Genomic and transcriptomic data production for helminths

John Martin (jmartin@wustl.edu)
Mitreva Lab (MGI, Washington University in St. Louis)
Makedonka Mitreva
MGI, Washington University in St. Louis

Method Article

Keywords: sequencing, assembly, contamination screening, genomic data production, transcriptomic data production, helminths, nematoda, platyhelminthes, genome annotation

Posted Date: May 17th, 2018

DOI: https://doi.org/10.1038/protex.2018.044

Abstract

Protocols for genomic and transcriptomic data production, assembly and quality control.

Introduction

A multi-step protocol for preparation of draft genomes, specifically applicable to Nematoda. This includes initial library preparation, genome and transcriptome assembly, assembly QC/contamination screening and gene prediction. Both 454 and Illumina sequencing platforms are included.

Procedure

A Genome sequencing library preparation Paired end short insert libraries 454 titanium fragment libraries are constructed with 5-10ug of DNA according to the manufacturer's recommendations \(Roche 454). Illumina small-insert paired-end libraries are prepared according to the manufacturer's protocol with the exception that multiple library enrichment reactions and size selection are performed after amplification and multiple size fractions \(300-400 and 400-500 bp) are collected. 454/Illumina 3 kb insert mate pair libraries 3kb mate pair libraries are created as follows: 1. DNA is sheared into 3kb fragments, blunt ended and ligated to the SOLiD Mate-Pair Cap Adapter \(ABI). 2. Ligated DNA is size fractionated to 2-4 kb fractions and then purified. 3. Circularization reactions are set up using 1 µg of the extracted fraction and 1.3 pmol of the Internal SoliD Mate Pair adaptor \(ABI). 4. Linear \(or non-circularized and nicked) fragments are removed, circularized fragments are nick-translated extending from gaps engineered within the cap adapter using 200 ng of library and 20 units of DNA polymerase I. 5. Nick-translation reactions are treated with S1 nuclease for 15 minutes. 6. Resulting products are blunt ended and immobilized using Dynal M270 Streptavidin beads \(Invitrogen). 7. For 454 sequencing, FLX Titanium paired-end library adaptors are ligated onto the immobilized DNA fragments and processed as recommended by the Manufacturers 3 kb span paired end library construction protocol \(Roche 454). For Illumina sequencing, blunt ended fragments are processed through an adenylation reaction. Illumina’s Truseq adaptors are ligated, the library is enriched with KAPA HiFi polymerase \(KAPA Biosystems) and a final dual SPRI size selection is performed to isolate 300-500 bp library fragments. 454/Illumina 8 kb insert mate pair libraries 8kb mate pair libraries are created as follows: 1. 8 kb span paired-end libraries are constructed for 454 sequencing according to the manufacturer’s recommendations \(Roche 454), except that the 6.5-9 kb fraction is extracted from the size selection gel and the extracted adaptor ligated DNA is purified using a Qiagen Gel Extraction Kit. For Illumina sequencing, 15 µg of high molecular weight DNA is sheared to a mean fragment size of 8 kb with a Hydroshear, blunt ended using DNA Terminator End Repair Kit \(Lucigen) and ligated with 20 µM Circularization Adapters \(Roche). 2. The ligated DNA is size-fractionated and the 6.5-10 kb fraction is purified using the Qiagen Gel Extraction Kit. 3. 300 ng of size selected DNA is circularized using 10U of Cre Recombinase. Linear \(or non-circularized and nicked) library fragments are removed. 4. The circularized library fragments are fragmented targeting a mean insert size of 300 bp. 5. The fragmented DNA is blunt ended using the DNA Terminator End Repair Kit \(Lucigen), processed through an adenylylation reaction \(NEB) and Illumina's TruSeq adaptors are ligated. 6. The adenylated fragments are immobilized with Dynal M270 Streptavidin beads and amplified with KAPA HiFi Polymerase \(KAPA Biosystems). 7. The final 300-500 bp library fragments are selected with a dual SPRI reaction. Genomes sequenced on the Roche/454 platform are assembled from a combination of fragment reads, 3 kb paired-end reads and 8 kb paired-end reads generated to meet the coverage criteria of 15x, 15x and 3x respectively, with a target of 30x coverage for the final assembly. Genomes sequenced on the Illumina platform had overlapping fragment reads, 3 kb and 8 kb paired-end reads and are sequenced to a depth of 45x, 45x, and 10x, respectively. **B** Genome assembly Assemblies are generated using the assembly workflows outlined in Fig. 1, with the specific method depending on the input material. Assemblies based on Roche 454 3kb, 8kb and fragment input followed the steps detailed in panel m1. Assemblies from Illumina 3kb, 8kb and fragment input used the workflow described in panel m2 and a reference guided assembly method shown in panel m3. See figure in Figures section. **Figure** **1:** The McDonnell Genome Institute

