‘Large’ Data Methods-An Introduction to Scalable Statistical Genomics in Linux

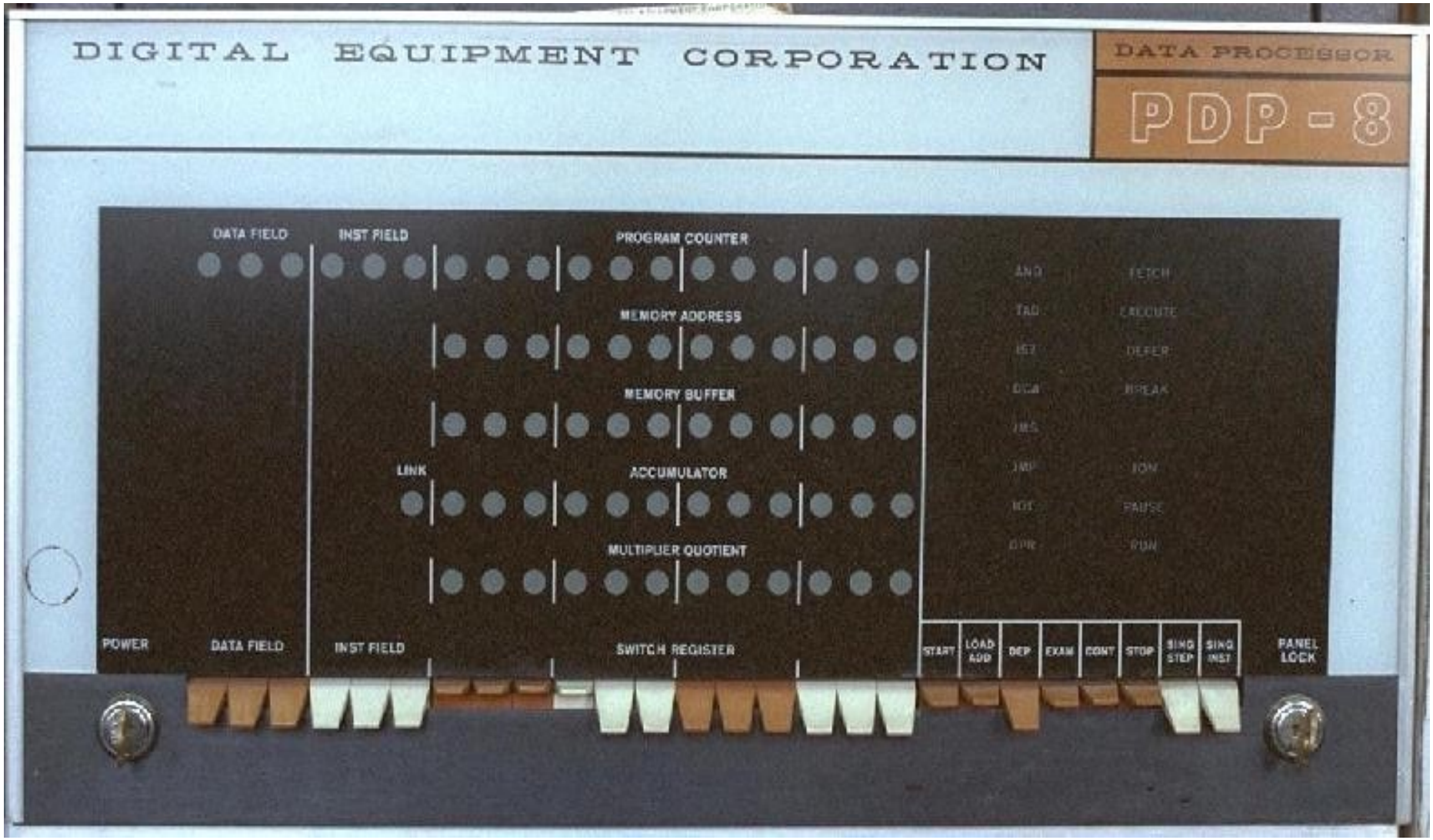
John McCallum
john.mccallum@plantandfood.co.nz
Marcus Davy
Samantha Baldwin

VISG
Virtual Institute of Statistical Genetics

digital

debian

VISG
Virtual Institute of Statistical Genetics



Statistical Genomics in 2012+

- Larger is getting cheaper and larger
- Context-specific SNP marker discovery
- Population pool methods
- Genotyping by sequencing
- Global methods
- Reference-based methods
- Largeness requires <u>scalable</u> computing
  - *device* < *desktop* < server < *cluster* < *cloud*
- Diverse tools, scripts available for UNIX (ie OSX,Linux)
- Web 2.0 options for social coding, analysis, sharing

GOALS

1. To introduce Unix as desktop and command-line environment + toolset for doing genetics
2. To explain a population genomic workflow with real data

Why Unix?

