Journal of Microbiological Methods 86 (2011) 368–375

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

A biologist's guide to *de novo* genome assembly using next-generation sequence data: A test with fungal genomes

Sajeet Haridas [a], Colette Breuill [a], Joerg Bohlmann [b], Tom Hsiang [c,*]

[a] Faculty of Forestry, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4
[b] Michael Smith Laboratories, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4
[c] School of Environmental Sciences, University of Guelph, Guelph, ON, Canada N1G 2W1

ARTICLE INFO

Article history:
Received 20 April 2011
Received in revised form 27 June 2011
Accepted 27 June 2011
Available online 3 July 2011

Keywords:
Next generation sequencing
Genome
Fungi
Assembly
Software
Illumina

ABSTRACT

We offer a guide to ***de novo* genome assembly**[1] using sequence data generated by the **Illumina** platform for biologists working with fungi or other organisms whose genomes are less than 100 Mb in size. The guide requires no familiarity with sequencing assembly technology or associated computer programs. It defines commonly used terms in genome sequencing and assembly; provides examples of assembling short-**read** genome sequence data for four strains of the fungus *Grosmannia clavigera* using four assembly programs; gives examples of protocols and software; and presents a commented flowchart that extends from DNA preparation for submission to a sequencing center, through to processing and assembly of the raw sequence reads using freely available operating systems and software.

1. Introduction

There has been a proliferation of new sequencing technologies and assembly software, which has made it feasible for many biological research labs to now attempt sequencing and assembly of small-sized eukaryotic genomes using their own resources (e.g. Ellwood et al. 2010). Many very good review and comparative articles have been written on technical aspects of sequencing as well as the assembly process (Claesson et al., 2010; Ma and Fedorova, 2010; Metzker, 2010; Miller et al., 2010), but these generally assume a level of understanding of bioinformatics and technical terms beyond that of many biologists. In this article, we offer a guide to the genome sequencing and assembly process that is written for biologists, and illustrate some key issues with experimental data from genome assembly of four filamentous fungal genomes.

More than 10 years after the start of the project, the first complete draft of the human genome was produced using conventional **Sanger sequencing**, and made available to the public in 2001 (International Human Genome Sequencing Consortium, 2001). In the decade that followed, there have been incredible advances in **high-throughput** sequencing technologies. These next generation sequencing (**NGS**) technologies (Medini et al., 2008) can produce massive amounts of data (ultra high throughput) in less time and at a lower cost, but with shorter sequence reads than conventional automated Sanger sequencing (Sanger et al., 1977). Each base position is sequenced many times on average to achieve deeper **coverage**, hence another nickname, "Deep Sequencing".

Three major sequencing platforms and processes have been developed that do not rely on the traditional Sanger Chain Termination method. These include 454 sequencing which involves DNA capture beads (www.454.com/products-solutions/how-it-works/index.asp), SOLiD™ System (www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html) which uses "microfluidic FlowChips", and Illumina Sequencing (www.illumina.com/technology/sequencing_technology.ilmn) which uses a "reversible terminator-based method". Solexa Inc. developed this last method, but the company was purchased by Illumina Inc. in 2007, and so this platform is sometimes referred to as Illumina-Solexa.

These platforms produce short reads, currently ranging in size from approximately 50 to 400 nucleotides depending on the platform and options selected. Software programs have been developed that can handle the large numbers of small reads, but require high levels of computational power. For example, the software program SOAPdenovo on a supercomputer with 32 cores and 512 GB of RAM was used to assemble the panda (*Ailuropoda melanoleuca*) genome using 56× coverage and reads with an average length of 52 bp generated by Illumina (Li et al., 2008). The presence of interspersed repeats larger than sequencing reads can confound assembly algorithms (Miller et al., 2010). Thus, various strategies have been adopted to take advantage of the strengths of different sequencing technologies. For example, DiGuistini et al. (2009) assembled a fungal genome *de novo* by combining reads from Sanger, 454 and Illumina platforms. The

* Corresponding author. Tel.: +1 519 824 4120; fax: +1 519 837 0442.
E-mail address: thsiang@uoguelph.ca (T. Hsiang).

[1] Items in bold at their first mention are defined in the glossary (Appendix A).

0167-7012/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2011.06.019

longer but fewer reads from Sanger and 454 sequencing combined with the depth of coverage from Illumina were used to address the issue of repeats. In addition, sequencing short reads of the ends of target-sized insert fragments (**paired-end sequencing**) can be used to bridge data over regions that are difficult to sequence.

Beyond the growth of the new technologies, there has been an even greater proliferation of new words and jargon associated with these technologies. There are detailed and informative articles on these new technologies, but they are sometimes written at a level that is too technical for biologists who wish to understand or use these technologies. The purpose of this article is to: (1) serve as a guide to the NGS technologies, particularly the Illumina sequencing platform; (2) define commonly used terms in genome sequencing and assembly; (3) provide examples of assembly of short-read genome sequence data for four strains of the fungus *Grosmannia clavigera* using four assembly programs; and (4) present a commented flowchart that starts from DNA preparation for submission to a sequencing center, through processing and assembly of the raw sequence reads using freely available operating systems and software.

This guide should be useful to biologists, particularly those working with fungi or other organisms whose genomes are less than 100 Mb in size, with no presumption of familiarity with the use of sequencing assembly technology or associated computer programs for understanding this article. However, the genome assembly process does require some basic skills in computational biology, particularly the use of the **Linux** operating system, and the use of **PERL** or other basic scripting language. Examples of particular protocols and particular computer programs are given without any implication that these are the optimal methods for genome assembly. We show comparisons of four freely available tools for *de novo* genome assembly from short reads, as well as assembly of data trimmed for read quality.

