Title

Development of genomic resources for four potential environmental bioindicator species: *Isoperla grammatica, Amphinemura sulcicollis, Oniscus asellus and Baetis rhodani*.

Authors

Hannah C Macdonald*1, Luis Cunha1,2 and Michael W Bruford1

Affiliation

1 School of Biosciences, Cardiff University, Cardiff, UK
2 Embrapa Florestas, Estrada da Ribeira km. 111, Colombo-PR, 83411-000, Brazil

Keywords: invertebrates, freshwater, terrestrial, *de novo* assembly

*Corresponding author

Hannah C Macdonald
Cardiff University
The Sir Martin Evans Building,
Museum Avenue,
Cardiff,
CF10 3AX
Email: BurtonHC@cardiff.ac.uk

Running Title: Genomic resources for four macroinvertebrates

Abstract

A low-coverage genome was generated for each of four environmental key-species of macroinvertebrate taxa for the primary purpose of microsatellite marker development. *De novo* assemblies and microsatellite markers were designed for the freshwater species *Isoperla grammatica, Amphinemura sulcicollis*, and *Baetis rhodani* but have not been completed for the common shiny woodlice *Oniscus asellus*. Here, the data is made available, and the methods and pipeline are described which led to the creation of this resource. As widespread and functionally important organisms, which are often neglected in favour of studies on vertebrates, this data will be a useful resource for further research.
Introduction

Macroinvertebrates are widespread, often dominant and functionally important members of their environment that, coupled with their relative ease of sampling make them ideally suited for use as indicator species for biomonitoring and conservation assessment (Pfrender et al. 2010; Buss et al. 2015; Cardoni et al. 2015), and are recognised as such by the Water Framework Directive (2000/60/CE) (European Commission 2000). However, their use in genetic approaches is still limited, often being neglected from studies because of the lack of data (Cardoso et al. 2011).

Four macroinvertebrate species were sequenced for the primary purpose of developing microsatellite markers for use in population genetics; these include three freshwater invertebrates (Amphinemura sulcicollis, Isoperla grammatica, and Baetis rhodani) and one terrestrial soil invertebrate, the common shiny woodlice, Oniscus asellus. They all represent dominant, widespread species and therefore can be used as biomonitoring tools that will be effective at large spatial scales, as policy demands (Statzner and Bêche 2010). Microsatellite markers within this group are scarce, for example, within the large and diverse groups of Plecoptera and Ephemeroptera, there are only five species with between 3-13 microsatellites each, therefore this data will be a valuable and considerable resource for future research. This data could be used for further study into these invertebrates, such as describing their mitochondrial genome (as in Stewart and Beckenbach (2006)), or studying their genome content; evolutionary analyses (e.g. divergent rates), and further investigation of their genetic features (as in Li et al. (2010)).

Data Access

Raw data is stored in NCBI’s Sequence Read Archive (SRA): NGS data for four invertebrates: Amphinemura sulcicollis, Isoperla grammatica, Baetis rhodani and Oniscus asellus (STUDY: PRJNA315680 (SRP072016)).

1. NGS sequence data (raw data sent from sequencing centres):

   - **Amphinemura sulcicollis**:
     
     SAMPLE: Amphi_NGS (SRS1349204)
     EXPERIMENT: Amphi_NGS (SRX1642982)
     RUN: Amphi_NGS (SRR3262386)
WTCHG_93433_274_1.fastq.gz
WTCHG_93433_274_2.fastq.gz
WTCHG_93434_274_1.fastq.gz
WTCHG_93434_274_2.fastq.gz

• **Isoperla grammatica:**

  SAMPLE: Iso_NGS (SRS1351356)
  EXPERIMENT: Iso_NGS (SRX1648180)
  RUN: Iso_NGS (SRR3262388)

  WTCHG_93433_273_1.fastq.gz
  WTCHG_93433_273_2.fastq.gz
  WTCHG_93434_273_1.fastq.gz
  WTCHG_93434_273_2.fastq.gz

• **Baetis rhodani:**

  SAMPLE: Baetis_NGS (SRS1351357)
  EXPERIMENT: Baetis_NGS (SRX1648181)
  RUN: Baetis_rhodani_NGS (SRR3262630)

  Beatis_L3_1.fq.gz.1.gz
  Beatis_L3_2.fq.gz.2.gz

• **Oniscus asellus:**

  SAMPLE: Woodlice_NGS (SRS1351401)
  EXPERIMENT: Oniscus_NGS (SRX1648318)
  RUN: Oniscus_NGS (SRR3263253)

  WTCHG_93433_275_1.fastq
  WTCHG_93433_275_2.fastq
  WTCHG_93434_275_1.fastq
  WTCHG_93434_275_2.fastq

2. Each freshwater species has a CONTIG file (after *de novo* assembly) deposited at DDBJ/ENA/GenBank under the accession’s listed below, any contigs under 200bp were removed.

