, on the whole, neglected to bring them up to speed. These issues may seem intuitive but much of the community's leadership is not conscious of it. Such issues cannot be readily resolved without substantial education. In this opinion article, I present some of the insights I gained from my own journey that I believe can help inspire the community to follow us as we hack our way through the multi-species genome sequencing trail.

## Insight 1: A brave new, informatics-led, capability

During the last few years, genomics has moved from being a capability led by – and limited by – wet lab techniques to one led by computational science. One of the first eukaryotic genomes ever assembled was *Drosophila melanogaster*[3], chosen partly because its genome architecture is much simpler than that of the human one. At the time, there was much discussion on whether sheer computational power would have been able to solve such a puzzle. The history of how this affected the human genome project is well known. There were two camps with many inspiring individuals in each. One side was the publicly-funded effort to sequence the human genome who believed that 'clone-at-a-time', overlap-based sequencing was the answer. One the major forces of innovation from that group was Eric Lander. Originally a mathematician, his team at the MIT was responsible for creating a completely automated sequencing pipeline[4]. As the pressure to publish was looming, the team realised that they did not have the computational software to generate the final genome assembly (until the skill and leadership of Jim Kent saved the day). The privately-funded side (TIGR and Celera), was led by a computational approach from the onset. Eugene Myers - having already worked on efficient computational approaches for biology (e.g. BLAST, suffix arrays and assembly) - proved that new computer science algorithms had the power to solve this problem much more efficiently[5]. Their computational capability drove the development for a new wet-lab technique called Whole Genome Shotgun (WGS) sequencing. The first camp, in other words, built computational approaches to match the wet lab technique. The second one built a wet-lab capability to match a novel computational approach. The majority of the molecular biologists at the time did not support this WGS approach as it was rightly considered to be of inferior quality. At the end of the day, however, it was the cross-talk of the two capabilities that not only produced a comprehensive human genome but also allowed many other species to have their genome sequencing completed. Indeed, the human genome project completely changed our perception of how biological research can scale in a world that transcends borders and how it should maintain its science-led focus (see 6). Even though we still follow most of these tenets, as shown from the vast majority of genome project publications, genome sequencing is an enterprise inherently focused on building resources. That does not mean that it is no longer science-led, it is just the science is not always the fields of biochemistry, ecology or genetics.

For the human genome project, computer science approaches were used to not only convert the wet-lab data to something the research community could make use of but to also guide the wet-lab techniques. By now, an entire coterie of associated industries such as information technologies (IT), informatics (the use of IT to handle large datasets) and data science is underpinning much of genomics. These sciences, often lumped into the rather over-used but useful umbrella definition of 'bioinformatics', have not only allowed us generate genomes but have now become an integral part of any downstream experiment using a reference sequence. Generally, the wider community is not aware of the true potential of informatics as it could be. The benefits go beyond being able to analyse a dataset: The epistemological understanding that comes with studying statistics and informatics can provide the skills for integrating the multi-disciplinary and ever-increasing amounts of data and the framework to make sense of a more synthesized knowledge. If we – as educators - allowed for aspects of programming and data science to become an integral part of the undergraduate curriculum - rather than the lip service that is currently common in most institutions – then we not only equip the next generation with a set of skills but we inspire a uniquely effective way of dissecting complex problems. Further, the high-throughput analysis of data is now a core requirement for any genomic experiment, yet often the analysis is

delegated to computer programs (bioinformatic software), which are, effectively, "black boxes". Such software are great for enhancing productivity but they ought not to be used before we understand all assumptions made on our behalf and explore their parameter space for each particular dataset[7]. Perhaps a way to resolve this is for the software engineering community to invest beyond core algorithms and produce high quality protocol papers that seek to explain what and how a software works while simultaneously providing a user-friendly interface that focuses on productivity (see Haas *et al.* 2013 for an example from a popular RNA-Seq assembly software[8]). A final point is that genomics – being data-rich - is ideal for exploratory research (i.e. generating new hypotheses). The varying quality of genomes and the inherent noise present in biology can actually be accommodated by approaches residing within the information science field (colloquially known as "big data science"). The information science field has been widely used in other disciplines and there are tangible and immediate benefits in experiments such as those using expression data (c.f. see a perspectives article by Hudson *et al.* 2012[9]). The only caveat is that, like all experiments, the quality of any such outcomes will depend on the quality of our resources. In order to avoid surprises, before we proceed with such experiments, we ought to first understand the process of generating such resources.

### Insight 2: A "life cycle" and a grand experiment

From a pedagogical perspective, one can compare genome projects to the life cycle of an insect (Figure 1). Like a developing insect, genome projects go through several stages of development: project design (often of an underestimated importance); DNA and RNA library preparation and sequencing (with rapidly evolving protocols); genome and transcriptome assembly (initially more than one before a consensus one is decided), structural annotation (e.g. "where are the genes and other features?"); functional annotation (e.g. "what does this gene do?"); manual curation of these two annotation types (often the most time-consuming step); and data dissemination (i.e. the steps that are visible to the public and perhaps the most important stage). Viewing this process as a life cycle provides not only the basis of an improved educational narrative but also some immediate insights. For example, genome project can go through multiple iterations of this "life cycle". Further, like insects, the fitness of each stage depends on the quality of all of the previous stages. Also notable is the fact that one cannot proceed unless a stage is completed and "frozen" (in sequencing centre jargon, i.e. no longer manipulated). For example, the annotation process cannot begin unless the assembly is completed. Again, this insight may seem intuitive but having it at the forefront of our thoughts while undertaking a genome project will afford us with some important advantages.