2. Materials and methods

2.1. Fungal species, DNA extraction and sequencing

This study considered four isolates of the fungus *G. clavigera*. We used a simple CTAB (cetyltrimethylammoniumbromide) DNA extraction protocol as described previously (DiGuistini et al. 2009). The DNA was not further processed before being sent to Canada's Michael Smith Genome Sciences Centre (GSC, Vancouver, Canada) for 75 bp reads in paired-end sequencing with a 200 bp insert size using an Illumina GAIIX. A single **sequencing lane** was specified for samples GC1, GC2 and GC4 while two lanes were obtained for GC3, to assess whether two lanes could improve the assembly.

2.2. Downloading and processing the sequence data

After sequencing was completed, the sequencing center uploaded the data to their FTP (file transfer protocol) website. The data were transferred using the Linux program SCP (secure copy) to a local high-performance computing cluster, known as WESTGRID (www.westgrid.ca), which is free for academic users. Some data were also transferred to a Microsoft Windows PC-based system using a Windows-based free data transfer program called WinSCP (winscp.net). Graphical assessment of the quality of read data was conducted using the software program FASTQC (www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), which can be run on a variety of operating systems with Java Runtime installed (java.com).

Trimming of reads based on read quality was done using the program Dynamictrim included in SolexaQA (Cox et al. 2010) which is "a Perl-based software package that calculates quality statistics and creates visual representations of data quality from **FASTQ** files generated by Illumina second-generation sequencing technology" (solexaqa.sourceforge.net). Assembly of the data was done using four different Linux-based *de novo* assemblers: Abyss (ver. 1.2.3, www.bcgsc.ca/platform/bioinfo/software/abyss; Simpson et al., 2009), SOAPdenovo (ver. 1.1.2, soap.genomics.org.cn/soapdenovo.html; Li et al., 2008), Velvet (ver 1.0.12, www.ebi.ac.uk/~zerbino/velvet; Zerbino and Birney, 2008), and IDBA (ver. 0.17, i.cs.hku.hk/~alse/hkubrg/projects/idba; Peng et al., 2010), using a broad range of **k-mer** values (from 27 up to 64). In general, the default selections for computer programs were selected and illustrated here, unless otherwise stated. The output from the program SOAPdenovo was further processed using GapCloser (soap.genomics.org.cn) which closes gaps (joins **contigs** into **scaffolds**) and provides better assemblies.

Assembled contigs that were smaller than 200 bp were removed (using PERL script remove_small_contigs.pl available at www.uoguelph.ca/~thsiang/pubs/supplement) before assembly statistics were calculated (using PERL script assemblystats.pl at the same website which gives file size, **assembly size**, number of contigs, **N50** and **N90**). The genomes of four isolates of *G. clavigera* were assembled (GC1, GC2, GC3 and GC4). The results for one of the genomes, GC1, are presented in the main text, and the results for the other three are available in supplementary data.

2.3. Estimating coverage (depth of sequencing) needed for assembly

To investigate how much data is needed for assembly, the dataset for one of the genomes (GC1) was divided into 10 portions, by systemically sampling every tenth sequence in the ~30 million read dataset. We used each of the resulting files which represented 10% of the total reads to incrementally add as input to the assembly programs Abyss, SOAPdenovo and Velvet using k-mer 39. Thus, ten assemblies representing 10%, 20%, 30% and so on up to 100% of the reads from one lane of sequence reads were produced for each of these three assembly programs.

2.4. Single-end vs. paired-end data

Since paired-end sequencing is thought to provide an additional layer of information that can improve the quality of assembly, we also tested paired-end data as single-end reads for assembly using the dataset of GC1. The ~30 million paired-end reads were treated as 60 million single-end reads, and we compared the assembly statistics of paired-end vs. single-end assemblies at various k-mer lengths in the different programs.

3. Results and discussion

3.1. Data download and quality of reads

The sequence data became available for download within three to four months after initial submission of the DNA samples. The data were provided in compressed **QSEQ** format. After downloading the files, they were converted to FASTQ format using PERL script qseq2fastq.pl available at www.uoguelph.ca/~thsiang/pubs/supplement. Each lane of sequencing provided between 25 and 35 million reads, of which 85%–90% passed a basic quality check using a **chastity filter** (Table 1). The quality of the reads was further assessed with the program FASTQC. All four samples showed consistent read quality that decreased steadily over the length of the reads (Suppl. Fig. S1a–d). This is normal for the Illumina technology according to the manufacturer (www.illumina.com/systems/genome_analyzer_iix.ilmn) where more than 85% of the bases achieve a threshold quality. The reads were used directly without further filtering for *de novo* assembly, or subjected to trimming (see Section 3.3) in attempts to improve the assembly.

Table 1
The number of reads in the raw QSEQ files obtained from the sequencing center and the number of reads from each lane that passed the chastity filter (quality check) for four genomes of the fungus *Grosmannia clavigera* with two lanes for GC3.

Genome	Reads	Reads after chastity filtering	Failed reads (%)
GC1	33,385,453	29,428,934	11.9
GC2	33,834,154	30,033,878	11.2
GC3.a	33,250,787	28,004,056	15.8
GC3.b	34,219,918	29,208,325	14.7
GC4	25,340,644	22,594,526	10.8

3.2. Assembly size, quality and coverage

Although previous work has assembled 30 Mb of the estimated 35 Mbp genome of *G. clavigera* combining several sequencing technologies (DiGuistini et al., 2009; 2011), we did not use these assembled data as a reference genome during the assembly process, which is referred to as **resequencing**. In resequencing, the sequenced genome can be used as a reference or scaffold for the assembly of subsequent genomes of the same species, and is a simpler process than *de novo* assembly. But for heuristic purposes, we only conducted *de novo* assembly of the newly sequenced data here.