• **Amphinemura sulcicollis:**

  SUBID: SUB1394726
  BioSample: SAMN04568201
  Accession: LVVV00000000
  Organism: Amphinemura sulcicollis Dwyryd
File name: amphi_kmer61.contig

- **Isoperla grammatica:**
  SUBID: SUB1397890
  BioSample: SAMN04568202
  Accession: LVVW00000000
  Organism: Isoperla grammatica Teifi
  File name: iso_kmer61.contig

- **Baetis rhodani:**
  SUBID: SUB1398024
  BioSample: SAMN04568203
  Accession: LVVX00000000
  Organism: Baetis rhodani Tywi
  File name: B_61.contig

3. Other data stored in Genbank:

- 132 Mitochondrial cytochrome c oxidase I (mtCOI) sequences for all four species (using barcoding primers from Folmer *et al.* (1994)) are available on genbank: Accession numbers KU955863-KU955994 (*Amphinemura sulcicollis* 31 sequences; *Isoperla grammatica* 29 sequences; *Baetis rhodani* 65 sequences; *Oniscus asellus* 6 sequences).

- 51 Microsatellite markers for *Isoperla grammatica*, *Amphinemura sulcicollis* and *Baetis rhodani* available on Genbank: between KR068997-KR069048 (Iso_1-18, Amp_1-21, B_1-13, respectively) and described fully in Macdonald *et al.* (2016) *in review*. Subsequent microsatellite marker development has not been completed for *Oniscus asellus*.

**Meta Information**

Data for the four draft genomes sequenced (Table 1) was generated at two different sequencing centres, which were compared for their cost effectiveness and yields. Libraries 1-3 (*A. sulcicollis, I. grammatica and O. asellus*) were sent to Oxford MRC Sequencing, multiplexed with five other samples (eight libraries) as part of collaboration at Cardiff University. Whereas *B. rhodani* was sent at a later date to Beijing Genomic Institute (BGI), along with two other samples, these three samples were labelled by BGI and multiplexed in one lane (see Table 1 for full details). The main goals of the experiment were to develop
enough genomic resources for each target species in order to retrieve enough high quality microsatellite markers.

**Library**

Multiple samples of each species were collected from sites around upland Wales, UK (Table 1) and stored in absolute ethanol. Genomic DNA was extracted from whole individuals using a High Pure PCR Template Preparation Kit for blood and tissue following the manufacturer’s instructions (Roche Diagnostics GmbH Mannheim, Germany). All samples were treated with RNase after DNA extraction. Individual samples were identified using Sanger sequencing, with standard barcoding primers from Folmer *et al.* (1994) and by comparing the sequences with data in Genbank. To assure sample quality, quantification was assessed using a Qubit and visualised on a gel (Figure 1). Nanodrop was used to assess contamination, where the 260/280 ratio were found to be between 1.8 and 2 and that the 260/230 ratio was between 2-2.2 across all analysed samples. The highest quantity and best quality samples were chosen; all species yielded DNA quantities required (which was 1-5μg of DNA normalized to a concentration of 50ng/μl) apart from *A. sulcicollis*, for which a vacuum concentrator had to be used. Samples showed high DNA integrity with no observed smearing on the electrophoresis gel (Figure 1).

The samples of *A. sulcicollis*, *I. grammatica*, and *B. rhodani* were all made up of only one individual; however the *O. asellus* sample is made up of two individuals pooled. This was because allozyme loci have been used to show that two genetically distinct sub populations of *O. asellus* exist (*O. asellus* and *O. occidentails*) (Bilton *et al.* 1999) within *O. asellus*, it was thought that mixing two individuals would give the highest chance of success at developing microsatellites for the largest amount of samples. However, this meant that a *de novo* assembly could not be performed on this species.