- It makes the web and our phones work
- OSX and current Linux desktops match or exceed Windows functionality
- Leading platform for scientific computing
- Secure
- Scalable *phone/desktop/server/cluster/cloud*
- Many flavours to suit your needs

freedom to work how you want, where you want, with (b)leading-edge tools

Agenda

- An intro to the Unix environment ***(1 h+)***
  - Using a modern linux desktop and command-line
  - Accessing servers and moving files
  - Documentation and help
  - Exploring file contents and formats
- Running Analyses -Variant detection & analysis ***(1 h)***
  - Getting and using third-party scripts and executables
  - Using bwa read mapper+samtools for variant analysis
  - Comparisom among populations with Popoolation2
- Visualization ***(1 h)***
  - Layering data in IGV
  - with reshape2+ ggplot2 in R
  - R and Galaxy ***(1/2 h)***
  - Rsamtools interfaces
  - Marker design in Galaxy

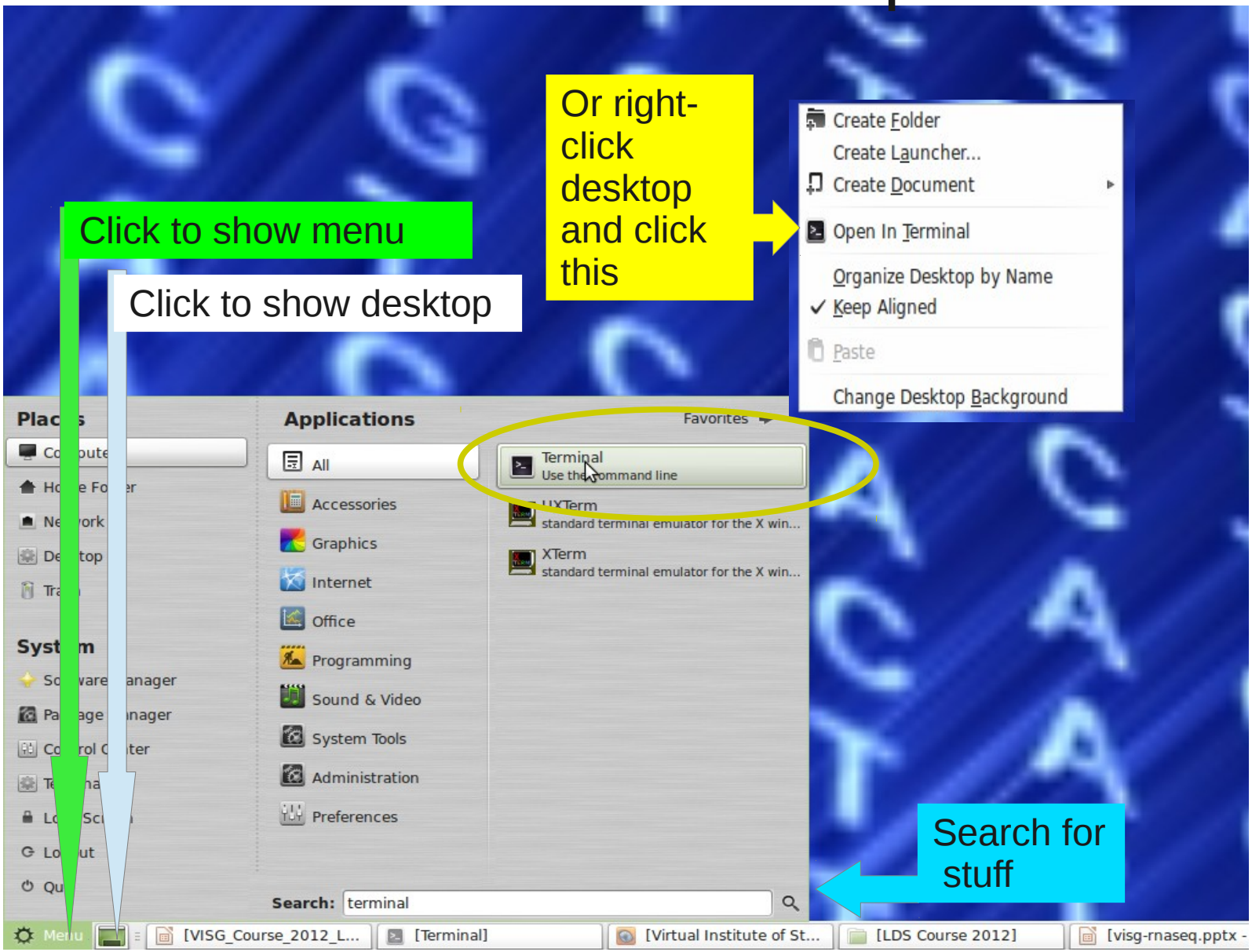
The Data

- Fragmented, barcoded PCR amplicons from flowering candidate genes from 7 onion populations, 2x 1/16 454 Ti plate
- Parallel-Tagged Sequencing Meyer et al 2008 -now rendered obsolete but basically same workflow from eg Nextera XT
- Raw data
  - Genomic Reference fasta sequence (homozygous reference)
  - Annotation of reference sequences from gmap (gff3)
  - Fastq files, one per population per plate segment
- Workflow
  - Read mapping with BWA SW
  - Manipulation and SNP calling with Samtools
  - Population genetic analyses with PoPoolation2
  - Visualization with IGV and R (reshape/ggplot2)
  - PCR Marker design in Galaxy

Biological Question

Which SNPs in these genes show evidence of strong population differentiation?

Accessing Terminal from Debian Linux Mint Desktop



> The bash shell

- 'Bourne-again-shell'
- A command-line (CL) interface to operating system
- a command interpreter
- **Command -option** <*value*> **argument(s)**
- Inputs and outputs from files or stdin/out

http://manuals.bioinformatics.ucr.edu/home/linux-basics

Exercise-Shell Orientation