The resulting assembly sizes across a wide range of k-mer values and using three assembly programs (Abyss, SOAPdenovo and Velvet) were between 27 Mbp and 30 Mbp for GC1 (Fig. 1) and the other three genomes (Suppl. Fig. S2a–c). The exception was for assemblies with k-mer <30 generated by SOAPdenovo, which were highly fragmented; during quality checking, a large number of contigs smaller than 200 bp were removed which resulted in much smaller assembly sizes (Fig. 1, Suppl. Fig. S2a–c). As expected, very small and very large k-mer values produced fragmented assemblies as shown by the N50 values (Fig. 2, Suppl. Fig. S3a–c), because of the lack of specificity (smaller k-mer) or the lack of sensitivity (larger k-mer). A fourth assembly program, IDBA, which does not require the user to specify a single k-mer for input, also produced similar assembly sizes (Table 2).

We used subsets of the read data to assess the effects of the number of reads on the resulting assembly quality using the criteria of assembly size and N50 values. Based on single lanes of 75 bp paired-end data, the overall assembly size did not increase after 30% of the reads (~10 million reads) were used for assembly (Suppl. Fig. S4), and quality of assembly as indicated by N50 values did not improve beyond 15 million reads (Fig. 3). Additionally, the use of two lanes of data for GC3 did not improve assembly statistics when compared to single lane assemblies for GC2 or GC4 (Suppl. Figs. S2a–c and S3a–c), and N50 was significantly lower with two lanes of data. Perhaps a saturation point is reached with increasing data, such that the errors in DNA sequencing inherent in this technology inhibit further improvements in assembly.

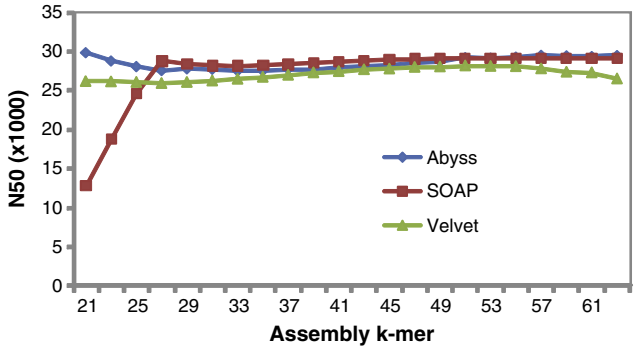

Fig. 1. Size of the assemblies for genome GC1 produced by the programs Abyss, SOAPdenovo and Velvet using various k-mer values.

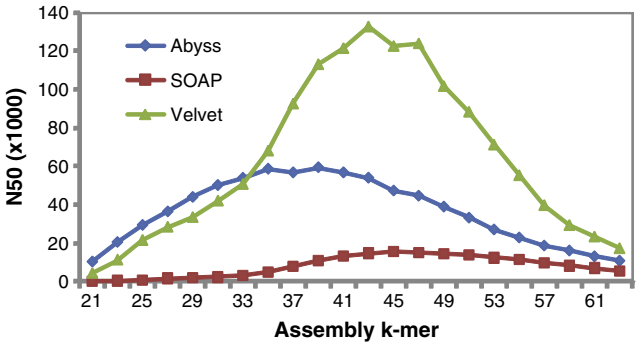

Fig. 2. N50 values for genome GC1 produced by the programs Abyss, SOAPdenovo and Velvet using various k-mer values.

The 25 to 35 million reads that we obtained from each lane of paired-end sequencing would theoretically provide approximately 125 to 175 fold coverage for a 30 Mb genome. However, the median coverage was much lower, and was approximately 86 fold for GC1. This discrepancy is possibly due to: (1) repetitive regions in the genome that were not properly assembled with this technology; (2) organellar DNA that was not removed from the sample before sequencing; (3) reads that could not be assembled; (4) and reads that were not all full length. Organellar DNA and genomic repeat structures tend to be over-represented during sequencing and are often misassembled (Haridas, 2010), and in GC1, the average coverage for these was ~2000 fold.

The coverage across the assembled contigs using Velvet k-mer 43 for GC1 is shown in Fig. 4. In this case, a small number of contigs that represent organellar genomes and nuclear genomic repeats were assembled from a large number of reads. These high coverage contigs and the few contigs with very low coverage (i.e. due to contamination, bad reads or poor assembly) should be assessed first and removed as required. A previously published mitochondrial genome (Haridas and Gantt, 2010) was assembled from a GS-FLX sequencing run where ~12% of reads were from the mitochondrial genome. Organellar DNA can be removed from total DNA before sequencing using a cesium chloride (CsCl) gradient, and this has been successfully done in 454 sequencing by Haridas (2010). Using the CsCl gradient to eliminate mitochondrial DNA can provide additional depth of coverage when using 454 sequencing; however, the incremental advantage for Illumina sequencing will likely be negligible due to the much higher number of reads produced.

Table 2
Top-ranking assemblies obtained using Abyss, SOAPdenovo, Velvet and IDBA for *Grosmannia clavigera* genomes GC1, GC2, GC3 and GC4 based on single lanes of 75 bp paired-end reads.

