Introduction

The Legacy Pipeline uses Make to pre-processes raw data from FASTQ inputs, align reads to a reference genome, and create input fasta files for metagenomic analysis.

Credits

Ana Duggan
Brian Golding
Hendrik Poinar
Katherine Eaton
Jessica Hider

Pipeline Steps

1) **Create reference genome indices for mapping (bwa, samtools)**
2) **Organize and rename input FASTQ sequences.**
3) **Format conversion (fastq -> bam)**
4) **Sequencing adapter removal and for paired end data merging (leeHom)**
5) **Read mapping to reference using (bwa aln)**
6) **Post-mapping filtering (samtools)**
7) **PCR duplicate removal (bam-rmdup, retrieveMappedProperlyPaired...)**
8) **Ancient DNA C-to-T damage pattern visualisation (mapDamage)**
9) **Metagenomics filtering (awk)**

Preparation

Remember that the purpose of keeping all the references in a single location is first and foremost consistency across users and projects.

1. As user poinarlab upload your reference.fasta[1] to
   /home/poinarlab/Reference_sequences
2. Index the reference for pipeline compatibility
   /home/ana/scripts/network-aware-bwa/bwa index reference.fasta
3. Index the reference for mapDamage compatibility
   samtools faidx reference.fasta
4. Make sure the read and execute permissions are open to all
   chmod +rx reference.fasta*

[1] Try to be informative with the file names, ex. Genus_species name plus GenBank accession number

Running The Pipeline

Updated pipeline usage[2]:
As per November 26, 2019 lab meeting discussion, this updated pipeline produces new files meant to ease data entry into metagenomic classifiers such as BLAST or Kraken, creates a text file summarising the exhaustion estimate for each library , and automatically adds a map quality filter in the final step.

1. Remember to save your .fastq.gz files pre-organised into folders such as Sample_ID

```
for f in `ls *R1_001.*fastq.gz`;
do
   mkdir Sample_${f%%_*} | mv ${f%%_*}* ./Sample_${f%%_*};
done
```

2. Use legacy.pl to generate a Makefile which contains all the commands necessary to process your data.

```
/home/ana/legacy.pl /path/to/data/folder[3]
/home/poinarlab/Reference_sequences/reference.fasta[4]  >Makefile
```

3. Launch make to process data

```
make -j N[5]
```

Output

a) **.sort.bam** → Paired R1/R2 reads (though file will be labelled as R1) data from compressed fastq files. Contains all sequenced reads that have been converted to bam format, trimmed and merged by leeHom, mapped against a given reference (and sorted based on mapping coordinates) but not yet filtered to remove unmapped reads.
b) **.EditDist.min24.fasta** → Reads combined from all sort.bam files, converted to fasta format, filtered for minimum length 24 bp and any lingering similarity to sequencing adapters.
c) **.AllUniq.EditDist.min24.fasta** → As above but with string duplicate reads removed. This is the file you should use as input into metagenomic classifiers.
d) **.mapped.bam** → Reads restricted to those that map to the given reference, are either merged and mapped or are unmerged but still properly paired, and filtered to remove reads with identical 5' and 3' mapping coordinates.

[2] For double stranded libraries sequenced with the dsLP adapters
[3] This has to be the folder containing all the Sample_ID folders
[4] Recall that you can map to multiple references sequentially by listing them one after another, but this should be done very carefully to avoid overloading the cluster and causing dropped jobs
[5] Where N is an appropriate integer based on cluster use (use top to check)

e) **.fld.txt** → List of insert sizes for all molecules in .mapped.bam, this can be used as input file for /home/ana/scripts/plotFLD.r, or a histogram program of your choosing.
f) **.Exhaust.txt** → Exhaustion estimate of library mapped to given reference as calculated by the biohazard bam-rmdup function[6].
g) **.min24MQ30.bam** → Mapped.bam file further filtered to minimum length of 24bp and minimum map quality of 30

Troubleshooting

Pipeline not working? Here are some suggestions:

- Make sure the reference is indexed properly.
- Make sure you're directing the pipeline to the appropriate folder containing your data. Issues often arise when you are in the folder the program is trying to find (want to be above Sample)
- Make sure that .sort.bam files aren't empty[7].
- If mapping to multiple references, best practice is to make sure that all the .sort.bam files have the same number of lines
- Always remember, tab complete is your friend 🙂

Shared Library Error

```
/home/ana/scripts/network-aware-bwa/bwa: error while loading shared libraries:
libzmq.so.4: cannot open shared object file: No such file or directory
```

Cause: The program LD cannot find the C-code library libzmq.
Fix:

```
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/lib/
```

Bam Header Error

```
[bam_header_read] bgzf_check_EOF: Invalid argument
[bam_header_read] invalid BAM binary header (this is not a BAM file).
```

Cause: The pipeline was interrupted before completing and the *.sort.bam files in the raw data directory are empty.
Fix: In the raw data directory for the offending sample, deleted the *.sort.bam files.