Genomic DNA for all four samples were sent to their respective sequencing centres for library preparation (DNA was sheared, Illumina adapters were ligated, libraries were controlled for quality, normalized, pooled) and sequencing on HiSeq run (Table 1).
Processing

For each library NGS created four raw Illumina read files (two libraries, each with two pairs), which was transferred to Linux, unzipped, and the two libraries were concatenated, leaving two files of two pairs (renamed from the raw file names in section Data Access, to Amp_1.fastq & Amp_2.fastq, Iso_1.fastq & Iso_2.fastq, Woo.fastq & Woo_2.fastq, and B_1.fastq & B_2.fastq for *I. grammatica*, *A. sulcicollis*, *O. asellus* and *B. rhodani*, respectively, Table 1). Quality control was preformed using Trimmomatic v0.32 (Lohse *et al.* 2012) and Musket v1.1 (Yongchao *et al.* 2013). Trimmomatic was used to cut adapters and other illumina-specific sequences from the reads. It was also used to remove reads of low quality and short length. In this case the threshold for quality window was set at 18 and the minimum length was 35bp (using phred33). Musket is multistage k-mer based corrector for Illumina short read data and was used to identify and remove any common Illumina errors for a higher quality de novo genome assembly.

For all aquatic (single sample) species SOAP de novo 2 (Luo *et al.* 2012) was then used to build de novo assemblies using short-reads. This was done in order to provide longer reads for microsatellite marker mining. Several de novo assemblies were run per species in order to test different Kmer values and best assembly metrics. Draft assemblies were chosen according to maximum contig and highest N50 value.

A de novo assembly was not attempted for the pooled sample of *O. asellus* due to the risk of chimeras, which is much higher for assemblies of mixed samples. Instead FLASH v 1.2.9 (Fast Length Adjustment of SHort reads) was used to merge paired ends creating reads of 300 bp ([http://ccb.jhu.edu/software/FLASH/MANUAL](http://ccb.jhu.edu/software/FLASH/MANUAL) [Date accessed: 02.03.16]).

*PrimerPipeline* ([http://www.scrufster.com/primerpipeline/](http://www.scrufster.com/primerpipeline/)) was then used to identify repeat regions within the data files and design forward and reverse primers for each microsatellite. It is a windows program incorporating MISA (MIcroSAellite identification tool, [http://pgrc.ipk-gatersleben.de/misa/](http://pgrc.ipk-gatersleben.de/misa/)) and Primer3 v.2.3.6 (Untergasser *et al.* 2012).

The full pipeline (including all scripts and annotations) is described in ‘Appendix 1 Script_NGS’ at the end of this manuscript.
Results

The NGS for all four species was very successful as the total number of reads (raw data) were very high (ranging from 71,727,142 to 123,076,504 reads), and they were of relatively high quality because the quality control sections of the pipeline did not remove too much (ranging from 0.3% and 11% of the total reads, see Table 1). *B. rhodani* data from BGI appears to be the most successful as it had the highest total reads and the lowest percentage of reads removed by quality control.

The *de novo* assemblies varied according to which kmer size was used (see
bioRxiv: Genomic Resource note

Table 2 for an example of how the kmer size affected the N50 in *A. sulcicollis*). For all three species that assemblies were performed for, kmer 61 was chosen as it produced the best assemblies. *A. sulcicollis* had the best N50 at 1,543, whereas *I. grammatica* had 568, meaning that on average the *de novo* assembly for *A. sulcicollis* produced larger contigs, therefore the best assembly.

Acknowledgements

The research was funded by the NERC DURESS project (Diversity in Upland Rivers for Ecosystem Service Sustainability, NE/J01481/1) within the Biodiversity and Ecosystem Service Sustainability (BESS) Thematic Programme. HCM was funded by the Cardiff University President’s Research Scholarships. Many thanks to Steve Ormerod and Hefin Jones from Cardiff University, for support and guidance.
Table 1. Details of the development of four separate libraries of macroinvertebrate using next generation sequencing.