Genome	Program	K-mer	Assembly size (bp)	Contigs	N50 (bp)	N90 (bp)
GC1	Abyss	39	27,735,442	3043	59,299	11,978
	SOAPdenovo	45	28,955,798	5837	15,504	2681
	Velvet	43	27,687,316	3616	132,842	18,547
	IDBA	27 to 64	28,963,689	1008	87,605	19,217
GC2	Abyss	37	27,595,753	2655	60,922	13,338
	SOAPdenovo	45	28,838,441	5877	14,407	2684
	Velvet	41	27,466,615	3343	123,452	17,498
	IDBA	27 to 64	28,902,903	1025	79,035	20,135
GC3	Abyss	39	28,450,218	2964	72,387	13,631
	SOAPdenovo	49	29,277,755	6545	15,336	2376
	Velvet	45	27,052,268	4630	65,030	7573
	IDBA	27 to 64	29,351,153	7138	24,483	2658
GC4	Abyss	39	28,189,576	3093	85,192	15,437
	SOAPdenovo	47	29,085,569	4301	26,336	3976
	Velvet	43	28,768,246	1241	236,827	23,574
	IDBA	27 to 64	28,952,419	984	103,412	20,192

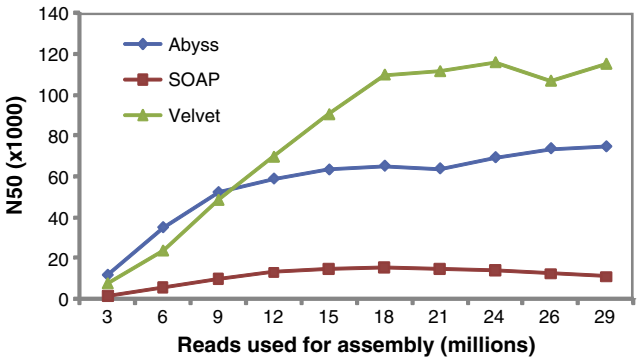

Fig. 3. N50 values for assemblies of GC1 produced by the programs Abyss, SOAPdenovo and Velvet using a subset of reads from one lane of 75 bp paired-end reads of Illumina data at k-mer = 39.

3.3. Effect of data trimming

Since quality of the last 10 to 15 bases of each 75 bp read was generally poor (Suppl. Fig. 1a), we trimmed the reads to 50 bp, either in the input file (for Abyss and Velvet), or we specified maximum input read length as 50 (for SOAPdenovo). While the resulting assembly sizes did not change significantly (Suppl. Fig. S5), the N50 did increase, with the largest gain seen in the Velvet assembly from 130 kb for GC1 (Fig. 2) to almost 180 kb after trimming (Fig. 5). Surprisingly, trimming based on per base quality using Dynamictrim did not improve assembly statistics (Fig. 6, Suppl. Fig. S6) in direct contrast to the experiences of Cox et al. (2010). However, in both cases, providing higher quality data (by trimming off poor data) provided higher specificity, and the programs were able to assemble larger contigs using smaller k-mer values. Due to the high depth of genome coverage in our Illumina sequencing runs, removal of this data did not significantly affect final genome assembly size (Suppl. Figs. S2a, S5 and S6).

3.4. Use of single end sequencing

When ~30 million 75 bp paired-end reads were treated as ~60 million 75 bp single-end reads, the assemblies were much more fragmented and the N50 values were significantly lower than obtained with paired-end data (Fig. 7, Suppl. Fig. S7). For example, GC1 had a max N50 of 24 kb when treated as single end data using Velvet (Fig. 6) in contrast to a max N50 of 133 kb (Fig. 2) when the same data were treated as paired-end data. With Abyss, the decrease was lower with 27 kb as single end (Fig. 6) vs. 59 kb as paired-end.

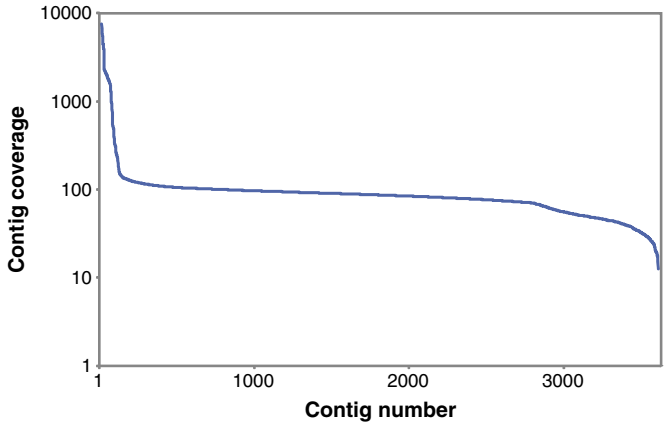

Fig. 4. The 3617 contigs of GC1 produced by Velvet using k-mer 43 arranged from highest coverage (most reads associated) to lowest coverage.

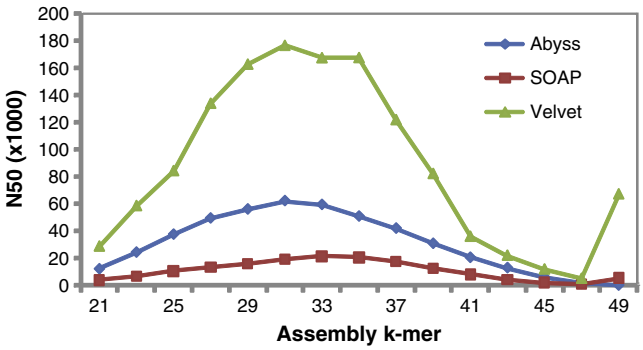

Fig. 5. N50 for assemblies of GC1 produced by Abyss, SOAPdenovo and Velvet using Illumina reads trimmed to 50 bases from 75.

With SOAPdenovo, the difference was the smallest with 14 kb as single end (Fig. 6) vs. 15 kb as paired end.

4. Recommendations

A flowchart summarizing the general steps in genome sequencing and assembly is presented, with further details in the following respective sections (Fig. 8).

