SciLifeLab Best practices in de novo genome assembly

Introduction

In genome assembly projects a genome is sequenced with short sequence reads which are then “glued” together, or in other words “assembled”, to recreate the genomic sequence. Nowadays this is usually done using some kind of Next Generation Sequencing (NGS) technology. This document includes recommendations that are good to think about before going into a genome assembly project.

The first question you need to ask yourself is, “Do I really need to assemble the genome?”. An assembled genome constitutes a huge resource and can be a great help in many projects, but it is also quite time consuming and expensive to get to the final assembled genome. A more light-weight option is to sequence and assemble the transcriptome. In this case mature mRNA are sequenced and assembled into complete transcripts. These can then be used to understand the repertoire of genes in the study organism, in differential expression studies, as phylogenetic markers etc. This can be enough for your needs. There are limitations of course. You are limiting yourself to what is transcribed in your samples, and you will know nothing about non-coding parts or be able to understand anything about the large scale structure or gene order of the genome.

A genome will give you more options than a transcriptome, but it is good to be aware of both possibilities and you should decide on an approach that will be best for your research questions.

First, do you need help?

In general, we would only recommend that you assemble and annotate the genome yourself if you know you will return to similar projects repeatedly and want to build up that competence in your group. If your group does not have that competence, we would recommend you to get help from us.

The SciLifeLab sequencing facilities in Stockholm and Uppsala where you order your sequencing data can, if the workload allows and the genome follows certain criteria, also assemble the genome for you, and we strongly suggest you accept this offer. However, your genome will not be annotated, and the team will only be able to help you if you also order the sequence data from them.

Another possibility is to turn to the NBIS Assembly and Annotation service. We can annotate the genome you received from the ScLifeLab sequencing facility or take responsibility for the

whole assembly and annotation process. You can reach the team at http://nbis.se/support/supportform/index.php. Please choose “genome assembly or annotation” under “subject”. We cannot answer the biological questions for you, but we can give you the time you need to focus on that.

Pre-assembly considerations

Here follows some aspects of genome assembly that are good to consider before ordering sequence data.

What do you want to use the assembly for?

If you are interested in comparing your genome with other genomes to find structural rearrangements, then you need the assembly to be as contiguous as possible. If you on the other hand only are interested in protein coding genes, and more specifically in the nucleotide sequence of these genes, then the contigs assembled only need to be long enough to include complete genes.

The difference between these two alternatives can be a lot of time and money. Long-range information, like long mate-pairs or longer sequences with PacBio RSII can significantly improve the contiguity of the genome assembly, but is also costly. Think about what you need for your purposes. Always go into a genome sequencing project with a hypothesis you want to test so you know why you are sequencing the genome.

Organism-specific details

Some properties of the organism you are studying will greatly influence the assembly process.

- Genome size - Larger genomes will require more sequence data, and thus increase sequencing costs. If possible, get an idea of the size of the genome you are studying before going into the assembly project. The genome size is also a very good number to bring to the sequencing center, as they can then help you calculate the amount of data you need to order. Also note that assembly program memory requirements increases with genome size, and that for larger genomes you will be restricted to run on a few select servers in Sweden.
- Repeat content - Sequence stretches that are found identically in different parts of the genome are called repeats. In particular for eukaryotes, repeats are very common, but the exact amount and distribution of the repeats differ between organisms. Assembly programs will be confused by these stretches as reads coming from these regions will be identical, and this can lead to incorrect assemblies. Longer sequence reads that go all through the repeat sequence into the unique sequence bordering the repeat will help greatly, as will sequencing libraries with greater insert sizes (e.g., Illumina mate-pairs, or PacBio long reads). A high repeat content will lead

to a more fragmented assembly, and if a high level of repeats are known from your study organism you should set your expectations accordingly.