Genome assembly pipelines 454 data Assemblies built using a combination of Roche 454 3kb, 8kb and fragment input data are constructed as follows \(Fig. 1 panel m1) 1. A combination of 3kb, 8kb and fragment 454 reads are subject to adapter removal, quality trimming and length filtering using a combination of the Flexbar[1] and Trimmomatic[2] tools \(Fig. 1). 2. Contaminant screening is done using the Bowtie2[9] aligner and a local contaminant database containing ribosomal RNA, bacteria and host sequence. 3. Cleaned reads are then assembled using the Newbler assembler[3] before being scaffolded with an in-house tool CIGA which links contigs based on cDNA evidence. 4. The resulting assembly is improved using another local tool named Pygap that uses Illumina short paired end sequences to help fill gaps between scaffolded contigs. 5. Finally the L_RNA_scaffolde[4] used 454 cDNA data to further improve scaffolding. Illumina data Assemblies constructed from 3kb, 8kb and fragment Illumina sequences followed this methodology \(Fig. 1 panel m2) 1. 3kb, 8kb and fragment Illumina sequences are subject to the adapter removal, quality trimming, length filtering and contamination screening process described for the 454 data above. 2. The cleaned reads are then assembled using the AllPaths-LG assembler[5] before being improved using the Pygap and L_RNA_scaffolder tools as described above. Reference guided, assisted assembly Finally we have a protocol for a reference guided, assisted assembly approach \(Fig. 1 panel m3) 1. Illumina 3kb paired end sequence data are subject to the same 'cleaning' procedure described above and is then mapped against the reference assembly using the bwa aligner[6] with default parameters. 2. Samtools mpileup is run on the alignment along with vcfutils.pl varFilter using suggested argument settings to identify the differences between the new input reads and the reference backbone. 3. We then subtract out the reference from the mapped region by omitting all SNP loci where the alternate allele frequency is 1. 4. We then used FastaAlternateReferenceMaker method of GATK \("http://software.broadinstitute.org/gatk/":http://software.broadinstitute.org/gatk/) and a bed file comprising only regions where the new data mapped to the reference to construct a new consensus populated with alleles from the new reads at each detected SNP locus. 5. In addition, reads that did not initially map to the reference are assembled using Velvet[7], with a kmer size of 39 chosen by the VelvetOptimizer \("http://bioinformatics.net.au/software.velvetoptimiser.shtml":http://bioinformatics.net.au/software.velvetoptimiser.shtml). 6. BLAT[8] is then used to compare the contigs created by Velvet to the contigs created by alignment to the reference and all Velvet contigs greater than 500 bp that mapped less than 50% of their length \(and at >80% identity) to an existing contig are added to the assembly. **C** Assembly QC / Contamination screening All assemblies are screened, to remove for contamination, before annotation. 1. Adaptor sequencs and contaminants are identified by compared contigs to a database of vectors, bacterial and host contaminants using Megablast. 2. High-scoring segment pairs \(HSPs) with E-value <0.01 and length >100 bp are picked. The final alignment length is from the first base of the first HSP to the last base of the last HSP. 3. The contig is removed if the identity is greater than 75% and the coverage is greater than 40% of the contig, or the contig is less than 2000 bp. 4. Any contigs which are on the border of the requirements and longer in length are manually reviewed as an extra measure against true genome contigs being removed. **D** Transcriptome sequencing and assembly Assembled RNAseq ata are used alongside EST data in the MAKER stage of gene prediction. 1. Transcriptome \(RNAseq) libraries are generated with the Illumina TrueSeq stranded protocol following the manufacturer's guidelines. 2. Raw reads are cleaned using an in-house tool that trims adaptor, quality trims and applies a length filter using Flexbar[1] and Trimmomatic[2]. 3. Low complexity sequence is masked using the filter_by_complexity tool in the seq_crumbs package \("http://bioinf.comav.upv.es/seq_crumbs/":http://bioinf.comav.upv.es/seq_crumbs/), and contaminating sequences are identified using Bowtie2[9] and TopHat2[10] before being removed using local code. 4. The cleaned, filtered RNAseq reads are assembled de novo with Trinity[11], using both left and right cleaned paired reads. 5. The output is filtered for the longest representative open reading frame, resulting in a 'best candidates' file. 6. Transcripts are merged using CD-HIT[12] with 98% coverage and identity. 7. The assembled contigs are assessed for quality by aligning \ (with TopHat2[10]) back to reference assembly to establish the percentage of reference aligned to by the reads and the percentage of reads that aligned to the reference. **E** Gene prediction Gene prediction is run on assemblies as follows: 1. For each assembly a repeat library is generated using RepeatModeler. Ribosomal RNA genes are identified using RNAmmer \("http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?rnammer":http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?