4.1. Preparing DNA for submission

Each sequencing center will have guidelines for preparation of DNA. Usually, DNA can be extracted with a variety of methods or commercial kits, and the short read sequencing process does not depend on extremely high quality and long length strands of DNA, except with large insert sizes that are required for **mate pair** sequencing. The sequencing center may request between 1 μg and 10 μg of genomic DNA, which has been checked for purity (OD260/280 from 1.7 to 2.0) and with minimal degradation (i.e. no extensive smearing in the gel lanes).

4.2. Depth of coverage and library fragment size

Illumina technology (HiSeq) can yield 80 to 100 million 100 bp reads from a single lane, equivalent to an empirically calculated >300 fold coverage for a 50 Mbp genome using paired-end sequencing. Even with a 50% coverage loss due to poor quality reads, unassembled reads and the over-sequencing of organelle DNA and genomic repeat structures, the amount of sequence data should still provide sufficient data for an attempt at *de novo* assembly.

The assembly statistics did not improve by providing more than 15 million reads (40× median coverage, 75× theoretical coverage) or by doubling the number of reads (Suppl. Figs. S2b and S3b). This suggests that increased genome coverage cannot be obtained by

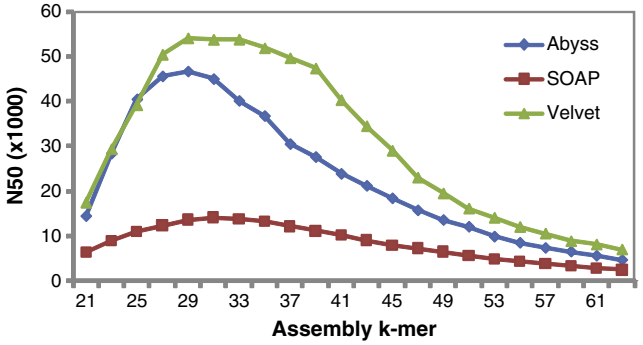

Fig. 6. N50 for assemblies of GC1 produced by Abyss, SOAPdenovo and Velvet using Illumina reads trimmed for quality using Dynamictrim.

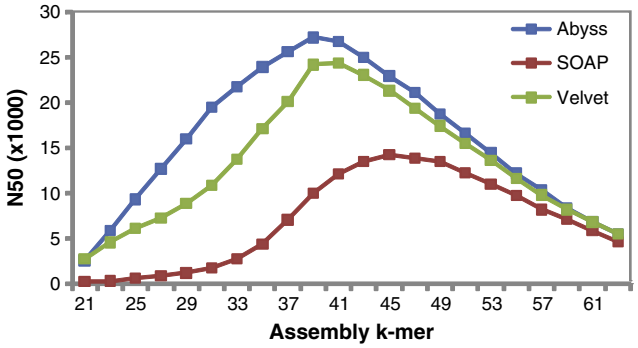

Fig. 7. N50 for assemblies of GC1 produced by Abyss, SOAPdenovo and Velvet using Illumina reads as single end data.

increasing read depth from a single sequencing technology. A better strategy would be to obtain additional information by using other fragment sizes (i.e. paired-end sequencing of another library with a different fragment size), and perhaps to combine different technologies (e.g. Illumina and 454 sequencing) to benefit from their respective strengths. Gnerre et al. (2011) of the Broad Institute, which is one of the major contributors to sequencing of fungal genomes, recommend a strategy using Illumina reads from two libraries: one from 180 bp fragments and another from 3000 bp fragments at around 45× coverage.

The methods presented here are also applicable to resequencing, where a genome has already been sequenced, and the assembly of the first genome is used as a scaffold for assembly or comparison of the subsequent genomes. For resequencing, the coverage needed is much less (10×–20× for Illumina data) than for *de novo* sequencing.

4.3. Hardware and software useful for genome assembly

A minimum of 8 GB of RAM (random access memory) for a desktop computer is needed for assembly of an average sized fungal genome (50 Mb) based on a single lane of 75 bp paired-end data. The amount of memory is dependent on the computer program used and >16 GB is recommended. Rather than a standalone desktop unit suitable for genome assembly (currently less than $1000 U.S.), we suggest the use of high performance computer clusters or grids which are often freely available or at a nominal cost to University and Government researchers. Examples are SHARCNET and WESTGRID in Canada, HECToR in the U.K., TeraGrid in the United States, and PRACE and DEISA in Europe.

There are commercially available programs such as CLC Genomics (www.clcbio.com) which has versions for Windows, Linux and Mac OS X, but most freely available software programs designed for assembly of very large amounts of sequence data are written for the Linux operating system. For Microsoft Windows-PC users, a relative easy and quick entry into the Linux world is the use of WUBI (wubi.sourceforge.net), which is a program that sets up Linux on the hard drive on a Windows-based computer without the necessity of creating a separate Linux partition. Another way of installing Linux on a Windows computer is through the use of virtual machines. These allow you to run multiple operating systems on a single PC simultaneously. An easy-to-use free tool to set up virtual machines called Virtualbox is provided by Sun Microsystems (www.virtualbox.org). In such a setup, resources are shared between the host and guest system and therefore the guest system will only be able to use a part of the available hardware resources including the RAM.