<table>
<thead>
<tr>
<th>Species</th>
<th>Library 1</th>
<th>Library 2</th>
<th>Library 3</th>
<th>Library 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Amphinemura sulcicollis</em> (Stephens, 1836)</td>
<td><em>Isoperla grammatica</em> (Poda, 1761)</td>
<td><em>Oniscus asellus</em> (Linnaeus 1758)</td>
<td><em>Baetis rhodani</em> (Pictet, 1845)</td>
</tr>
<tr>
<td>Genus</td>
<td>Amphinemura</td>
<td>Isoperla</td>
<td>Oniscus</td>
<td>Baetis</td>
</tr>
<tr>
<td>Order</td>
<td>Plecoptera</td>
<td>Isopoda</td>
<td>Ephemeroptera</td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>Insecta</td>
<td>Malacostraca</td>
<td>Insecta</td>
<td></td>
</tr>
</tbody>
</table>

Meta Information

| Sequencing centre | The Oxford Genomics Centre (WTCHG) / High-Throughput Genomics (Oxford, UK) | BGI (Shenzhen, China) |
| Platform | Illumina | |
| Model | HiSeq 2500, Rapid run | HiSeq 2000 (PE91) |
| Analysis type | DNA | DNA |

Library

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Whole-genome shotgun sequencing of genomic DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shared lane</td>
<td>One lane (by itself)</td>
</tr>
<tr>
<td>Sample type (mtDNA seq name available in genbank)</td>
<td>One individual (95A1)</td>
</tr>
<tr>
<td>Source of material (Taxon)</td>
<td></td>
</tr>
<tr>
<td>Sample Location#</td>
<td>275070E 342810N</td>
</tr>
<tr>
<td>Catchment</td>
<td>Dwyryd Upland Wales, UK</td>
</tr>
<tr>
<td>Insert length</td>
<td>450</td>
</tr>
<tr>
<td>Max read length</td>
<td>300</td>
</tr>
</tbody>
</table>

Results

<p>| Total reads (before QC) | 79,196,610 | 71,727,142 | 81,811,132 | 123,076,504 |
| Total reads (after QC) | 71,796,770 | 64,628,924 | 74,813,127 | 122,669,217 |
| % removed | 10.3 | 11.0 | 9.4 | 0.3 |</p>
<table>
<thead>
<tr>
<th>Species</th>
<th>Library 1</th>
<th>Library 2</th>
<th>Library 3</th>
<th>Library 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Amphinemura sulcicollis</em> (Stephens, 1836)</td>
<td><em>Isoperla grammatica</em> (Poda, 1761)</td>
<td><em>Oniscus asellus</em> (Linnaeus 1758)</td>
<td><em>Baetis rhodani</em> (Pictet, 1845)</td>
</tr>
<tr>
<td>Mode of assembly</td>
<td><em>De novo</em> assembly</td>
<td><em>De novo</em> assembly</td>
<td>Flash</td>
<td><em>De novo</em> assembly</td>
</tr>
<tr>
<td>Best kmer size</td>
<td>61</td>
<td>61</td>
<td>/</td>
<td>61</td>
</tr>
<tr>
<td>N50</td>
<td>1,543</td>
<td>568</td>
<td>/</td>
<td>850</td>
</tr>
<tr>
<td>Total No. of contigs</td>
<td>91,245</td>
<td>356,623</td>
<td>/</td>
<td>144,347</td>
</tr>
<tr>
<td>Total scaffold length</td>
<td>182,044,631</td>
<td>304,248,447</td>
<td>/</td>
<td>162,525,650</td>
</tr>
<tr>
<td>Longest scaffold</td>
<td>148,977</td>
<td>16,928</td>
<td>/</td>
<td>47,828</td>
</tr>
</tbody>
</table>

# Coordinates are in EPSG:27700 - OSGB 1936 / British National Grid.
**Table 2.** Shows information on all *de novo* assemblies performed with *Amphinemura sulcicollis* data.

