Time-saving workflows and easy parallel processing

Workflows

- Pipelines
- Scripts
- Data cleanup
- Analysis steps
- Producing graphs and charts
- etc…

Workflows

- Correct
- Reproducible
- Documented

- Easy is better than tedious
- Fast is better than slow

Workflows

- Correct*
- Reproducible
- Documented

- Easy is better than tedious
- Fast is better than slow

* http://software-carpentry.org/blog/2013/02/correctness-isnt-compelling.html
http://www.davidhbailey.com/dhbpapers/icerm-report.pdf

`make`

- Produces (*makes*) files using recipes
- Recipes are plain text files named *`Makefile`*
- Language agnostic
- Only does the work necessary
- Stops on error

- Simple to start, allows complexity

`Makefile` recipes

- Filename(s) to make = *targets*
- Necessary input file(s) = *prerequisites*
- A set of commands to run = *actions*

Makefile recipes

- Filename(s) to make = *targets*
- Necessary input file(s) = *prerequisites*
- A set of commands to run = *actions*

```
seqs_aa.fasta: seqs_na.fasta
   transeq -sequence seqs_na.fasta \
           -outseq seqs_aa.fasta   \
           -frame 1 -clean
```

Makefile recipes

- Filename(s) to make = *targets*
- Necessary input file(s) = *prerequisites*
- A set of commands to run = *actions*

```
seqs_aa.fasta: seqs_na.fasta
  transeq -sequence $<                    \
          -outseq $@                      \
          -frame 1 -clean
```

```
seqs_aa.fa: seqs_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

seqs_aa.aligned.fa: seqs_aa.fa
   muscle -quiet < $< > $@

seqs_aa_freq.tsv: seqs_aa.aligned.fa
   # CountAAFreq.pl only takes Nexus
   fasta2nexus < $< > seqs_aa.nxs
   perl CountAAFreq.pl seqs_aa.nxs $@ 0.25 0.5
   rm seqs_aa.nexus
```

```
seqs_aa.fa: seqs_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

seqs_aa.aligned.fa: seqs_aa.fa
   muscle -quiet < $< > $@

seqs_aa.nxs: seqs_aa.aligned.fa
   fasta2nexus < $< > $@

seqs_aa_freq.tsv: seqs_aa.nxs
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

```
seqs_aa.fa: seqs_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

seqs_aa.nxs: seqs_aa.fa
   muscle -quiet < $< | fasta2nexus > $@

seqs_aa_freq.tsv: seqs_aa.nxs
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

```
%_aa.fa: %_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

seqs_aa.nxs: seqs_aa.fa
   muscle -quiet < $< | fasta2nexus > $@



seqs_aa_freq.tsv: seqs_aa.nxs
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

```
%_aa.fa: %_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

%.nxs: %.fa
   muscle -quiet < $< | fasta2nexus > $@


seqs_aa_freq.tsv: seqs_aa.nxs
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

```
%_aa.fa: %_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

%.nxs: %.fa
   muscle -quiet < $< | fasta2nexus > $@

%_aa_freq.tsv: %_aa.nxs
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

```
%_aa.fa: %_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

%.nxs: %.fa
   muscle -quiet < $< | fasta2nexus > $@

# Keep intermediate alignments, for speed
.PRECIOUS: %.nxs

%_aa_freq.tsv: %_aa.nxs
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

Advanced features

- Variables

```
NAME := Thomas
hello:
   echo 'Hi, my name is $(NAME).'
```

- Using $ signs in your recipes

```
check_balance:
   echo 'Your balance is $$17.03.'
```

- Targets don’t have to be files

Advanced features

- Recipes don't have to have actions

```
all: gag_aa_freq.tsv env_aa_freq.tsv
```

- Prerequisites don't have to be recipes

```
%_aa_freq.tsv: %_aa.nxs CountAAFreq.pl
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

Advanced features

- Recipes don't have to have actions

```
all: gag_aa_freq.tsv env_aa_freq.tsv
```

- Prerequisites don't have to be recipes

```
%_aa_freq.tsv: %_aa.nxs CountAAFreq.pl
   perl CountAAFreq.pl $< $@ 0.25 0.5

%_aa_freq.tsv: CountAAFreq.pl %_aa.nxs
   perl $^ $@ 0.25 0.5
```

Validation

- Charts in R

Assertions

- State assumptions
- Error if assumption doesn't hold
- Catch problems with data earlier than later
- Avoid "information leakage"*

- Assertions are an old programming tool

* http://vincebuffalo.org/2012/03/08/the-beauty-of-bioconductor.html#information_leakage_and_statistics_at_every_level

Assertions

```
HVTN505.renamed.fa: HVTN505.fa
   rename-seqs < $< > $@
   [ -z `grep '^>' $@ | grep -E --invert \
        '^>505\.\d{4}a_(WG|RH|LH)\d{2}'` ]
```

`Makefile` gotchas

- Hard tabs vs. spaces
- Force updates after changing `Makefile`
- Change default behavior on error:

```
SHELL := /bin/bash
export SHELLOPTS := errexit:pipefail
.DELETE_ON_ERROR:
```

Parallel processing

How to do more than one thing at a time

What can be parallelized?

- Easiest is separate input files
- Single input files can be split up

- Independent tasks
- All prerequisites available

```
%_aa.fa: %_na.fa
   transeq -sequence $< -outseq $@ \
           -frame 1 -clean

%.nxs: %.fa
   muscle -quiet < $< | fasta2nexus > $@

# Keep intermediate alignments, for speed
.PRECIOUS: %.nxs

%_aa_freq.tsv: %_aa.nxs
   perl CountAAFreq.pl $< $@ 0.25 0.5
```

Parallelizing with make

gag_na.fa

env_na.fa

```
make gag_aa_freq.tsv env_aa_freq.tsv
make {gag,env}_aa_freq.tsv
```

Parallelizing with `make`

- What if you had those files for 20 patients?

Pt100_gag_na.fa, Pt100_env_na.fa

Pt101_gag_na.fa, Pt101_env_na.fa

…

```
make Pt{100,101,…}_{gag,env}_aa_freq.tsv
```

Parallelizing with make

- What if you had those files for 20 patients?

Pt100_gag_na.fa, Pt100_env_na.fa

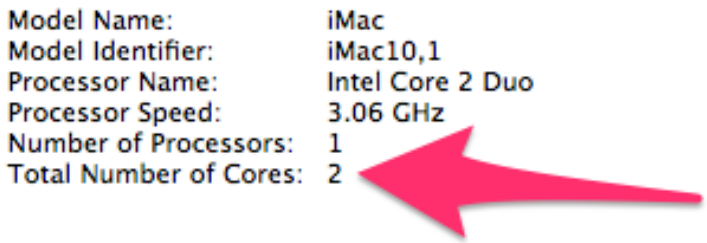
Pt101_gag_na.fa, Pt101_env_na.fa

…

```
make --jobs=24 \
  Pt{100,101,…}_{gag,env}_aa_freq.tsv
```

How many cores?

- Apple menu → About This Mac → More Info → System Report → “Total number of cores”



- `lscpu | grep -E 'Core|Socket'`

```
Core(s) per socket:     12
Socket(s):              2
```

I’ve got 99 problems, but a `Makefile` ain’t one.

Parallel processing power tools

No `Makefile`? No problem.

```
for file in *.fasta; do
  do_something -in $file -out $file.new
done
```

No `Makefile`? No problem.

```
for file in *.fasta; do
  do_something -in $file -out $file.new
done

parallel do_something -in {} -out {}.new \
  ::: *.fasta
```

No `Makefile`? No problem.

```
for file in *.fasta; do
  do_something -in $file -out $file.new
done

parallel do_something -in {} -out {}.new \
  ::: *.fasta
```

Parallel NCBI BLAST+

```
parallel \
    --halt 2 \
    --recstart '>' -N1 \
    --pipe \
    blastn \
        -task blastn \
        -db ./db/nucleotide/viroverse \
        -query - \
        -outfmt 6 \
        -max_target_seqs 25 \
    < input.fa \
    > results.tsv
```

Parallel NCBI BLAST+

```
parallel \
    --halt 2 \
    --recstart '>' -N1 \
    --pipe \
    blastn \
        -task blastn \
        -db ./db/nucleotide/viroverse \
        -query - \
        -outfmt 6 \
        -max_target_seqs 25 \
    < input.fa \
    > results.tsv
```

Parallel NCBI BLAST+

```
parallel \
    --halt 2 \
    --recstart '>' -N1 \
    --pipe \
    blastn \
        -task blastn \
        -db ./db/nucleotide/viroverse \
        -query - \
        -outfmt 6 \
        -max_target_seqs 25 \
    < input.fa \
    > results.tsv
```

Parallel NCBI BLAST+

```
parallel \
    --halt 2 \
    --recstart '>' -N1 \
    --pipe \
    blastn \
        -task blastn \
        -db ./db/nucleotide/viroverse \
        -query - \
        -outfmt 6 \
        -max_target_seqs 25 \
    < input.fa \
    > results.tsv
```

Parallel NCBI BLAST+

```
parallel \
    --halt 2 \
    --recstart '>' -N1 \
    --pipe \
    blastn \
        -task blastn \
        -db ./db/nucleotide/viroverse \
        -query - \
        -outfmt 6 \
        -max_target_seqs 25 \
    < input.fa \
    > results.tsv
```

Parallel NCBI BLAST+

```
parallel \
    --halt 2 \
    --recstart '>' -N1 \
    --pipe \
    blastn \
        -task blastn \
        -db ./db/nucleotide/viroverse \
        -query - \
        -outfmt 0 -out results-{#}.blastn \
        -max_target_seqs 25 \
    < input.fa
```

Getting these tools

- parallel
  - Installed on themis
  - I can install it elsewhere or help you do it
- make
  - Part of Apple's Xcode
  - In the Mac App Store

Resources

- Manuals
  - http://www.gnu.org/software/make/manual/make.html
  - http://www.gnu.org/software/parallel/parallel_tutorial.html
- `man make`
- `man parallel`
- Ply me with donuts, or just ask nicely