4.4. Handling the raw data

A single lane of genome sequencing data (Illumina GAIIX paired-end 75 bp) can take up ~2 GB file size in bz2 compressed format. When uncompressed to text format, this file can exceed 7 GB in size. Several easy-to-use Microsoft Windows-based programs such as WS-FTP or WinSCP are useful for downloading the data, and for Windows operating systems, 64-bit versions are required to handle such large files. These data are often provided as FASTQ files or QSEQ files (Suppl. Figs. S8 and S9), and may be compressed. Regular editing programs such as Word Processors are unable to handle such large files. Scripts

4.1) Prepare DNA for submission
- regular DNA extraction methods are sufficient

⇓

4.2) Select a depth of coverage and specify library fragment size
- a single lane of 100 bp reads may be sufficient
- typical insert sizes are 200 bp to 10 kb
- give these instructions to sequencing center which creates the library

⇓

4.3) Set up or obtain access to computing facilities
- set up Linux on a PC with 16 Gb RAM or
- obtain access to a High Performance Computing Cluster

⇓

Generation of sequence data at sequencing center

⇓

4.4) Download and convert the raw sequence data
- use software to download the gigabytes of data
- convert the data format for processing

⇓

4.5) Choose assembly program
- choose between several (freely) available software programs (Linux based)
- decided whether to trim data for quality

⇓

4.6) Choose k-mer for assembly
- lower values have higher sensitivity, while higher values have higher specificity
- scripts can be used to produce a wide range of k-mer assemblies (examples given)

⇓

4.7) Accept that your Illumina assembly is incomplete
- current technological limitations cannot overcome issues of repetitive regions

Fig. 8. A flowchart summarizing the general steps in genome assembly of Illumina data.

written in computer program languages such as PERL are useful for modifying the data files, for example to convert FASTQ to FASTA format, or to trim by quality score. A variety of PERL scripts for genome assembly work are available at www.uoguelph.ca/~thsiang/pubs/supplement.

4.5. Data analysis, trimming and assembly

We found no advantage in trimming data based on quality of **base calls** (individual nucleotides in read sequences) using Abyss or SOAPdenovo. However, there were improvements to assembly statistics using trimmed data with Velvet. Of the four programs used in this study, SOAPdenovo is the only one available as a precompiled binary and may be the simplest to use for a new Unix/Linux user, but Abyss, Velvet and IBDA generally provided better assemblies in our tests. These genome assembly programs are frequently upgraded, and new programs are constantly being produced. The choice of assembly program will depend on the hardware limitations of the user, since genome size and amount of data will influence the amount of memory required to process the data. For each program, there are user groups (listserver.ebi.ac.uk/pipermail/velvet-users, groups.google.com/group/bgi-soap, groups.google.com/group/abyss-users, groups.google.com/group/hku-idba) who discuss issues with each program. The developers are usually involved in answering the questions.

In terms of quality of assembly by the different programs, Velvet at its best performing k-mer value gave the largest N50 values for three of the four genomes, followed by IDBA, Abyss and then SOAPdenovo. All four programs provided very similar assembly sizes indicating that all four programs are capable of uncovering the appropriate unique set of data; but because the N50 values differ, this demonstrates that they are not equally good at fitting all the sequence reads together. Another consideration is that while reads might be joined together, they might actually not belong together. This would need to be analyzed by comparison to a reference genome. Preliminary comparisons of our *de novo* assemblies to the published genome of *G. clavigera* (DiGuistini et al., 2009) suggest that the assemblies produced by Abyss were more consistent with the published data. Additionally, Abyss also performed better than Velvet when they were provided larger datasets in GC3 (Table 2, Suppl. Fig. S3b). However, the assemblies produced by Velvet were superior in certain cases especially for GC4.

4.6. Choice of k-mer

The choice of k-mer is considered to be a tradeoff between specificity and sensitivity. Longer k-mers will provide higher specificity, but may not assemble all the available data. A smaller k-mer is more sensitive and may allow joining of more reads, but may not be able to resolve spurious overlaps. As the quality of input data is improved (as shown by trimming in this paper), a smaller k-mer can be used to assemble the data effectively. While the assembled genome size does not change significantly over a range of k-mers, suggesting that most of the genome has been assembled at all these k-mers, the sizes of contigs (as indicated by N50) are largest across a much smaller k-mer range. The choice of k-mer is program and data specific and several iterations may be required to find the optimum k-mer for a particular dataset. In our studies with 75 bp reads, we used k-mers for Abyss, Velvet and SOAPdenovo ranging from a low of 21 to 63 which was the maximum accepted by the version of SOAPdenovo we used. The program IDBA requests a minimum and a maximum k-mer, and computes over this range.

4.7. Genome composition may hinder genome assembly

Genome assembly is complicated by several factors. The most basic of these are the presence of small repeat clusters (<500 bp), small tandem duplications (1000 bp) and large tandem duplicated clusters and repeats (>2000 bp). Some of these can be resolved by using longer read lengths such as those provided by 454 and Sanger sequencing. Others require the use of large insert paired-end sequencing or mate pair sequencing to bridge across the expected repeat size. Multiple insert size libraries will be required to bridge across small and large duplications and repeat clusters, which cannot be assembled using whole genome sequencing techniques. Additional factors that complicate genome assembly such as allelic differences and paralogous genes are more difficult to address.

Haploidy and clonality, which are common states for many fungi, provide ideal materials for genome sequencing, because in these states, problems with allelic differences, gene duplications, and variations in populations are avoided. If insufficient DNA is available from a single organism because of technical hurdles in clonal propagation, then multiple genotypes may be required to provide sufficient DNA. This may cause problems in assembly since single base pair differences may prevent joining of sequences, and these genome composition issues together increase the difficult of genome assembly.

4.8. Time and cost considerations

In December 2010, the costs associated with genome sequencing based on quotes and prices available on the web were as follows:

(1) Library construction, $800–$1000 per library from genomic DNA
(2) Paired-end sequencing $2500–$3500 per lane for 75 to 100 bp read lengths.
(3) A single lane which can provide over 60 million single reads can be subdivided through the use of tags or barcodes to individual libraries (also known as indexing), such that more than one library (e.g. genome) can be run in a single lane, but library preparation and sequencing costs are slightly higher.