- Heterozygosity - Assembly programs in general try to collapse allelic differences into one consensus sequence, so that the final assembly that is reported is haploid. If the genome is highly heterozygotic, sequence reads from homologous alleles will be too different to be assembled together and these alleles will then be assembled separately. This means that heterozygotic regions might be reported twice, while less variable regions will only be reported once, or that the assembly simply fails at these variable regions. Highly heterozygotic genomes can lead to more fragmented assemblies, or create doubt about the homology of the contigs. Large population sizes tend to lead to high heterozygosity levels, so marine organisms often have high heterozygosity levels and are often problematic to assemble. If you have the possibility to sequence inbred individuals, then this is recommended.
- Ploidy level - If you have the possibility to sequence haploid tissue (true for bacteria and many fungi), this will essentially remove any problems caused by heterozygosity (see above) and is preferable. Diploid tissues, which will be the case for most animals and plants, is fine and usually manageable, while tetraploidy and above has the potential to greatly increase the number of present alleles, which likely will result in a more fragmented assembly (once again, see heterozygosity above). Diploid assemblers of long reads are available and used at SciLifeLab, however please keep in mind that correct assembly of diploid genomes might require higher coverage.
- Size of the organism - If the organisms you are working with are small then you might have problems extracting sufficient amounts of DNA. Pooling of individuals is usually not recommended and should only be done if the individuals are close to identical in their genome sequence, otherwise you risk creating a DNA sample that is too heterogeneous and hard to assemble. If you know you will have problems like these, please contact the sequencing platform to get the best possible and up to date advice.
- Symbionts and parasites - Sequences from other organisms can contaminate your sequencing data, and make assembly more complex.
- GC content (highly relevant for bacterial genomes). Most of the sequencing technologies have strong bias when reading through either AT or GC rich DNA. For such genomes, use of PacBio reads is highly recommended.

Do you want to annotate the genome?

Annotation is the process in which genes and other features of the genome are inferred. Without annotation your newly assembled genome will just be a long stretch of nucleotides and you will not know where the genes are or what they do.

If you want to annotate the genome, which almost always is the case, we strongly recommend you also order RNA-seq (sequenced mRNA) when you order your sequence data. The RNA-seq will be aligned to the assembled genome in the annotation process and used to structurally infer the genes in the genome. This will greatly improve the quality of the annotation and is for many organisms a must to achieve a good quality annotation, in

particular if your organism is only distantly related to other sequenced and annotated genomes.

No replicates are needed, but we recommend sequencing different tissues if possible or different life stages so that as many genes as possible are expressed in your extracted RNA. If possible, extract RNA from the same individual as in the genomic DNA extraction. For organisms with a lot of genetic variation in the populations it might be very hard to align RNA-seq reads from one individual to the genome of another.

DNA extraction

Quality of the input DNA is crucial for success of mate-pair and PacBio sequencing. For these applications the sequencing facilities will require that you submit a gel-picture of the DNA, as well as 260/280 and 260/230 absorption ratios. When estimating amount of DNA in sample, please bear in mind that NanoDrop and NanoVue tend to over-estimate the overall DNA concentration with up to 500%. It is strongly recommended to use either PicoGreen or Qubit measurements. Is those are unavailable at your facility, aim for 5x higher NanoDrop assessment than required by the sequencing facility. For more information on sample prep for PacBio sequencing, please read the following information: https://portal.scilifelab.se/genomics/sites/default/files/Instructions%20for%20sample%20preparation%20for%20PacBio%20RSII%20instrument.pdf; watch the recorded webinar: https://s3.amazonaws.com/files.pacb.com/webinar/2015-09-Uppsala/Webinar-PacBio-Uppsala-Recording-Sept2015.mp4 , or download a PowerPoint presentation: programs.pacificbiosciences.com/e/1652/ala-Presentation-Sept2015-pptx/3gr93d/490791557.
If unsure about the extraction method, kindly consult the sequencing facility prior to onset of the project.