<table>
<thead>
<tr>
<th></th>
<th>Kmer_55</th>
<th>Kmer_61</th>
<th>Kmer_71</th>
<th>Kmer_81</th>
<th>Kmer_85</th>
<th>Kmer_91</th>
<th>Kmer_101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffold number</td>
<td>88,343</td>
<td>91,245</td>
<td>89,130</td>
<td>87,731</td>
<td>88,028</td>
<td>87,347</td>
<td>92,636</td>
</tr>
<tr>
<td>In-scaffold contig number</td>
<td>834,812</td>
<td>832,323</td>
<td>832,173</td>
<td>842,707</td>
<td>850,059</td>
<td>864,994</td>
<td>1,170,180</td>
</tr>
<tr>
<td>Total scaffold length</td>
<td>184,467,063</td>
<td>182,044,631</td>
<td>180,880,141</td>
<td>183,600,553</td>
<td>181,179,087</td>
<td>179,755,162</td>
<td>172,156,150</td>
</tr>
<tr>
<td>Average scaffold length</td>
<td>2,088</td>
<td>1,995</td>
<td>2,029</td>
<td>2,092</td>
<td>2,058</td>
<td>2,057</td>
<td>1,858</td>
</tr>
<tr>
<td>Filled gap number</td>
<td>177,030</td>
<td>183,386</td>
<td>171,646</td>
<td>170,306</td>
<td>161,405</td>
<td>153,310</td>
<td>173,854</td>
</tr>
<tr>
<td>Longest scaffold</td>
<td>149,091</td>
<td>148,977</td>
<td>140,148</td>
<td>149,117</td>
<td>147,505</td>
<td>147,516</td>
<td>113,833</td>
</tr>
<tr>
<td>Scaffold and singleton number</td>
<td>606,125</td>
<td>580,009</td>
<td>605,194</td>
<td>626,038</td>
<td>645,302</td>
<td>671,404</td>
<td>950,155</td>
</tr>
<tr>
<td>Scaffold and singleton length</td>
<td>297,424,877</td>
<td>287,874,638</td>
<td>295,491,963</td>
<td>300,591,981</td>
<td>302,518,974</td>
<td>303,469,211</td>
<td>305,381,813</td>
</tr>
<tr>
<td>Average length</td>
<td>490</td>
<td>496</td>
<td>488</td>
<td>480</td>
<td>468</td>
<td>451</td>
<td>321</td>
</tr>
<tr>
<td>N50</td>
<td>1,550</td>
<td>1,543</td>
<td>1,482</td>
<td>1,525</td>
<td>1,466</td>
<td>1,448</td>
<td>1,141</td>
</tr>
<tr>
<td>N90</td>
<td>195</td>
<td>193</td>
<td>192</td>
<td>196</td>
<td>198</td>
<td>203</td>
<td>102</td>
</tr>
<tr>
<td>Weak points</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Figure 1.** Photograph of a 3% ethidium-bromide stained agarose electrophoresis gel under UV light, showing genomic DNA of two individuals each species of the three species *Isoperla grammatica*, *Amphinemura sulcicollis* and *Oniscus asellus* that were sequenced first, compared to three concentrations of lambda DNA (left to right: 16.5 ng/µL, 34 ng/µL and 67 ng/µL).
References


Appendix 1: NGS script

Key: (Commands in red, Annotations and Instructions in Black)

1.0. Check and Unzip

#To check data files are unaffected from download/upload use md5sum to check the unique identity of the file (compare the md5 number, it has to be exactly the same format)

```
md5sum file1 > file1_md5.txt
```

#And check the first line of the sequence

```
head filename
```

#E.g. md5sum formats for *B. rhodani* raw data files:

```
#aa278e7dd0de7af2e12aaf0d4ba9fc97  Beatis_L3_1.fq.gz.cut/Beatis_L3_1.fq.gz.1.gz
#1ba32f3d3a72be4877311163ce07d3c6  Beatis_L3_1.fq.gz.cut/Beatis_L3_1.fq.gz.2.gz
#26ea5697b1c950aba2d83095f143f0c  Beatis_L3_2.fq.gz.cut/Beatis_L3_2.fq.gz.1.gz
#f7fee1ecf2b928cf4504e6d48f417636  Beatis_L3_2.fq.gz.cut/Beatis_L3_2.fq.gz.2.gz
```

#Check permissions, the following code changes the permissions of the file called est_trimmer.pl.

```
chmod 777 est_trimmer.pl
```

#Unzipping

#E.g. Rawdata files end in “gz” so they need to be unzipped:

```
#WTCHG_93433_274_2.fastq.gz
#WTCHG_93433_274_1.fastq.gz
#WTCHG_93434_274_1.fastq.gz
#WTCHG_93434_274_2.fastq.gz
```