rnammer) and transfer RNAs are identified with tRNAscan-SE[13]. 2. Non-coding RNAs, such as microRNAs, are identified by searching against Rfam[14]. 3. Repeats and predicted RNAs are then masked using RepeatMasker. 4. Protein-coding genes are predicted from the masked assembly using a combination of several ab initio programs: SNAP[15], FGENESH \(Softberry, Corp), Augustus[16] and the MAKER pipeline[17], which aligns mRNA, EST and protein evidence from the same species or cross-species to aid in gene structure determination and modifications \(Fig. 2). 5. SNAP and Augustus models are generated where possible using the MAKER pipeline and species-specific evidence. A consensus gene set from the above prediction algorithms is generated, using a logical, hierarchical approach developed at MGI. See figure in Figures section. **Figure** **2:** The McDonnell Genome Institute Gene-finding pipeline. High confidence gene selection A high confidence gene set is created from MAKER[17] output: 1. Quality Index \(QI) criteria are calculated as follows: \(i) length of the 5' UTR; \(ii) fraction of splice sites confirmed by an EST; \(iii) fraction of exons that overlapped an EST alignment; and \(iv) fraction of exons that overlapped EST or protein alignments. 2. These decision-making steps are followed to define the set of high confidence genes: a) Genes are screened for overlaps \(<10% overlap is allowed). b) If QI\[2] and QI\[3] are >0, or QI\[4] is >0, then the gene is kept. c) Genes are retained if they matched Swissprot[18] using BLAST \(E<1e-06). d) Genes are retained if they matched Pfam[19] using RPSBLAST \(E<1e-03). e) RPSBLAST is run against CDD[20] \(E<1e-03 and coverage >40%). Genes that met both cut-offs are kept. f) If no hit is recorded the gene is retained if it had ≥ 55% identity to the genes database from KEGG[21], and and a bitscore of ≥35. Additional curation of gene sets Depending on the nature of the final gene set in relation to the assembly quality some gene sets underwent an additional manual review of short genes lacking definitive evidence. After the high confidence gene selection steps described above, shorter single and double exon genes and genes annotated as hypothetical \(with no KEGG nor InterPro homologies) are further scrutinized. A manual review of the Annotation Edit Distance \(AED, from MAKER) is considered in combination with the QI scores \(all provided by MAKER), enabling analysts to make a more informed decision about whether to keep or discard each such gene.

Anticipated Results

raw sequence data, genomic and/or transcriptomic assembly and a high confidence gene set.

References

1. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology \(Basel) 1, 895-905, doi:10.3390/biology1030895 \(2012). 2. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120, doi:10.1093/bioinformatics/btu170 \(2014). 3. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380 \(2005). 4. Xue, W. et al. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics 14, 604, doi:10.1186/1471-2164-14-604 \(2013). 5. Butler, J. et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18, 810-820, doi:10.1101/gr.7337908 \(2008). 6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760, doi:10.1093/bioinformatics/btp324 \(2009). 7. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821-829 \(2008). 8. Kent, W. J. BLAT–the BLAST-like alignment tool. Genome Res 12, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 \(2002). 9. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature methods 9, 357-359, doi:10.1038/nmeth.1923 \(2012). 10. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36, doi:10.1186/gb-2013-14-4-r36 \(2013). 11. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644-652, doi:10.1038/nbt.1883 \(2011). 12. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152, doi:10.1093/bioinformatics/bts565 \(2012). 13. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a

program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research 25, 955-964 \(1997). 14. Nawrocki, E. P. et al. Rfam 12.0: updates to the RNA families database. Nucleic acids research 43, D130-137, doi:10.1093/nar/gku1063 \(2015). 15. Korf, I. Gene finding in novel genomes. BMC bioinformatics 5, 59, doi:10.1186/1471-2105-5-59 \(2004). 16. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research 34, W435-439, doi:10.1093/nar/gkl200 \(2006). 17. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188-196, doi:10.1101/gr.6743907 \(2008). 18. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research 28, 45-48 \(2000). 19. Finn, R. D. et al. Pfam: the protein families database. Nucleic acids research 42, D222-230, doi:10.1093/nar/gkt1223 \(2014). 20. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. Nucleic acids research 43, D222-226, doi:10.1093/nar/gku1221 \(2015). 21. Kanehisa, M. The KEGG database. Novartis Found Symp 247, 91-101; discussion 101-103, 119-128, 244-152 \(2002).