Sequence technologies

At the moment, there are two dominating sequence technologies in de novo genome assembly: Illumina and Pacific Biosciences. In general, assembly programs are made to work with one of these technologies, not both. There are programs that accept both, but in comparisons the programs that focus on one single technology seem to perform best. This could change quickly though, please contact the NBIS assembly and annotation team to get the best up to date advice.

The two technologies are very different, here are their main properties.

Illumina:

- Short sequence length (up to 300)
- High yield
- Low error rate
- Paired-end and mate-pair sequencing (for long range information)
- Price per nucleotide low

Pacific Biosciences (PacBio):

- Long sequence length (1000-40000 bp, or even more)
- Low yield
- Higher error rate (in particular for longer reads, but in practice rarely a problem for de novo assemblies)
- Single end only
- Price per nucleotide high
- Can contain base modification information, such as methylation
- Require less hands-on bioinformatics.

Comparison, pros and cons:
The long sequence length of PacBio makes it ideally suited for de novo genome assembly. The reads are often long enough to go through repeats, and heterozygosity is also less of a problem as the reads are long enough for the assembly program to effectively differ between alleles.

Illumina on the hand has shorter reads, but gets the necessary long range information through mate-pair sequencing. Here fragments of up to 40 kbp are sequenced from two directions. The sequences will be short, but approximate distance between the reads will be known and also simply the fact that they are known to be connected can be used to improve the contiguity of the assembly. In practice, including mate pair sequences and not only paired end sequences in the assembly makes a huge difference in assembly contiguity.

That being said, longer reads are always preferable, and PacBio has the upper hand there. The problem is the cost. We will therefore suggest a few different solutions, based on genome size, but if the funding is there, PacBio is the recommended option for any genome.

Assembly programs

In particular when working with Illumina data, it is important to understand that the choice of assembly program will greatly influence the data you need to order. So you should choose assembly program before ordering your data. One example is Allpaths-LG, that needs a library with short insert-size (overlapping reads) and one mate-pair library to even start. Please keep this in mind and discuss your assembly strategy in detail with the sequence provider.

Assembly recipes

Here follows a few recipes for genome assembly based on genome size. PE=Paired End, MP=Mate Pair. These are the current recommendations, but please note that this can change quickly. Please contact the sequencing platforms and/or the NBIS assembly and annotation team for up to date advice.

1 single bacterial genome - 1-2 PacBio cells

This will often result in a closed bacterial genome, for a low cost.

Population of bacteria (>12) - PacBio for the reference genome, rest with Illumina.

Eukaryote < 50 Mbp - PacBio only

Eukaryote 50-500 Mbp - Illumina only solution

- 2x150 bp PE overlapping reads, 50-90x coverage
- 2x150 bp MP, insert-size 3 kbp, 50x coverage
- + more MP libraries of larger insert size if possible

This recipe is based on the assembly program Allpaths-LG, which given this type of data generally does the best job possible.

Eukaryote > 500 Mbp

- 2x150 bp PE overlapping reads, 50x coverage
- 2x150 bp PE, insert size 650 bp, 50x coverage
- 2x150 bp MP, insert-size 3 kbp, 5-15x coverage
- 2x150 bp MP, insert-size 8 kbp, 5-15x coverage
- + 20 Kb insert mate-pairs if possible, at typically very low coverage (ie. 1x)

This recipe is also based on the assembly program Allpaths-LG, but more data are needed due to the larger genome size.

Eukaryote > 50 Mbp - PacBio only solution

- 20 kb insert libraries (number of libraries depends on the genome size)
- aim for 30x coverage with reads over 10 kb (overall coverage of 50-70x per haploid genome)

This recipe is based on the assembly program HGAP for smaller genomes and FALCON for large diploid genomes.

Version history

Version 1.2 - 2016-04-27
Version 1.1 - 2016-01-08
Version 1.0 - Aug 2015 by Henrik Lantz, NBIS assembly and annotation team. Continued development by the SciLifeLab De Novo Application group.

Contact person

Henrik Lantz (henrik.lantz@bils.se)