GigaBlog

Extended Q&A with Assemblathon2 Author Keith Bradnam

Scott Edmunds

Published August 20, 2013

Citation
Edmunds, S. (n.d.). In *GigaBlog*. GigaBlog. https://doi.org/10.59350/vmb4z-g1e81

Keywords
Technology, Assemblathon, Competitions, Genome Assembly, Genomics
Feature Image

A *lot* has already been written about last months Assemblathon2 paper in *GigaScience* (see the growing list of articles here), but for the box-set completists interested in squeezing every last bit of insight into the project and how it was put together, there was a lot of additional material left over from the recent Biome Q&A with Keith Bradnam that we thought it could be useful to post in a (hopefully final) blog posting. Keith is a project scientist in the Korf lab at UC Davis where he has juggled investigating intronic-based signals in gene expression and running the Assemblathon. This is the second Q&A we have collected in Gigablog, after Xin Zhou provided an behind-the-scenes overview of his insect "squishome" metabarcoding paper, and we hope to provide more of these in the future that you will be able to search using the Q&A tag.

How did the Assemblathon come about?

The Assemblathon developed as an offshoot of the Genome 10K project, which is aiming to sequence the genomes of 10,000 vertebrates. There is almost no point in attempting to sequence so many genomes if we are not reasonably sure that we can accurately assemble those genomes from the initial sequencing read data. As there are so many different tools out there to perform *de novo* assembly, it seemed to make sense to try to benchmark some of them.

Why is it important to assess genome assembly?

There are many areas of genomics where a researcher can find a plethora of bioinformatics tools that all try to solve the same problem. However, different software tools often produce very different answers. Even the same tool can generate very different answers if you explore all of its parameters and configuration options. It is not always obvious how the results from different tools stack up against each other, and it is not always obvious as to which tools we should trust the most (if any).

If you wanted to know who made the best chili in your local area, then you could organize a 'chili cook off'. As well as deciding an overall winner, you could also award prizes for specific categories of chili (best vegetarian chili, best low-fat recipe etc.). What we can do for chili we can also do for genome assemblers.

Contests like the Assemblathon can help reveal the differences in how different genome assemblers perform, and can also pinpoint the specific areas where one program might outperform another. However, just as tasting chili can be a very subjective experience, there can be similar issues when evaluating genome assemblers. One of the objectives for the Assemblathon was to try to get a handle on what 'best' means in the context of genome assembly.

As the majority of genomes are currently sequenced using the same Illumina technology, are things pretty standardized?

The landscape of sequencing technology is constantly shifting. Illumina has a strong grip on the sequencing market at the moment; however, there are many other players out there (Ion

Torrent, Roche 454, Pacific BioSciences, ABI SOLiD), not to mention others who may or may not shake up the industry in the near future (e.g. Oxford Nanopore, Nabsys).

There can also be a lot of diversity *within* a single platform. The first Illumina (then Solexa) sequencing reads were only 25 bp long; the latest Illumina technology is promising maybe a 10-fold increase in length. Writing software to optimize the assembly of 25 bp reads is a different problem than optimizing the assembly of 250 bp reads. So even if we only had a single platform, there would still be a need to evaluate and reevaluate assembly software.

What was the rationale for using a different approach to Assemblathon 1, where you used synthetic data?

It helps to put a jigsaw together when you have the picture on the box to help you! Most of the time that people perform genome assembly, they don't have that picture. So it can be hard to know whether you even have all of the genome present (let alone whether it is accurately put together). In Assemblathon 1, we wanted to know what the answer was going to be *before* people started assembling data, so we used an artificial genome that was created in a way that tried to preserve some properties of a real genome.

However, for Assemblathon 2 there was a lot of interest in working with real world data. The genome assembly community wants to be solving problems that can actually help others with their research.

Why did you pick the three species that you did (budgie, boa constrictor and cichlid fish)?

A major factor was simply the availability of suitable sequencing data. But it also seemed a good idea to use some fairly diverse species, the genomes of which might pose different types of challenge for genome assemblers (different repeat content, heterozygosity etc.).

What were the main findings of Assemblathon2?

To paraphrase Abraham Lincoln: you can get an assembler to perform well across all metrics in some species, across some metrics in all species, but you can't get an assembler to perform well across all metrics in all species.

Were you surprised by the huge variations in the results?

Yes and no. Personally speaking, I was expecting to see variation in performance between assemblers but I thought that some of the bigger 'brand name' assemblers might have shown more consistency when assessed across species.

What were the challenges of coordinating such a large project?

Coordinating 91 co-authors on *any* project is tough, but in the case of Assemblathon 2 most of those co-authors were from competing teams. Therefore you have to be careful to ensure that suggestions to change any detail of the project/paper are all in the best scientific interests of everyone concerned. Data distribution of the raw input data was also a challenge and was helped greatly by the Pittsburgh Supercomputing Center that helped make the data available

for downloading. In many cases, it was easier for collaborators to use the 'sneakernet' approach and just send us hard drives in the post.

The way you have embraced pre-prints and blogging, and the way the more interactive peer-review came about is still quite unusual for biology. What is the appeal of making your work available in arXiv, and the open peer-review that *GigaScience* carries out?