The production time from DNA submission until receiving the raw sequence reads between 2009 and 2011 has varied between two to five months, based on several genome sequencing projects using different sequencing companies. The exponential rate of advance of technologies (sequencing and software) may mean affordable Next-Gen sequencers and packaged bioinformatic software within few years to produce a complete fungal genomic sequence in your own lab. But for now, we offer these suggestions depending on the budget available for *de novo* sequencing of small eukaryotic genomes.

Acknowledgments

This work was partially funded by grants from Genome Canada, Genome BC and Genome Alberta (grant to JB and CB) in support of the Tria Project (http://www.thetriaproject.ca), and by funds from NSERC (grant to TH). Our thanks to G. Robertson, Y. Wang, L. Lah, J. McLaughlin, L. Jewell, P. Goodwin, and I. Birol for helpful comments on the manuscript.

Appendix A. Glossary

Assembly size — The number of nucleotides used in contigs or scaffolds assembled from reads.

Base calls — The nucleotide (A,T,C or G) that is "called" or determined for each position of a read based on the raw sequencing data output. The probability that a base is called incorrectly, known as the quality score, is associated with each base call. These data, both the base calls and quality scores, are provided in FASTQ or QSEQ files.

Chastity filtering — The sequencing machine detects signals for each type of base (A,T,C,G) at each read cycle. It then compares the intensity of the signals relative to each other to assess whether the

base can be accurately assigned. Chastity for a given base call is defined as "the ratio of the highest of the four (base type) intensities to the sum of highest two". For Illumina data, the default pipeline quality filter has a threshold of CHASTITY >= 0.6. This filter is used to identify clusters with a low signal to noise ratio, often as a result of two adjacent clusters being so physically close together that their signals cannot be measured independently (illumina.ucr.edu/ht/documentation/file-formats-old). This type of filtering can be used to screen out sequences which do not pass the threshold (implemented within some assembly programs), or by custom PERL scripts written for this purpose, such as qseq2fastq.pl or qseq2fasta.pl available at www.uoguelph.ca/~thsiang/pubs/supplement. Note that chastity filtering is also called purity filtering, or purity.

Contig — A contig (from the word contiguous) is a set of overlapping joined DNA segments (see scaffold). The largest possible size of a single contig (often referred to as a supercontig) is a chromosome.

Coverage, Genome Coverage, Depth of Coverage — The average number of times each base position is sequenced. This can be estimated empirically by the (number of reads × read length)/(genome size). For example, a single lane of 75 bp paired-end data in the Illumina GAIIX platform may produce 40 million reads. For a genome of 50 Mb size, this is calculated as (40,000,000 × 75)/(50,000,000) = 60× coverage or 60-fold coverage.

De novo **assembly** — Sequencing and assembly of a genome without prior assembly information for the same or closely related species (see resequencing).

Deep sequencing — See next generation sequencing.

DNA sequencing — The process of determining the nucleotide order in a DNA fragment.

FASTQ — A data format for output from sequencing platforms, which includes sequence data and quality score for each position. An example is given in Suppl. Fig. S8.

Fragment size — See insert.

High throughput sequencing — Compared to the standard Sanger sequencing method and manual sequencing by electrophoretic gels, the newer sequencing technologies are able to produce millions of sequences at a time.

Illumina sequencing — Sequencing using the Illumina/Solexa platform where DNA molecules are first attached to primers on a slide and amplified to form local clonal colonies. ddNTPs are added and non-incorporated nucleotides are washed away. The DNA is extended one nucleotide at a time and a camera takes images of the fluorescently labeled nucleotides. The dye along with the terminal 3′ blocker is chemically removed from the DNA, allowing a next cycle.

Illumina sequencers — Sequencing machines and systems produced by Illumina, including: HiSeq, GAIIx, GAIIe or iScan-Seq. These offer the same applications, biochemistry and data, but differ in output.

Indexing — Single lanes can be subdivided to sequence different libraries using short sequence tags that are adapted to DNA fragments before sequencing. This allows the output to be sorted into different sequence pools for processing.

Insert — In paired-end sequencing and mate pair sequencing, the ends of a fragment are sequenced, and depending on the size of the fragments, these ends allow the sequence to bridge duplications and repeat clusters.

K-mer — In genome assembly, this refers to the number of perfectly matching adjacent nucleotides among reads that are needed to make contigs (see Section 4.6 for choice of k-mer).

Linux — An open-source version of the Unix operating system, created by Linux Torvalds. Many freely available software titles for DNA analysis particularly sequence assembly are available for the Linux platform.

N50 — A median weighted statistic often used to assess assembly quality. N50 is calculated by sorting all contigs from smallest to largest and determining the contig size at which 50% of all bases in the assembly are contained in contigs or scaffolds larger and smaller than this value. A PERL script written to calculate this and other assembly statistics (assemblystats.pl) is available at www.uoguelph.ca/~thsiang/pubs/supplement.

N90 — A weighted median statistic such that 90% of the entire assembly (in terms of number of base pairs) is contained in contigs or scaffolds equal to this value expressed as base pairs.

Next Generation Sequencing (NGS) — The high throughput sequencing platforms mostly developed in the 2000s such as 454, Solid and Illumina are referred to as "Next Generation" or "Second Generation" sequencing technologies. "Third Generation" sequencing technologies are currently under development.