---

[6] This removes the need to end make call with 1>stdout.txt
[7] This can happen when the job ends unexpectedly. If the sort.bam files are empty the rest of the pipeline won't proceed. Delete all pipeline-produced files and start over

Tool References

- bwa Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 Download: http://bio-bwa.sourceforge.net/bwa.shtml
- SAMtools Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 Download: http://www.htslib.org/
- leeHom Renaud, G., Stenzel, U., & Kelso, J. (2014). leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic acids research, 42*(18), e141. https://doi.org/10.1093/nar/gku699 Download: https://github.com/grenaud/leeHom
- Bam-rmdup Stenzel, Udo. Bio-hazard tools. MPI EVA Bioinformatics. https://github.com/mpieva/biohazard-tools/ Download: http://haskell.org/ghc) and http://haskell.org/cabal)
- Make
- mapDamage Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics (Oxford, England)*, *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193. Download: https://ginolhac.github.io/mapDamage/

Appendix

Detailed Usage and Command Parameters

1. **Log on to the shared poinar lab workspace**

```
ssh poinarlab@info.mcmaster.ca
```

2. **Navigate to the testing directory**

```
cd /home/poinarlab/Docs/pipeline_test
```

3. **Run the pipeline script to generate the test Makefile**

```
/home/ana/legacy.pl \
/home/poinarlab/Docs/pipeline_test/data
\ /home/poinarlab/Reference_sequences/Yersinia_pestis_CO92_whole_genome.fasta \
> Makefile
```

4. **Use a text editor (ex. nano or vim) to inspect the Makefile)**
5. **Run the pipeline example.**

```
cd output/
make -j 1 -f ../Makefile 2>&1 | tee Makefile.log
```

Pipeline - Mapping

**/home/ana/scripts/BCL2BAM2FASTQ/fastq2bam/fastq2bam **

-r EB34M2 \
-o /dev/stdout \
/home/poinarlab/Docs/pipeline_test/data/Sample_EB34M2/EB34M2_S288_L002_R1_001.fastq.gz \
/home/poinarlab/Docs/pipeline_test/data/Sample_EB34M2/EB34M2_S288_L002_R2_001.fastq.gz | \

E.g. of more detailed info
Why-because bam files take up less space...easier processing
How-fastq2bam

**leeHom **

--ancientdna \
--log
/home/poinarlab/Docs/pipeline_test/data/Sample_EB34M2/EB34M2_S288_L002_R1_001.fastq.gzleehom.txt \

-o /dev/stdout \
/dev/stdin | \

**/home/ana/scripts/network-aware-bwa/bwa bam2bam **

-n 0.01 \
-o 2 \
-l 16500 \
-g /home/poinarlab/Reference_sequences/Yersinia_pestis_CO92_whole_genome.fasta \
/dev/stdin | \

**/home/ana/scripts/samtools-patched/sam sort **

-o /dev/stdin \
/home/poinarlab/Docs/pipeline_test/data/Sample_EB34M2/EB34M2_S288_L002_R1_001.fastq.gz.sort \
>/home/poinarlab/Docs/pipeline_test/data/Sample_EB34M2/EB34M2_S288_L002_R1_001.fastq.gzYersinia_pestis_CO92_whole_genome.sort.bam

#this happens for lane 1 and then again for lane 2

**/home/gabriel/libbam/insertSize **

EB34M2_Yersinia_pestis_CO92_whole_genome.mapped.bam | \
sort \
>EB34M2_Yersinia_pestis_CO92_whole_genome.FLD.txt

**/home/ana/scripts/samtools-patched/sam view **

-b \
-m 24 \
-q 30 \
-o EB34M2_Yersinia_pestis_CO92_whole_genome.min24MQ30.bam \
EB34M2_Yersinia_pestis_CO92_whole_genome.mapped.bam

Pipeline - Metagenomics

**/home/ana/scripts/samtools-patched/sam merge **

/home/poinarlab/Docs/pipeline_test/data/Sample_EB34M2/EB34M2_S288_L002_R1_001.fastq.gzYersinia_pestis_CO92_whole_genome.sort.bam \
/home/poinarlab/Docs/pipeline_test/data/Sample_EB34M2/EB34M2_S288_L001_R1_001.fastq.gzYersinia_pestis_CO92_whole_genome.sort.bam | \

**/home/ana/scripts/samtools-patched/sam view **

-m 24 \
/dev/stdin | \

```
        cut -f 1,10 | \
        agrep -v -1 AGATCGGAA | agrep -v -1 TTCCGATCT | sed 's/^/>/g' | sed 's/\t/\n/g'
        >EB34M2_EditDist.min24.fasta

/home/keaton/scripts/NGSXRemoveDuplicates \
        --fasta EB34M2_EditDist.min24.fasta \
        --output EB34M2_AllUniq.EditDist.min24.fasta \
        --stats EB34M2_DeDupStats.txt
```

Pipeline Mods

Length Filter

Change the minimum length filter from 24 bp to 35 bp:

```
sed 's/-m 24/-m 35/g' Makefile | sed 's/min24/min35/g' > Makefile.min35.mk
```

Super Speed

Bump up the number of cores used per sample during mapping:

```
sed 's/bam2bam/bam2bam -t 10/g' Makefile > Makefile.threads10.mk
```