Personally, I am a strong advocate that results from tax-payer funded research should become publicly available a.s.a.p. I think we can demonstrably show — by the volume of blog posts that have now discussed the Assemblathon 2 project — that the resulting conversation about genome assembly has been a useful one. I really hope that more journals adopt open peer-review and encourage the use of pre-print servers.

Has the experience changed the way you plan to publish your own work in the future, or review others work?

I think it has made me even more committed to support open access publishing and open peer-review (I made the decision about a year ago to start signing my reviews, even when not required). It has also made me more determined to write more on my personal blog about bioinformatics and genomics. Blogging and tweeting about your work can raise your visibility within your field, can help hone your writing skills, and may even help you compete with others when applying for jobs.

Titus Brown, who openly blogged about the paper whilst in peer review, suggests that scientists do a very poor job of communicating the uncertainty in genome assemblies 'an indictment of much of computational biology'. Do you have such a bleak view? What do we need to do to improve this?

To a large degree, I agree with Titus. The genomics community often talks about 'drowning in data', but we are also 'drowning in tools to analyze that data'. This is especially so in certain areas, e.g. the number of programs available that map short-reads back to genomes is overwhelming. Sometimes it seems that we are in an endless race to keep on publishing new tools without pausing to really check how well the existing tools work (or whether they work at all). Maybe we need a dedicated journal, or online resource, that does nothing other than assess the comparative performance of bioinformatics tools. We should also be encouraging students — i.e. the next generation of bioinformaticians — to perhaps be more distrustful of these tools. Just because a particular piece of software gives you an answer, doesn't make it the right answer (or the only answer).

On top of what we are doing wrong with genome assembly, what are we doing right? Are there any positives you have found from the project?

One of the issues that has partly limited the field of genome assembly is that we have mostly stuck with the same sequence file formats (FASTA and FASTQ). These formats limit us to only displaying *one* version of a genome sequence, even when we know about variants that might be present. This is a bit like having to tick a single box on questionnaires that ask you about

your ethnicity. If you are half Japanese and half Swedish, which box do you tick? Choosing one option means you lose a whole lot of information. Currently, a genome assembly in FASTA format is an assembly that has to tick that one box over and over again.

Fortunately, an initiative led by the Broad Institute has led to the development of a new 'FASTG' sequence format. This format allows a genome assembly sequence to describe haplotype and other variation. For example, a contig might contain a region that consists of ACGT followed by 2 or 3 Cs, optionally followed by a G. The FASTG format allows you to capture this variation and this will allow for a much better representation of genome assemblies. A true genome sequence is usually a mix of two parental genomes, and we should be representing this.

The results from each of the teams on each of the genomes was very mixed, but would you say there were some entries that were better overall? One criticism of Assemblathon is that to get so many groups to continue to take part, you've been reluctant to promote winners and losers. How do you respond to this, and can you tell us who the real winner was?

We were very diplomatic with how we described the outcome of Assemblathon 1, where some assemblers performed consistently better than others. But in Assemblathon 2, there was so much variation it seemed hard to say that any single team should be declared a winner. At best we can say that in a given species, and when considering a specific genome assembly metric, that there were winners i.e. some assemblers will give you longer contigs for the fish genome, some others might capture more of the genes in the snake genome, and others will give better coverage of the bird genome. These may all be different assemblers. This is not necessarily what people were hoping to hear, but that's what happened. Hopefully, this information will still be valuable to people though.

There has been a lot of criticism lately of large consortia driven projects such as ENCODE. Having had to coordinate over 20 groups, and 90 authors on this paper, has this changed your perspective on this? What have you learned from the whole experience, and do you think you will tackle anything on this scale again?

Assemblathon 2 is a bit different to projects like ENCODE in that it was a contest that was predominantly organized by a fairly small group (three of us at the UC Davis Genome Center). The large number of co-authors primarily reflect the composition of the 21 competing teams that took part. I have learned a lot from the experience, especially not to underestimate the challenges involved when coordinating so many people! A quick tip to anyone else in a similar position: using Google Docs to coordinate feedback on a draft manuscript is a viable — i.e. sanity-restoring — alternative to instead emailing many PDFs or Word Docs back and forth and having to somehow assimilate the resulting melee of comments.

It's hard to imagine doing something like this again, but who knows.

Do you think there will be an Assemblathon3? If so, what do you think you would do differently next time?

I'm in the process of writing a blog post on this very topic (to be posted at http://assemblathon.org). There are many things that could, and perhaps should, be done differently. For starters, if the community embraces the FASTG format that I mentioned earlier, it would make sense to use this as the format of choice for Assemblathon 3. Perhaps more importantly, we should find a different lead author!

Further reading:

1. Keith R Bradnam *et al.*, Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species *GigaScience* 2013 **2**:10 http://dx.doi.org/10.1186/2047-217X-2-10

UPDATE 22/8/13 Keith has now posted his "Is there going to be an Assemblathon 3?" blog posting here.

The post Extended Q&A with Assemblathon2 Author Keith Bradnam appeared first on GigaBlog.