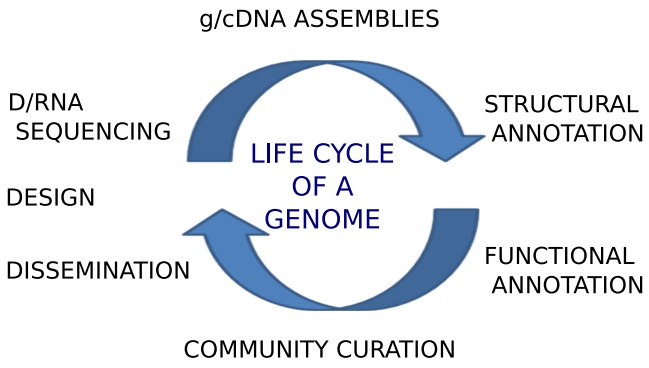


**Figure 1. The iterative process of generating a genome sequence can be seen as a life cycle.**

Before elaborating on that, it is important to first point out that creating a draft genome sequence is a scientific experiment. There is at least one question (often a biological one, the nature of which depends on the discipline of the research leaders), but at the very least involves investigating an organism's genetic blueprint. From a computer science point of view the question is straightforward: what is the correct genome sequence for this species and what are the parts that are important for its function. Further, there are a variety of possible methods and approaches that can be used, there is a risk of failure and at the end there will hopefully be more questions than answers. Therefore, like all scientific experiments, a good project design is essential. This ought to be led by someone - or a team - possessing in-depth knowledge of every step of the process. A review by Richards and Murali 2015[10] outlines many of the common issues a team has to consider when working with insect genome projects. For example, DNA availability and quality, genome size and polymorphism are some of the most important aspects that have led to the poor quality of a number of genome projects. As we complete more genome projects, further capturing and sharing that knowledge is something that the community sorely needs.

By perceiving genome projects as an experiment with a life cycle, one can begin identifying a number of useful insights. For the sake of brevity I will expand on only a couple of the most important ones that can help steer genome projects to be more likely to succeed.

First, most genome projects want to address a particular question which varies between disciplines. It may be to perform a quantitative genetic study, fully ascertain a gene family which hosts a number of recently duplicated and near-identical members, or to build a more accurate phylogenetic framework and identify genes that are key innovations. Each one of these aims requires a genome of a different quality and therefore the project design ought to focus on those outcomes. For example, quantitative genetic studies depend on long scaffolds so markers can be associated with causal genotypes. Gene family ascertainment needs not only a high base-level accuracy but also characterisation of any gene family member turnover (i.e. Copy Number Variants) that may exist within a species[11]. Phylogenomic studies, on the other hand, require neither of these two characteristics: rather multiple species have to be analysed and putative key genes need to be painstakingly curated and characterized. Striving for perfection by achieving all these characteristics could be attempted in the first iteration but rarely do genome project teams have the diversity of skill, time and money to achieve it in a timely fashion.

Second, striving for perfection often manifests as a lack of discipline in keeping with the original experimental design when faced with access to new technological advances. This is even more critical when one considers that the speed of innovation in this field is extraordinary and that new technologies are less well-tested both in terms of wet-lab techniques and the dry-lab algorithms meant to analyse them. For example, even though new approaches such as long read sequencing[12,13], linkage maps[14,15] or chromatin interaction data[16] are an under-used approach and can be of great value to improving genome sequence contiguity (i.e. scaffolding), they are high-risk, time consuming and expensive. Unless the original design included them, in practice they will end up delaying genome projects by months if not years. If one accepts the iterative nature of genomes, one can strive for timely incremental improvements.

Third, we ought to remember that the life cycle moves forward only when we are satisfied with the quality and therefore when we are not satisfied we have to backtrack. Most genome sequencing groups learn early on to appreciate the need to complete each stage to a satisfactory level before proceeding to the next stage. Quality assessment using pre-defined metrics is standard practice in data science. More experienced workers also learn that once a stage is satisfactorily completed ("frozen") and the next one started, one must under no circumstances go back. For example, once the structural annotation is completed, any perturbation of the assembly will invalidate the genome sequence and co-ordinates that gene models depend on. Certainly, if the stage is not satisfactory then it is expected that we go back one step (or even back to stage 1, project design) and start over. For example during the *Helicoverpa* genome project, we found that when there are high levels of polymorphism, the error levels of the 454 technology were prohibiting in completing an assembly at an N50 higher than 30 kb; short Illumina reads (ca. 100 bp at the time) were far more suitable and could be coupled with 454 mate pair libraries to produce a genome assembly of an N50 exceeding 1,000 kb. Had we decided to continue with the life cycle then we would have invested enormous effort in annotating a fragmented genome that was not suitable for our aims.

Fourth, the life cycle enforces the notion that quality of any one step is dependent on the outcomes of the previous step. For example, using the best possible starting material and extensive quality control of the sequencing data can add far more value to an assembly than any wet-lab or even dry-lab investment. This may also appear self-evident but, as mentioned by Richard and Murali, many non-model insect projects have limitations due to biology or availability of samples which no assembly algorithm can account for. For example within the *Heliconius* community, a multi-species genome project undertaken by the Discovar team, showed that even this new approach provides results inferior to the well tested Allpaths-LG approach (created by the same team), even in the hands of the authors. This is postulated to be due to these butterflies exhibiting high levels of complex polymorphism and repeat structures (Owen McMillan pers. comm. September 2015). Instead, a current protocol using a more traditional Allpaths-LG[17] strategy coupled with a dense linkage map has allowed for a far superior assembly. At the end of the day, a combined understanding of the computational approaches used in genomics and the genomic architecture of a species is required to generate an excellent assembly; this will likely require more iterations of the life cycle.