Figures

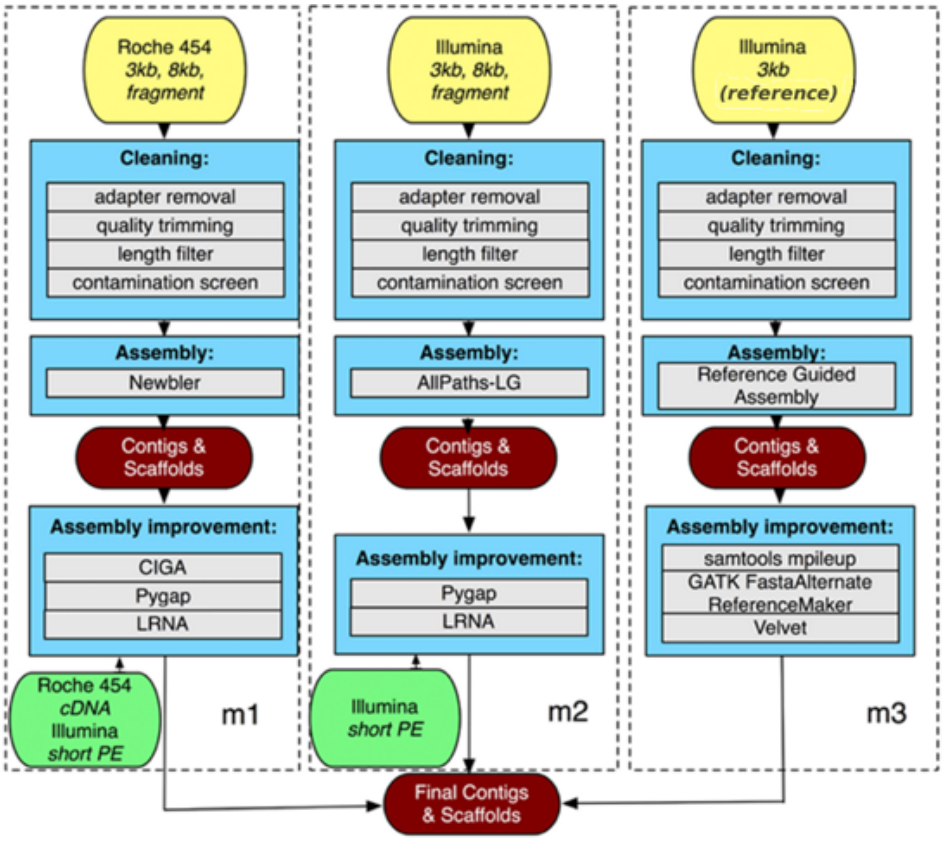

Figure 1

The McDonnell Genome Institute Genome assembly pipelines.

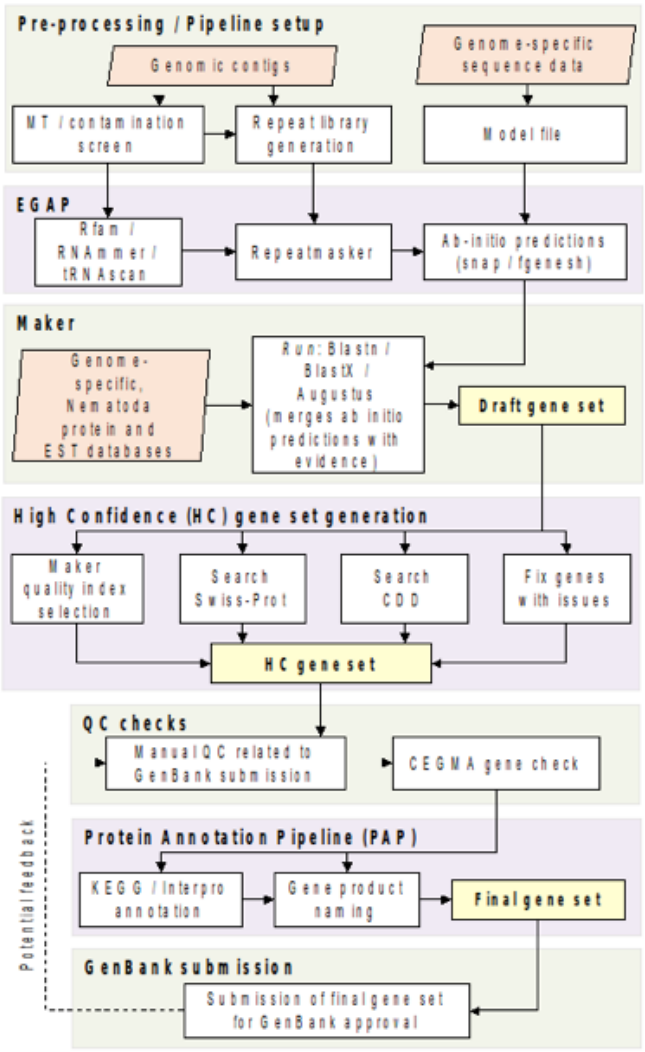

Figure 2

The McDonnell Genome Institute Gene-finding pipeline.