`whoami`	Who are you?
`pwd`	Where are you?
`ls`	List the files
`ls -l`	Long listing
`cd /`	Go to root directory
`ls`	List the files
`ls -la ~`	Long listing of home plus hidden files
`cd`	Go Home
`ls -l .`	List files in this dir
`ls ..`	List the files in parent dir

Get the Workflow and Data

- **Browse**:https://github.com/cfljam/VISG-course-2012

```
git clone https://github.com/cfljam/VISG-course-2012.git
## get the archive
ls -l ##get a directory listing
ls -R * ##recursively list directory contents
tar -tvf VISG_LDS.tgz ##check the archive
tar -xvf VISG_LDS.tgz ##unpack it
ls -l ls -l 00.raw/ ##check there are data files
```

- Browse the directories and archives with the file manager

Security==Permissions

- Permissions protect you and the system from you and others
-  your sysadmin :-)
- If you must....on this install
  - **sudo su ##to become admin**
  - **sudo <do-something requiring admin rights> ## one-off**
- Each file/directory has
  - Owner
  - Group membership
  - Permissions
  -

Permissions links owner group size date name

```
-rw-r--r-- 1 visg_user visg      875 Oct 17 11:36 README.md
drwxr-xr-x 2 visg_user visg     4096 Oct 17 11:36 supplementary_QC
drwxr-xr-x 2 visg_user visg     4096 Oct 17 11:36 supplementary_samtools_usage
-rw-r--r-- 1 visg_user visg 22160684 Oct 17 11:37 VISG_LDS.tgz
```

Help!

`help` #list shell commands

`help cd` #help for cd command

`man ls` #read the man pages for ls, q to exit

`<command>` #may give help eg bwa

`<command>` --help # for some programs

Google..

Exercise CL Navigation & Wildcards

`cd VISG-course-2012/00.raw`	Move to raw data dir
`ls P<tab>`	List with f*ilename completion*
`<tab> <tab>`	*All the options*
`ls Pool1_BARCODE*`	*List the Pool1 files*
`ls Pool1_BARCODE?.fastq`	*List the Pool1 fastq files*
`ls Pool?_BARCODE[468].fastq`	*List all barcode 4,6,&8 reads*
`history`	*see all the history*
`history | tail`	*Last few items*
`<up arrow>`	*Back in history*
`<down arrow>`	*Forward in history*
`< mouse double click>`	*Copy*
`<shift ins>`	*paste*