## Insight 3: Sharing is caring

Significant resources and team effort are required for updating a genome version. In my experience, the greater of the challenges is how to co-ordinate the release of a new genome so that the community has access to the latest science and the genome team is rewarded for their contribution in way appreciated by their funding bodies and employers. There are not many insect genomes that have completed the life cycle multiple times but a flagship example is that from the silkworm, *Bombyx mori*, which had three publications. The first paper provided an early draft of the genome that turned out to be of limited broad utility but was published in Science[18]. The second was published shortly afterwards by a second, competing, group[19]. Even though it was of higher overall accuracy, that paper appeared in the journal of DNA Research (Oxford University Press). The third publication was the result of intense political activity, leadership and labour from both teams. It delivered a genome resource that was of higher quality than any other non-model species published at the time (and for the time being, it still is). However, this last iteration was published in the domain-specific journal of Insect Biochemistry and Molecular Biology[20]. Further, it is unlikely that any funding body would support any such activity today: once a genome is "published" it is deemed complete. Even though some of the raw data was made available through GenBank, it is important to note that each silkworm paper came with its own data repository and database. Indeed, there is still no "one stop-shop" for silkworm genomics and there is no support for the community to provide feedback for particular scaffolds or genes (i.e. "curation"). By all accounts, even though silkworm genomics is very much alive, it seems that the relevant informatic community is no longer active and I fear that most insect genome projects are – by default rather than choice - following this protocol.

So we are faced with two issues: how to provide informatic support beyond that initial publication and how to create a sustainable publishing model that allows for a genome project life cycle. In this particular case, I believe these two issues have solutions that can address both. One option, currently undertaken by the community, is for the papers exhibiting a new genome version to be submitted as a technical advance to a low impact factor but useful journal. This may, however, prevent engaging the best bioinformaticians and also undervalues the contribution of bioinformatics. Another option is for genome project teams to address a different, novel and important research question. That way the value of an improved genome resource can be properly showcased. This is time-consuming, however, and will therefore result in significant delays for making the data accessible to the community. A third option is to decouple publications from resources while at the same time respecting the value of genomics. We can achieve that if we shifted our focus from the "impact factor route" (21; c.f. the San Francisco Declaration on Research Assessment from the American Society for Cell Biology) and focus on "real world impact". This is what an increasing number of government and research institutions (such as universities) are being asked to focus on. In genomics,

we can make use of an updated version of the "Fort Lauderdale Agreement". First, the data is made available before publication (in a controlled fashion) and the community is offered the opportunity to edit and improve it. Importantly, the community should be able to use it for downstream experiments and - if a journal editor agrees and their work does not fall under genomics but say biochemistry or molecular ecology – publish their findings (e.g. see 22 for an example from the *Helicoverpa* Genome Project). The assembly and annotation are benchmarked by this process and the community acquires both awareness and training. Eventually the first "genome paper" is submitted to a journal. This showcases not a new technical capability or a competition for being "the first but not the best" but rather a broad body of work from a relative large section of the stakeholders. The subsequent genome versions can be linked to either new experimental work on this one species or multi-species comparative genomic insights. Under the leadership of the Baylor College of Medicine, this is exactly the model that the i5k community has chosen. Even though a considerable number of genome projects have been "completed but unpublished", they are nonetheless available either freely (e.g. the Mediterranean fruit fly is on NCBI) or upon request (for example see https://www.hgsc.bcm.edu/i5k-pilot-project-summary) and there are significant real world impacts as scientists across the world are collaborating using these new resources.

Overall, our community is excellent at producing and disseminating the outcomes of top quality research, however, what the broader community values most is the dissemination and maintenance of data. Except terminal data (e.g. assemblies), primary (e.g. sequencing reads) and annotation data are also extremely valuable for conducting further experiments. A reliable and user-friendly IT platform for dissemination is the most effective way to reduce the bioinformatic bottleneck that is manifesting in many labs. Traditionally, data dissemination occurred in tandem with publication (e.g. via GenBank). Sadly, this is often limited to what occurred to get a particular paper accepted in that one publication and we cannot rely on journal editors to ensure that up-to-date data are available. Informatics has certainly empowered the community by providing it with a number of tools such as those based on "GMOD toolkit" and Content Management Systems[23,24], Ensembl[25], InterMine[26] or even entire infrastructures that can support the knowledge discovery process from beginning to the end[27]. Provision of resources is also not limited by a lack of effort (c.f. the Nucleic Acid Research and Database journals) but issues such as lack of funding, maintenance, exchange of data from other resources or communication with the relevant community. As a consequence, their utility or lifespan can be limited and the invested informatic effort wasted. We need a system that has the interoperability of the UCSC Browser (the version developed for cancer research; 28), the web-services of InterMine (created originally for Drosophila; 26), the data richness of the ENSEMBL project[25] and the ecosystem of iPlant[27]. At the same time it hosts a dedicated team knowledgeable on insect biology and tasked with not only managing the data for the insect community but also building awareness for best practices, providing training and enforcing quality control. Without this resource, every new insect genome that is funded will be of limited value. The major issue for ensuring long-term sustainability is that sequencing centres and science leaders cannot guarantee the long-term provision of the required computational infrastructure. Even though centres such as the NCBI can host raw data and finalised gene models, they cannot provide a community portal with domain specific tools and resources. This is another area where the i5k consortium has shown leadership: in collaboration with the National Agricultural Library (NAL) of the USDA the insect community has now access to a dedicated team which is deploying an increasing number of tools (including the GMOD toolkit) and provides basic computational resources and training[29].