#The following command will unzip everything ending in .gz in the background. The above file goes from a fastq.gz file to just a fastq file.

```
 gunzip *.gz &
```

#For unzipping program files e.g. musket. If ends in tar.bz the command is:
bioRxiv: Genomic Resource note

```bash
tar -xvjf
#If ends in tar.gz the command is:
tar -xvzf

#I had two libraries with a forward and reverse, put the two forwards into one file and the two backs in one file, just to make it simpler. The two 1's together and the two 2's.

#The following command tells Linux to concatenate the files called
‘WTCHG_93433_273_1.fastq’ and ‘WTCHG_93434_273_1.fastq’, and name the combined file iso_1.fastq, and do it all in the background (&). Note, you have to be in the directory that the files are in or tell Linux where to find them e.g. home/c1135170/Hannah/
cat WTCHG_93433_273_1.fastq WTCHG_93434_273_1.fastq > iso_1.fastq &
#Do the same with the other pair
cat WTCHG_93433_273_2.fastq WTCHG_93434_273_2.fastq > iso_2.fastq &
#E.g with B. rhodani:
cat Beatis_L3_1 fq.gz.1 Beatis_L3_1.fq.gz.2 > B_1.fastq &
cat Beatis_L3_2.fq.gz.1 Beatis_L3_2.fq.gz.2 > B_2.fastq &
```

3.0. Trimmomatic

#First need to download Trimmomatic, you can find here:
[http://www.usadellab.org/cms/?page=trimmomatic](http://www.usadellab.org/cms/?page=trimmomatic)

#Right click, ‘copy link address’ for the Binary.

`wget`

#Right click to paste:
[http://www.usadellab.org/cms/uploads SUPPLEMENTARY/Trimmomatic/Trimmomatic-0.32.zip](http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.32.zip)

#It downloads to the folder you’re in and it’s called Trimmomatic-0.32.zip

`unzip Trimmomatic-0.32.zip`
#now just called Trimmomatic-0.32

#Now ready to run Trimmomatic

#Code means: Using Trimmomatic PE (paired end) which can be found here (pathway) do “–phred33” to these two files (B_1.fastq & B_2.fastq) then rename them
trimmomatic_B_1.fastq.gz (for paired) and trimmomatic_B_1_unpaired.fastq.gz for unpaired, and the same with the other pair. The nohup at the beginning is there so I can close the window and it will still carry on running.

nohup java -classpath /home/c1135170/Hannah/app/Trimmomatic-0.32/trimmomatic-0.32.jar org.usadellab.trimmomatic.TrimmomaticPE –phred33 B_1.fastq B_2.fastq trimmomatic_B_1.fastq.gz trimmomatic_B_1_unpaired.fastq.gz trimmomatic_B_2.fastq.gz trimmomatic_B_2_unpaired.fastq.gz

ILLUMINAFLIP:/home/c1135170/Hannah/app/Trimmomatic-0.32/adapters/TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:18 MINLEN:35 &

#E.g. Nohup.txt

#Read Pairs: 61528668 Both Surviving: 61132622 (99.36%) Forward Only Surviving: 96495 (0.16%) Reverse Only Surviving: 298538 (0.49%) Dropped: 1013 (0.00%)

#Unzipped file (see section 1.0)

#To install, go to the Musket folder and press ‘make’:

cd ../app/cd musket-1.1/ls

make
#Then you can remove the original zipped file

rm musket-1.1.tar.bz

#2.0. Run Musket using file outputs from Trimomatic:

nohup /home/c1135170/Hannah/app/musket-1.1/musket -k 21 2192141955 -p 32 -omulti corrected -inorder trimmomatic_B_1.fastq.gz trimmomatic_B_2.fastq.gz trimmomatic_B_1_unpaired.fastq.gz trimmomatic_B_2_unpaired.fastq.gz 1>out.txt 2>error.txt

#They are named corrected.0, corrected.1, corrected.2, and corrected.3, in order of how the files were listed in the command above. Rename the files:

mv corrected.0 musket_B_1.fastq
mv corrected.1 musket_B_2.fastq
mv corrected.2 musket_B_1_unpaired.fastq
mv corrected.3 musket_B_2_unpaired.fastq

# FLASH (used for O. asellus only)

#To find Flash: http://ccb.jhu.edu/software/FLASH/

#To download

wget http://sourceforge.net/projects/flashpage/files/FLASH-1.2.9.tar.gz

#To unzip

tar -zxvf FLASH-1.2.9.tar.gz

# Command asks flash to merge paired ends <musket_woo_1> < musket_woo_1 > [-m minOverlap - varied] [-M maxOverlap - 100] [-x mismatchRatio-varied] [-p phredOffset] [-o prefixOfOutputFiles] [-d pathToDirectoryForOutputFiles] [-f averageFragment Length- 300] [-s standardDeviationOfFragments- varied] [-r averageReadLength- 150].