Exercise-Explore

`ls *.fastq` #***list to stdout***

`ls *.fastq > somefile` #***redirect to file***

`cat somefile` ***#to stdout***

cat somefile | head # ***pipe file to head***

cat > somefile #***read from stdin/ctrl d save***

In Unix...everything is a File

Regular files-human readable text

Directories

Executable files

Compiled

Special text files

Symbolic links -'shortcuts'

Exercise -View, Browse and Search

- `cd /VISG/00.raw ##move to  raw data dir`
- `head Pool1_BARCODE2.fastq ##see top of file`
- `tail Pool1_BARCODE2.fastq ##see bottom of file`
- `less Pool1_BARCODE2.fastq ##view with the less pager`
  - **h help screen**
  - **g *top of file***
  - **G bottom of file**
  - **/<pattern> *search for pattern***
  - **q quit less**
- `grep @GYSS Pool1_BARCODE2.fastq | head ##get readnames`
- `grep -c @GYSS Pool1_BARCODE2.fastq ##count reads`

Gotchas-Symbols, Whitespace , Names

- Stick to `A-Za-z0-9_` for naming files
- Non-printing characters
  - Spaces and tabs
  - Line endings: Unix=LF, Win =CR/LF
- In shell environment many characters have special meaning e.g..

comment
#! *shebang*
> *redirect to*
< *input from*
| *pipe*
$ variable expression

/ path delimiter
\ quote next character
" strong quote
“” weak quote
` ` evaluate

Formats

- Input and outputs should stick to standard common formats
- Read the specifications!!!!!
- Fasta – Raw nucleotide/peptide format
- Fastq – Raw sequence information + quality
- 'Stream Formats'
  - Sam/Bam format – Sequence Alignment/Map format
  - GFF – General feature format-annotations
- Tools for format conversion & filtering
  - Unix tr, awk, sed, perl
  - Programming Libraries Python, Perl, R etc
  - Galaxy

Fasta format

- Text-based format for storing nucleotide/peptide sequence(s)
- Restricted to *IUPAC alphabet* letters

No spaces!

```
>gi|63055|emb|V00385.1| Part of the chicken ovalbumin X gene
ACTGTGTCTTAGCACTCACTGCTTTGCTTCCTTCTTACAGGACAGATCAAAGATTTGCTTGTATCAAGCT
CCACTGATCTTGATACAACGCTGGTCCTTGTTAATGCCATCTACTTCAAAGGGATGTGGAAGACAGCATT
TAATGCAGAAGACACTCGAGAAATGCCCTTCCATGTAACAAAGGTAGGGGACGTAGTCACCGCTTCTGGG
...
```

Header

Content

Newline wrap usually at 60 - 80 characters

http://en.wikipedia.org/wiki/FASTA_format

VISG
Virtual Institute of Statistical Genetics

Fastq format

```
@HWUSI-EAS582_157:6:1:1:1501/1
NCACAGACACACACGAACACACAAAGACATGCCCATATGAAGAT
+
%.7786867:778556858746575058873/347777476035
@HWUSI-EAS582_157:6:1:1:1606/1
NCTGGCACCTTGATTTTGGACTTCCCAGCCTCCAGAACTGTGAG
+
%194898888879898836689888864899878889888588
@HWUSI-EAS582_157:6:1:1:453/1
NCTGCTTGCACCCCTGAAGTCACTGATCACATTTCAGGGTCACC
+
%/868998988888867668888986644788988413488885
@HWUSI-EAS582_157:6:1:1:1844/1
NGATTGACATTGGCAAAGAGGACAACTGATTGCAAACTTCACAC
+
%-7;::::::;86499;75574586::635:62687666887879
@HWUSI-EAS582_157:6:1:1:1707/1
NAGGCTCAGGCGCACGGCCTACATCGTCGCTGTCGGCCAAGGGG
+
```

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/ (Detail)
http://wiki.genomequest.com/index.php/NGS_Reads

Fastq format header

```
@HWI-EAS209_0025_FC427:6:1:1041:14884#ACAGTG/2
AATTTGTTTGTTTGTTGTTTATTTTTTTGTTAGTTTCGTTTGTTGTTTGTTTGGATTTCCTCTGTGTTGAGTATTT
+HWI-EAS209_0025_FC427:6:1:1041:14884#ACAGTG/2
_________QZNUSUISNQW__U^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Header
Sequence
Header
Quality

- Illumina header contains several fields