## Insight 4: The human touch

The NAL team goes beyond merely hosting data and developing tools: they provide a platform for the community to assess the quality of genomes and edit the results of the automated bioinformatic processes of annotation. This manual checking and editing, i.e. "curating", is an important check on the automated approaches on the underlying data that any experiment will end up relying on and it has featured in all major genome projects. In the early days of genome projects, the automated annotation 'freeze' was the stage where significant community outreach and involvement was sought. This often took the form of Annotation Jamborees and these were driven – and funded - by the leaders of the consortium. There, community members would meet and edit the computational predictions using the Apollo annotation system[30], discuss research questions and co-ordinate project activities. These events are now mostly associated with the Sanger era where the costs to create a genome sequence were orders of magnitude larger than the costs associated with hosting a meeting. However, these meetings played a critical role in not only improving on the computational predictions but also forming a genomics community and educating researchers on how to use the genome[31]. As genome project costs have been driven down we had to invent new ways of co-ordinating work. One solution has been the International Arthropod Genomics Workshop but that lacks the immediacy and cannot deal with the enormous volume of data and diversity of species in a timely manner. Perhaps not surprising, informatics came to the rescue with a number of 'community curation' tools developed. In the insect world, the clear winner has been the Web Apollo software (also known as WebApollo) a plugin of the JBrowse genome browser[32,33]. Except for offering a real-time, internet-enabled implementation of genome viewing and editing, this informatic capability is underpinning NAL's effort to help in forming, educating and maintaining genomic communities. Our greatest challenge in this space, however, is that we are one step behind: even though we are excellent in collecting and curating genomic data, we will still have to learn how to efficiently collect and curate a vast amount of new types of information such as those derived from epigenetics, population genetics and even ecology.

## Future directions

If there was one final take home message it would be that while the genomics community is currently reaping the benefits of a number of technological advances, it is also about to be faced with a paradigm shift due to not only the number of genome sequences being made available but also the types of data that are becoming cheaper and increasingly common. Certainly, we need more

scientists to learn how these data are derived and how to work with them more effectively but we – the informatics community - also need to educate more of them of how to produce high-quality, long-living projects that meet best practice. In my opinion, the challenge to deliver such an outreach activity is the development and the delivery of a high quality, unified course that will perhaps be tailored for each taxonomic domain or discipline. It is true that genome analysis and bioinformatics is a research discipline that takes years to master but – like statistics – it is also a useful set of tools that empowers everyone who chooses to invest the time to acquire some basic knowledge. Even further, it is such an exciting time to be a biological data scientist that those who decide to view computational biology as a skill to be mastered, while excelling in their chosen biological discipline, may also drive many of the next generation of synthesis in biology.

## Competing interests

No competing interests were disclosed.

## Grant information

The author is supported by the Hawkesbury Institute for the Environment (Western Sydney University); no grants were involved in supporting this work.

## Acknowledgements

I would like to thank Scott Cain, Alex Feltus, Monica Munoz-Torres, Nassib Nassar, Konrad Paszkiewicz, Gil Smith, Rob Waterhouse, Jennifer Wortman, Yannick Wurm and many others for inspiring me to write this review. I am also grateful to Alex Watson-Lazowski and Stephen Richards for comments on an earlier version of this paper.

## References

1. i5K Consortium: **The i5k Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment.** *J Hered.* 2013; **104**(5): 595–600.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Edwards OR, Papanicolaou A: **A Roadmap for Whitefly Genomics Research: Lessons from Previous Insect Genome Projects.** *J Integr Agric.* 2012; **11**(2): 269–280.
Publisher Full Text
3. Myers EW, Sutton GG, Delcher AL, *et al.*: **A whole-genome assembly of *Drosophila*.** *Science.* 2000; **287**(5461): 2196–2204.
PubMed Abstract | Publisher Full Text
4. Watson JD, Berry A: **DNA: The Secret of Life**. Alfred A. Knopf; 2003.
Reference Source
5. Myers EW, Weber JL: **Is Whole Human Genome Sequencing Feasible?** In *Theoretical and Computational Methods in Genome Research*. Springer US; 1997; 73–89.
Publisher Full Text
6. Collins FS, Morgan M, Patrinos A: **The Human Genome Project: lessons from large-scale biology.** *Science.* 2003; **300**(5617): 286–290.
PubMed Abstract | Publisher Full Text
7. Papanicolaou A, Stierli R, Ffrench-Constant RH, *et al.*: **Next generation transcriptomes for next generation genomes using *est2assembly*.** *BMC Bioinformatics.* 2009; **10**: 447.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Haas BJ, Papanicolaou A, Yassour M, *et al.*: ***De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.** *Nat Protoc.* 2013; **8**(8): 1494–1512.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Hudson NJ, Dalrymple BP, Reverter A: **Beyond differential expression: the quest for causal mutations and effector molecules.** *BMC Genomics.* 2012; **13**: 356.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Richards S, Murali SC: **Best Practices in Insect Genome Sequencing: What Works and What Doesn't.** *Curr Opin Insect Sci.* 2015; **7**: 1–7.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Briscoe AD, Macias-Muñoz A, Kozak KM, *et al.*: **Female behaviour drives expression and evolution of gustatory receptors in butterflies.** *PLoS Genet.* 2013; **9**(7): e1003620.
PubMed Abstract | Publisher Full Text | Free Full Text
12. English AC, Richards S, Han Y, *et al.*: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.** *PLoS One.* 2012; **7**(11): e47768.
PubMed Abstract | Publisher Full Text | Free Full Text
13. McCoy RC, Taylor RW, Blauwkamp TA, *et al.*: **Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements.** *PLoS One.* 2014; **9**(9): e106689.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Heckel DG: **Comparative Genetic Linkage Mapping in Insects.** *Annu Rev Entomol.* 1993; **38**: 381–408.
Publisher Full Text
15. Rastas P, Paulin L, Hanski I, *et al.*: **Lep-MAP: fast and accurate linkage map construction for large SNP datasets.** *Bioinformatics.* 2013; **29**(24): 3128–34.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Burton JN, Adey A, Patwardhan RP, *et al.*: **Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions.** *Nat Biotechnol.* 2013; **31**(12): 1119–25.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Gnerre S, MacCallum I, Przybylski D, *et al.*: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proc Natl Acad Sci U S A.* 2011; **108**(4): 1513–1518.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Xia Q, Zhou Z, Lu C, *et al.*: **A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*).** *Science.* 2004; **306**(5703): 1937–1940.
PubMed Abstract | Publisher Full Text
19. Mita K, Kasahara M, Sasaki S, *et al.*: **The genome sequence of silkworm, *Bombyx mori*.** *DNA Res.* 2004; **11**(1): 27–35.
PubMed Abstract | Publisher Full Text
20. International Silkworm Genome Consortium: **The genome of a lepidopteran model insect, the silkworm *Bombyx mori*.** *Insect Biochem Mol Biol.* 2008; **38**(12): 1036–1045.
PubMed Abstract | Publisher Full Text
21. Verma IM: **Impact, not impact factor.** *Proc Natl Acad Sci U S A.* 2015; **112**(26): 7875–6.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Jones CM, Papanicolaou A, Mironidis GK, *et al.*: **Genomewide transcriptional signatures of migratory flight activity in a globally invasive insect pest.** *Mol Ecol.* 2015; **24**(19): 4901–4911.
PubMed Abstract | Publisher Full Text
23. Ficklin SP, Sanderson LA, Cheng CH, *et al.*: **Tripal: a construction toolkit for online genome databases.** *Database (Oxford).* 2011; **2011**: bar044.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Papanicolaou A, Heckel DG: **The GMOD Drupal bioinformatic server framework.** *Bioinformatics.* 2010; **26**(24): 3119–24.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Hubbard T, Barker D, Birney E, *et al.*: **The Ensembl genome database project.** *Nucleic Acids Res.* 2002; **30**(1): 38–41.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Kalderimis A, Lyne R, Butano D, *et al.*: **InterMine: extensive web services for modern biology.** *Nucleic Acids Res.* 2014; **42**(Web Server issue): W468–472.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Goff SA, Vaughn M, McKay S, *et al.*: **The iPlant Collaborative: Cyberinfrastructure for Plant Biology.** *Front Plant Sci.* 2011; **2**: 34.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Goldman M, Craft B, Swatloski T, *et al.*: **The UCSC Cancer Genomics Browser:**