#Several different combinations tried to

../app/FLASH-1.2.9/flash musket_woo_1.fastq musket_woo_2.fastq -m 15 -M 100 -x 0.1 -p o merged -d -f 300 -s 50 -r 150 1>flash.out 2>flash.err

#10% retained and matched
bioRxiv: Genomic Resource note

../app/FLASH-1.2.9/flash musket_woo_1.fastq musket_woo_2.fastq -m 10 -M 100 -x 0.1 -o merged2 -d -f 300 → 40 -r 150 1>flash2.out 2>flash2.err &

#15% retained and matched

../app/FLASH-1.2.9/flash musket_woo_1.fastq musket_woo_2.fastq -m 25 -M 100 -x 0.1 -o merged3 -d -f 300 → 40 -r 150 1>flash3.out 2>flash3.err &

#9.5% retained and matched

../app/FLASH-1.2.9/flash musket_woo_1.fastq musket_woo_2.fastq -m 20 -M 100 -x 1 -o merged4 -d -f 300 → 40 -r 150 1>flash4.out 2>flash4.err &

#100% retained and matched

../app/FLASH-1.2.9/flash musket_woo_1.fastq musket_woo_2.fastq -m 20 -M 100 -o merged4 -d -f 300 → 40 -r 150 1>flash4.out 2>flash5.err &

#mismatchRatio: default 0.25 . 12.78% retained and matched

#Use merged extended frags to feed straight into MISA

#To continue must convert fastq file to fasta file, and removes spaces at the same time, using:

awk 'BEGIN{a=0}{if(a==1){print;a=0}}/@/ {print;a=1}' myFastqFile | sed 's/^@/>/' > myfastafile

########################################################################

6.0. SOAPdenovo2

#Download soapdenovo, (same way as Trimmomatic and Musket), find here:

http://sourceforge.net/projects/soapdenovo2/files/SOAPdenovo2/

#To download:


#Unzip

tar -xvzf SOAPdenovo2-src-r240-4.tar

#Compile by navigating to the folder that the 'makefile' is in and type:

Make
#First make config file to use with soap de novo, red needs to change depending on the data, especially the pathways to the musket output files so soap de novo knows where the files are. In this example the config file was named ‘iso_config.text’.

max_rd_len=150

[LIB]
#average insert size
avg_ins=300

#if sequence needs to be reversed
reverse_seq=0

#in which part(s) the reads are used
asm_flags=3

#in which order the reads are used while scaffolding
rank=1

#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=/home/c1135170/Hannah/Isoplera/musket_iso_1.fastq
q2=/home/c1135170/Hannah/Isoplera/musket_iso_2.fastq
q=/home/c1135170/Hannah/Isoplera/musket_iso_1_unpaired.fastq
q=/home/c1135170/Hannah/Isoplera/musket_iso_2_unpaired.fastq

#For kmers less than 63, use the following command, be in the same folder as the config file:

nohup ../app/soapdenovo/soapdenovo2-src-r240/soapdenovo-63mer all -s iso_config.txt -K 55 -R -o iso_kmer55 1>iso_kmer55.log 2>iso_kmer55.err

# Command explained: nohup ../app/soapdenovo/soapdenovo2-src-r240/soapdenovo-63mer (telling it where to find soapdenovo) all -s iso_config.txt (name of the config file we made) -K 55 (kmer size 61) -R -o iso_kmer55 1>iso_kmer55.log 2>iso_kmer55.err (names of the output files)

#For kmers above 63:

Nohup /home/c1135170/Hannah/app/soapdenovo2-bin-LINUX-generic-r240/soapdenovo-127mer all -s iso_config.txt -K 71 -p 10 -R -o iso_kmer71 1>kmer71.log 2>kmer71.err