Paired-end sequencing — Both ends of fragments ranging in size from 200 bp to 5 kb are sequenced, to provide positional information, and to detect structural variations such as chromosomal rearrangements, copy number variation, and indels. For paired-end sequencing with large inserts, particularly to bypass hard-to-sequence regions and to join contigs into larger scaffolds, these are referred to as mate pair sequencing, and require higher quality less fragmented DNA. (www.illumina.com/systems/genome_analyzer/paired_end_module.ilmn).

PERL — A computer scripting language (practical extraction and reporting language) that was originally developed for manipulating/extracting text, is now commonly used in a large variety of computer related tasks, including bioinformatics. The commands are written in text file, which is executed by the PERL program. PERL is freely available on a wide variety of operating systems, and is easily installed. The scripts (such as examples given here) are mostly interchangeable between operating systems.

Phred Score — Phred is a base calling program that uses output from a sequencing machine to give the base and a score reflecting the quality of the call and the probability that the base is called incorrectly. A Phred score of 10 means a 10% chance of an incorrect base call, while a Phred score of 50 means a 1 in 10^5 chance of an incorrect base call. This analysis is usually done by the sequencing center, and the data provided in the form of FASTQ or QSEQ files to the user.

Quality, Base Calls — "Quality scores measure the probability that a base is called incorrectly". www.illumina.com/truseq/quality_101/quality_scores.ilmn. These scores are represented by single symbols in QSEQ and FASTQ files, and since there are more score types than 26 letters and 10 single digits, other assorted characters such as # are also used. See base calls and Suppl. Figs. S8 and S9.

Quality, Assembly Quality — The quality of a sequence assembly is often judged by two major criteria: N50 and assembly size.

QSEQ — A data format for output from sequencing platforms, which includes sequence data and quality score for each position. An example is given in Suppl. Fig. S9.

Read — See raw read.

Raw Read — The nucleotide sequence produced by a sequencing instrument. Read sizes have been steadily increasing for Illumina machines, from 35, 50, or 75 bp to 110 bp more recently.

Resequencing — Sequencing of a genome for which there is already an assembly (same or very similar species). This may require less sequencing information since scaffolds and predicted genes may already be available. This is done to examine intraspecific variation and for use in population genetics. (See *de novo* assembly).

Sanger Sequencing — Also known as Chain Termination Sequencing, or Dideoxy Sequencing, this technique involves the use of DNA polymerase to synthesize DNA fragments of varying lengths in four different reactions (A,T,C,G) from the same start point, and assessing the terminal base for each fragment size to piece together the sequence order.

Scaffold — In genomic mapping, a scaffold is a series of contigs that are in the right order but not necessarily connected in one continuous sequence (see contig), with N's connecting the sequences. The output from assembly programs may contain N's since they use paired-end

data to position the reads and contigs, but in between sequences may be absent.

Sequence Read — See raw read.

Sequencing lane — Sequencing using the Sanger method was originally done by running the four reactions out in gel lanes, and assessing the identity of each base position based on labeling of nucleotides. The terminology of "lane" is still used whether these are now "virtual lanes", generated within computers, or whether the lanes are actually plates or flowcells, and so on.

Sequencing plate — See sequencing lane.

Tags — See indexing.

WGS, Whole Genome Sequencing — Sequencing of an entire genome, in contrast to sequencing of single genes or DNA fragments.

Appendix B. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.mimet.2011.06.019.

References

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., O'Toole, P.W., 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. 38, e200.

Cox, M.P., Peterson, D.A., Biggs, P.J., 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinforma. 11, 485.

DiGuistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M., Mardis, E., Marra, M.A., Hamelin, R.C., Bohlmann, J., Breuil, C., Jones, S.J., 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol. 10, R94.

DiGuistini, S., Wang, Y., Liao, N.Y., Taylor, G., Tanguay, P., Feau, N., Henrissat, B., Chan, S.K., Hesse-Orce, U., Alamouti, S.M., Tsui, C.K., Docking, R.T., Levasseur, A., Haridas, S., Robertson, G., Birol, I., Holt, R.A., Marra, M.A., Hamelin, R.C., Hirst, M., Jones, S.J., Bohlmann, J., Breuil, C., 2011. Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. Proc Natl Acad Sci USA 108, 2504–2509.

Ellwood, S.R., Liu, Z., Syme, R.A., Lai, Z., Hane, J.K., Keiper, F., Moffat, C.S., Oliver, R.P., Friesen, T.L., 2010. A first genome assembly of the barley fungal pathogen *Pyrenophora teres* f. *teres*. Genome Biol. 11, R109.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Williams, R.L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S., Jaffe, D.B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108, 1513–1518.

Haridas, S., 2010. The genome of *Trametes cingulata* and the search for a lignin depolymerase. Ph.D. Thesis, University of Minnesota, Minneapolis, MN, USA. 178 pages.

Haridas, S., Gantt, J.S., 2010. The mitochondrial genome of the wood-degrading basidiomycete *Trametes cingulata*. FEMS Microbiol. Lett. 308, 29–34.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, R., Li, Y., Kristiansen, K., Wang, J., 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713–714.

Ma, L.J., Fedorova, N.D., 2010. A practical guide to fungal genome projects: strategy, technology, cost and completion. Mycol: Int J Fungal Biol. 1, 9–24.

Medini, D., Serruto, D., Parkhill, J., Relman, D.A., Donati, C., Moxon, R., Falkow, S., Rappuoli, R., 2008. Microbiology in the post-genomic era. Nat. Rev. Microbiol. 6, 419–430.

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46.

Miller, J.R., Koren, S., Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. PNAS 74, 5463–5467.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.

Peng, Y., Leung, H., Yu, S.M., Chin, F.Y.L., 2010. IDBA — a practical iterative de Bruijn graph de novo assembler. Res Comput Mol Biol LNBI 6044, 426–440.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.