**update 2015.** *Nucleic Acids Res.* 2015; **43**(Database issue): D812–D817.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Poelchau M, Childers C, Moore G, *et al.*: **The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes.** *Nucleic Acids Res.* 2015; **43**(Database issue): D714–D719.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Lewis SE, Searle SM, Harris N, *et al.*: **Apollo: a sequence annotation editor.** *Genome Biol.* 2002; **3**(12): RESEARCH0082.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Elsik CG, Worley KC, Zhang L, *et al.*: **Community annotation: procedures, protocols, and supporting tools.** *Genome Res.* 2006; **16**(11): 1329–33.
PubMed Abstract | Publisher Full Text

32. Skinner ME, Uzilov AV, Stein LD, *et al.*: **JBrowse: a next-generation genome browser.** *Genome Res.* 2009; **19**(9): 1630–8.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Lee E, Helt GA, Reese JT, *et al.*: **Web Apollo: a web-based genomic annotation editing platform.** *Genome Biol.* 2013; **14**(8): R93.
PubMed Abstract | Publisher Full Text | Free Full Text

# Open Peer Review

## Current Referee Status: ✓ ? ✓

**Version 1**

Referee Report 15 March 2016

**doi:**10.5256/f1000research.8139.r11817

**Ian Holmes**
Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA

This is a very nice opinion piece surveying the current community practices around genome projects and the shortcomings thereof. Using the metaphor of the insect lifecycle, the paper discusses the “genome lifecycle” (sequencing, assembly, annotation… re-sequencing, re-annotation, etc) and various lessons drawn from real-life case studies (e.g. informatics is key, perfection is the enemy of the good, we need to decouple data dissemination from publication to some extent, we need plans for sustained computational infrastructure, we need new collaborative tools).

I agree with the positions espoused here and I find this piece a very insightful distillation of the challenges facing the community as funding pay-lines become tighter and we transition from an era of quick genome-project headlines to one in which the community can (with luck) collectively curate and maintain data, rather than letting data-silos decay.

I had one minor suggested edit which is that the line “until the skill and leadership of Jim Kent saved the day” could use a citation (presumably to Kent & Haussler, 2001[1]).

**References**

1. Kent WJ, Haussler D: Assembly of the working draft of the human genome with GigAssembler.*Genome Res*. 2001; **11** (9): 1541-8 PubMed Abstract | Publisher Full Text

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

***Competing Interests:*** No competing interests were disclosed.

Referee Report 15 January 2016

**doi:**10.5256/f1000research.8139.r11819

**John W. Davey**
Department of Zoology, University of Cambridge, Cambridge, UK

This article draws attention to several important issues related to the production and use of reference genomes and associated data sets, particularly gene annotations. It outlines the typical process for a

genome project in the era when large consortia could acquire substantial funding for such a project and produce high-impact genome papers. It correctly notes that the process of generating genomes is rapidly changing, as sequencing costs are falling and tools are improving, allowing small groups to produce genomes for a fraction of previous costs, but also with reduced impact. This change has many scientific and political implications for the production of genomes, which the article attempts to summarise. However, I do not think it does a good job of summarising or addressing these implications. I hope the following criticisms will help to bring these important issues into the clearest possible light.

The article highlights the conflict between the nature of a reference genome as a resource that requires long-term, communal effort and infrastructure, and the nature of science funding, which requires the repeated completion of substantial short-term goals leading to high-impact first- and last-author papers. It claims that genome projects are best undertaken with particular biological questions in mind, in order to deliver the most relevant possible resource at the time and to avoid perfectionism and the distraction of new sequencing technologies and assembly tools. It also rightly insists that genome projects require complex computational analyses that need to be understood to some extent by those producing and using the genome so that potential errors in analyses using the genome are well understood and can be fixed where possible. It calls for the active building and maintenance of communities of researchers working on particular genomes, partly through the training of scientists in genome assembly and annotation techniques. It also calls for the decoupling of publications from resources by rewarding efforts for their real world impacts, and for increasing the likelihood of real world impacts by making genomes available early and encouraging communities of researchers to use them. Finally, it highlights the importance of core teams such as that at the National Agricultural Library for disseminating training, hosting data, and providing tools and platforms in order to support individual groups working on genome projects, and the need to secure long-term funding for these teams.

All of these points are important and worth making strongly (with a few caveats), both for those already involved in genome assembly and annotation and those new to the field. But they do not come across clearly in the article, as they are hobbled by inconsistent use of various concepts and by several bad examples that hurt the case being made.

Firstly, the concept of a community is very unclear. There are 44 references to communities in the article, including the non-model species community, the genomics community, the insect community, the insect genomics community, the research community, the wider community, the software engineering community, the i5k community and the informatics community, but mostly just to 'the community'. There are then 39 references to what 'we' are doing or should do. Who are we? Which community or communities does 'we' refer to in each case, and who is the article addressed to? It is not clear who needs to do what differently in order to improve the situation, or where the problems really lie. To give one example, "Third, the research community is not just the end-user but also part of the project team; we have, on the whole, neglected to bring them up to speed. These issues may seem intuitive but much of the community’s leadership is not conscious of it." Who have neglected to bring the research community up to speed? i5k? Informaticians? Which community's leadership is not conscious of the problems, the i5k community, the informatics community, the research community? The article at points seems to be attempting to address a general audience, but other times is directed to the i5k community, and at points seems to be saying i5k should be doing things differently, but at others recommending the i5k model as best practice. While these things are not necessarily incompatible, the article would be much easier to read if it was much clearer about the groups it is addressing, and the social structures that would improve the process of generating genomes.

Secondly, the contrast between using genomes to answer questions and providing genomes as

resources is passed over, with both being claimed as important while not addressing the conflict between them. 'Genomics' is sometimes used to refer strictly to the production of genome sequences and perhaps annotations, but sometimes to research done using these sequences and annotations. The concept of a genome project being an experiment is confused in the same way; sometimes it seems to refer to the genome assembly as an experiment itself, and sometimes to the genome as used to conduct an experiment that answers a biological question (which can drive the genome project design). And the same confusion arises over the encouragement for scientists to learn 'data science' and related fields; sometimes this is directly related to genome assembly, sometimes to research using the genome.

Clarity on these issues is important, because at present the confusion obscures some hard problems. For example, while it is highly desirable to direct genome projects towards particular biological questions (to maximise the chance of funding and high-profile papers, and to circumscribe the limits of the genome project itself), and to engage the widest possible relevant community in the production of genome resources (to make sure the genomes are used correctly, to increase impact, and hopefully to increase quality of assembly and annotation), the article doesn't bring out explicitly the fact that these goals are antithetical. As more groups become involved in a genome project, the number of relevant biological questions increases, and the quality of the genome must increase to accommodate them, making it harder to design and manage the project (especially if the entire community is to be involved in not only the annotation but also the assembly, and do research on the genome along the way, as recommended in Insight 3).

Also, while the article makes several welcome calls for better genomics education, the confusions over what the relevant communities are and the distinction between use and provision of genomes make it very unclear what the nature and extent of this education should be. Is the article arguing that bioinformaticians should do the assemblies but educate biologists in the limitations of genomes they deliver; or that bioinformaticians should train biologists to assemble and annotate genomes themselves; or that the role of bioinformatician should disappear and biologists should do it all themselves, given that all biology is computational these days; or that biologists should use more data science techniques on their research but leave the genome assembly up to dedicated bioinformaticians? The article seems to be arguing for variations on these possibilities at different points.

I don't know what the answers are to these problems, but at least they should be brought out clearly in the article, rather than left obscure. With this in mind, I will turn to individual comments on the insight sections.

Insight 1: there is an important point here, which is that results may vary greatly depending on how software is used, and a more basic one, which is that computation is (and always has been) required to produce genomes and so those who wish to produce a genome need to engage with computational analyses. But these points are obscured by more confusions and irrelevant or inaccurate points.

The opening point, that genomics has moved from being led by and limited by wet lab techniques to being led by computational science, is highly debatable, and I personally don't agree, unless perhaps if 'genomics' here means biology in general. Genomics in the sense of genome assembly continues to be led by the available sequencing technologies, not by computation - the two current assembly methods referred to in the paper, Allpaths and Discovar, were both designed to fit an available sequencing technology, they did not prompt the development of the technology. In long read sequencing too, the technology is driving the algorithms, not the other way around. While the Celera assembler is a great achievement and is being heavily used to assemble long read sequences, it is far from the case that Pacific Biosciences and Oxford Nanopore are designing their machines to fit the design of the Celera assembler. And the human genome example doesn't support the case at all, given that, as noted, 'the

WGS approach [was] rightly considered to be of inferior quality' and the private genome ended up incorporating a lot of the public mapping data; the article ends up concluding that it was the 'cross-talk of the two capabilities' that was important, contradicting the initial point of the paragraph.

The second paragraph attempts to make the case for biologists to develop computational skills, but the range of terms used just further obfuscates the issue. What is the difference between information technologies, informatics, data science, information science and "big data science"? How exactly are they related to genomics and bioinformatics? What distinct 'epistemological understanding' does statistics and informatics provide that biology does not, and what is a 'framework to make sense of a more synthesized knowledge' (and why should a researcher want it)? If the point is to say biologists would benefit in general if they improved their computational skills, that may be true, but is isn't really relevant to an article about genome assembly (and it is mildly insulting to say biologists need to improve their statistical skills, given that they invented statistics). If the point is to say biologists need to engage with genome assembly and annotation, that's quite a different issue, and doesn't need to be backed up by the general case for computational training. Reducing the generalities about computation and increasing the specificities about how biologists need to engage with genome assembly and annotation would help here.

Insight 2: again, several very different points are mixed up into one here. Genome projects to date do tend to follow a life cycle as described, and can be iterated. But the points that follow, especially those about experiments, are confused. It is true that a genome sequence can be used to test hypotheses, and that the relevant hypotheses can often direct the design of a genome project. But in what sense is 'creating a draft genome sequence' an experiment? What is the hypothesis being tested by the assembly process itself? In what sense is 'investigating an organism's genetic blueprint' a hypothesis-driven experiment? It's possible to make the analogy (perhaps every time an assembler compares two reads, it conducts an experiment to test whether the reads overlap or not?) but it is not very enlightening, and it is not necessary for making the case that good project design is essential, that a variety of methods can be used and that there is a risk of failure - many things other than experiments share these properties. The sentence about the computer science point of view is even more confusing; the question "what is the correct genome sequence for this species" does not require an experiment in the traditional sense, and "what are the parts that are important for its function" isn't really a computation-only question at all.

Further, the advice here is quite convoluted - "when we are not satisfied we have to backtrack", but "More experienced workers also learn that once a stage is satisfactorily completed... one must under no circumstances go back", however, "if the stage is not satisfactory that... we go back one step". Clearly satisfaction is key here, but our satisfaction can change - and if our satisfaction about an earlier stage is changed by what we discover at a later stage, does that mean "one must under no circumstances go back"?

While genome projects to date have followed a life cycle as described, and perhaps initial versions of a genome may need to follow this process, I'm not convinced that a strict adherence to this model for future iterations is helpful. Insisting that every stage of the life cycle is completed by the whole community step by step in order to lead to a paper of lower and lower impact is surely the model we want to get away from. There is decades of research in software engineering refining or rejecting completely this kind of waterfall model in favour of more incremental approaches; while there is still controversy over this, it seems likely that genomics could benefit from moving in this direction as well.

In theory, there is no reason why genomes can't be patched and updated piecemeal as small assembly errors are fixed, or scaffolds are ordered, or single gene families are annotated, with infrequent major releases rolling together these patches. This is standard practice in the software industry and for the

human genome. I don't claim this is the only way to do things, or that there aren't problems with this approach, and it is true the infrastructure is not in place to do this efficiently for non-model species. But that doesn't mean we should restrict ourselves to the existing life cycle model; adhering to this model is one of the causes of the problems the article is trying to address (big version releases lead to problems in acquiring funding, managing large communities, rewarding individual contributors, deciding on publication strategy...). Why not just change the model?

Finally, the point about genome assembly often being limited by the biology of the organism is valid, but the example is a poor fit and should be removed. The Heliconius Discovar assemblies were never intended to provide reference-quality assemblies, as the Allpaths-LG assembly was, and the biology of the organism was not ignored, as the paragraph implies; in fact, the Discovar assemblies were specifically intended to test the Discovar assembler on a set of highly heterozygous genomes, and improve the assembler to deal with this data. The assemblies were preliminary and were never optimised because the Discovar team left the Broad and did not complete the project, so it isn't fair to compare the assemblies. A better example to support this point would be the Plutella xylostella genome, where considerable heterozygosity remained after ten generations of inbreeding and thorough fosmid sequencing was required to produce a genome of reasonable quality.

Insight 3: the issues described here (how to provide informatic support beyond the initial publication and how to create a sustainable publishing model that allows for a genome project life cycle) are real, but the solutions provided are not very realistic, are already fairly standard practice, or do not address the issues. Three options are presented: submit new genome versions as low-impact technical papers; use the new version to address a new biological question, or (the preferred option) to decouple publications from resources and respect the value of the genomic resources. This last option might well be a good idea, but the proposals for achieving it fall short.

Most of the points made (releasing data early, engaging the community and allowing them to publish before the genome is published in its own right, showcasing a wide variety of analyses in the eventual genome paper) are to do with the initial release of the genome, not how to maintain the genome beyond its initial publication. There isn't much new in these points, given that this is the template set by the human genome project, but that doesn't necessarily mean it's not worth highlighting them again. However, it should be noted that this very fluid use of data, where the community edits and improves the assembly and annotation, makes maintaining a strict life cycle with frozen stages even harder.

The only point this paragraph does make about later versions of the genome is that they should be linked to new experimental work or multi-species comparative genomic insights - which is just the second option that was passed over earlier. Also, the first option is passed over because it is unappealing to the best bioinformaticians, but why should a bioinformatician working under the standard publishing model where first-author papers are required be more interested in the proposed model where a large range of community analyses, some perhaps previously published (and so lowering their impact or making them inadmissable for further publication), are put into one paper?

The problem is correctly identified as the conflict between the publishing model for individual scientists and the need to build communal resources, but the text doesn't propose anything meaningful to address this, beyond insisting that it would be good to separate publications from resources. But how is this to be done? Which communities need to change what they are doing, and how they value work, to achieve this? What metrics should we be using and recommending to faculty in hiring computational biologists, other than publications? While touching on this issue, the article does not really address it, and does not extended 'real world impact' beyond the use of the data by other researchers. If this is the limit, what is

wrong with the current system where impact is measured by the proxy of citations?

Finally, the whole manuscript would benefit from more attention to detail. For example, "Second, this draft does indeed contain many of the instructions of how to generate an organism, but a genome sequence alone does not decipher it. It merely transcribes so we can conduct experiments with it. Deciphering will require both good experimental design and the capability to integrate such experiments." - what do the two consecutive 'it's refer to? The organism, then the genome sequence? How does a genome sequence transcribe? What is being deciphered? What are the experiments being integrated with? "this number is increasing exponentially" - is it exponential? "an achievement which may be currently underutilised but whose importance cannot be understated" - surely overstated, but the hyperbole doesn't help here anyway. It's not convincing to just say the work is important; why is it so important?

I am sorry to be so critical, especially in public. I hope this level of detail will be taken as a mark of respect for Dr Papanicolaou's expertise and passion for this subject, which I agree is a very important topic that needs to be engaged with by all involved. I thank him for stepping forward to raise these issues and hope that this review will be taken constructively and lead to improvements in the piece.

The following typos or omitted words should be fixed:
the genomes projects of a larger part the tree of life
One the major forces of innovation
dataset: The
explain what and how a software works
For example, genome project can go through
allowed us generate genomes
Richard and Murali
for their contribution in way appreciated by
on the automated approaches on the underlying data
that lacks the immediacy and
Except for offering a real-time -> Because?
may also drive many of the next generation of synthesis in biology. -> syntheses?

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

***Competing Interests:*** Dr Papanicolaou and I are colleagues who have both been involved in Heliconius genomics for many years and were both part of the Heliconius Genome Consortium.

Referee Report 12 January 2016

**doi:**10.5256/f1000research.8139.r11905

**Stephen Richards**
Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

This is an excellent review of the genome project process and life cycle, that most valuably shows the reader how a genome project fits into the larger goal of biological research around a species. Dr. Papanicolaou's insights remind us how genome scale data "completely changed our perception of how

biological research can scale in a world that transcends borders" and provides passionate enthusiasm and advice for those researchers and communities without genomes who wish to join this new world.

Insight 1 tells us about the power and necessity of bioinformatics for not just assembly and annotation, but the genome wide analyses required to gain the most biological insight, but additionally warns us against relying on the "black box" that software can become without understanding how it works. I for one have fallen for this by following the conventional wisdom about what a particular software package does, but finding out only by inspection of the code that something else entirely is going on.

Insight 2 places the genome project in it's rightful place as part of the experimental life cycle. The emphasis on experimental design in genomics is important - all to often in the past this critical step is ignored either due to cost reasons in collecting a sufficient data set compared with the urge to do something and call it preliminary data or simply out of bravado and not planning, with the result of very poor genome assembly quality and unreliable or un-interpretable downstream analyses. Dr. Papanicolaou outlines the importance of things like community curation to simply enable the community to look closely at the data and gene models - something that is vital to get an idea of how much to trust any conclusions coming out. but at the same time to match genome quality to the desired experimental requirements, to freeze genomes and annotations, and to get on with it and publish. This is excellent advice, and many a genome projects publication has been stalled for multiple years in the pursuit of better quality without the contaminant investment of resources, or in slow transition between the various steps of analysis dues to poor planning and training for the entire process of the life cycle. This part of the review is required reading for anyone contemplating a genome project, directing the thoughts of the reader to consider the longer term plan for the genome for his or her's lab, experiment or even for a larger community, and tailoring the experimental plan to fit.

The insights on Data Sharing, and the requirement for pre-publication data sharing are critical, and point the reader to resources that will enable placing new genome datasets in public repositories with long funding horizons, stable futures, and academic reach around the board. More interesting to this reviewer, was the discussion on the difficulties in funding and publishing the improvement of draft genomes in the future. Although this is getting technically easier with he advent of longer read sequencing technologies, the manuscript is correct in noting the difficulties in publishing a fourth improved draft genome compared with the third - it is hard to say it is a significant improvement to our state of knowledge when closing say 75% of the gaps. I believe in the future we will still be interested in "effectively finished" archival genomes, and that these will be worth data notes in lower impact journals, but the option of "decouple publications from resources while at the same time respecting the value of genomics " to me seems like the correct way forward as we one day hope to have sequenced all species on the planet - i.e. to read the primary biological data for life on earth. Whilst we realize the genome sequence of the 10,000th bird species may not make the highest profile journal, not to have this sequence in the natural history museums of the future seems unthinkable.

The human touch insight is dedicated to the need for researchers to look at data to correct gene models, to understand the limits of the dataset. New tools allow this to be done in a co-ordinated manner with groups of researchers from around the world, with the result that research can be accelerated around the world with the sharing of a single genome. This is particularly true today with the use of RNAi and Crispr gene manipulation techniques. In the milkweed bug community RNAi was the mainstay of comparative developmental research, but relied on degenerate PCR to identify genes and design probes. A draft genome quickly gave this research community the information to design all the probes they needed, but human curation was still needed to checkoff the number of genes in a family had changed from the Drosophila model, or that the automated gene model had got the sequence right before committing to a

wet lab experiment, and that phylogenetic trees had confirmed that the researcher was manipulating the gene he or she thought she was, and not a paralog or a gene from a different but related family.

Overall Dr. Papanicolaou has written an excellent guide to the genome project, the reading of which will profit anyone contemplating a genome project. It is well written, and whilst I have a few differences of opinion on minor points, they are in no means enough to prevent indexation. Overall I believe this manuscript merits immediate indexation with no modification necessary.

Bonus points for remembering and reminding us of the role of Jim Kent.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

***Competing Interests:*** No competing interests were disclosed.