```
HWI-EAS209  Unique machine identifier
0025        Run number
FC427       Unique flowcell identifier
6   Lane  number
1   Tile number
1041        X coordinate within tile
14884       Y coordinate within tile
#ACAGTG     illumina barcode multiplexing index tag
/2  Pair number (1 or 2)
```

VISG
Virtual Institute of Statistical Genetics

Sequence alignment map format

- sam – text format
- Bam – binary version

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7  30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref 9  30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref 9  30 5H6M       *  0   0 AGCTAA *      NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC *       NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Header

Alignments

http://samtools.sourceforge.net/SAM1.pdf
http://en.wikipedia.org/wiki/SAMtools
http://samtools.sourceforge.net/
http://samtools.sourceforge.net/samtools.shtml

VISG
Virtual Institute of Statistical Genetics

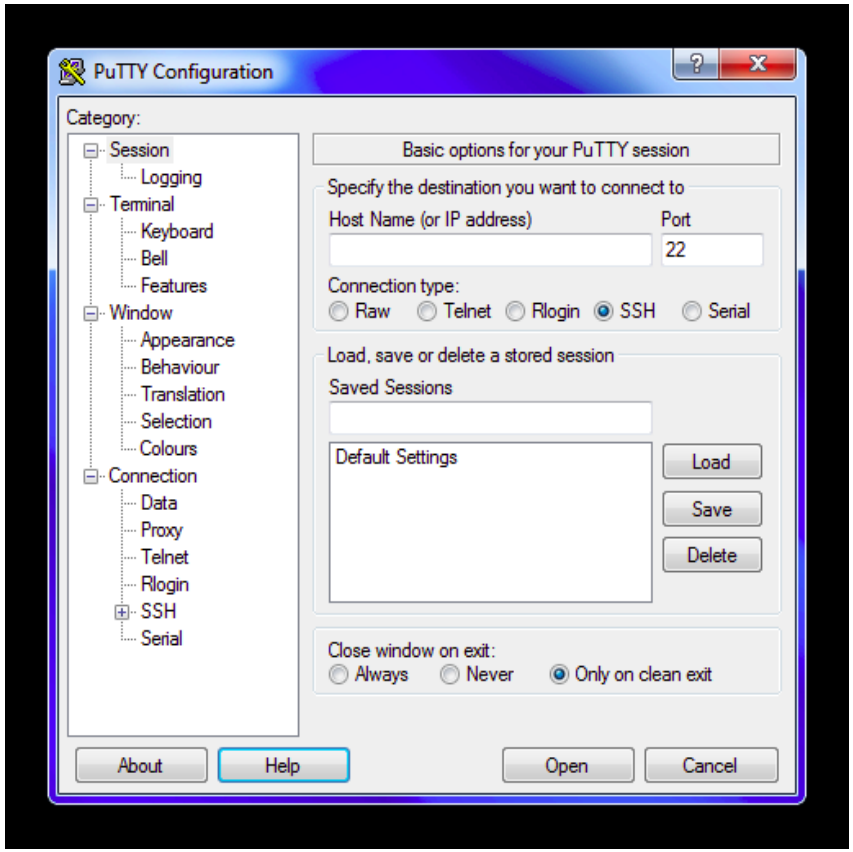
GFF3 Format

- Generic feature format
- http://www.sequenceontology.org/gff3.shtml
- Stream format-'one liners' of information about a feature

CO_Pool1_contig00004	reference	cDNA_match	355	575	100	-	.	ID=JR851107.path1;Name=JR851107;Target=JR851107 500 720;Gap=M221
SOC1_Pool2_contig00003	reference	cDNA_match	325	565	100	+	.	ID=JR848637.path1;Name=JR848637;Target=JR848637 2 242;Gap=M241
SOC1_Pool2_contig00003	reference	cDNA_match	1320	1401	98	+	.	ID=JR848637.path1;Name=JR848637;Target=JR848637 243 325;Gap=M42 I1 M40
VIN3-like_Pool2_contig00005	reference	cDNA_match	321	933	100	+	.	ID=JR853510.path1;Name=JR853510;Target=JR853510 1 613;Gap=M613
SOC1_Pool2_contig00003	reference	cDNA_match	325	565	100	+	.	ID=JR848637.path1;Name=JR848637;Target=JR848637 2 242;Gap=M241

Important Freeware Tools

- Linux -Live CDs/DVDs/USBs-use as installers
- All platforms-Oracle VirtualBox
- OSX
  - Terminal
- Windows
  - Putty
  - Xming
  - Winscp
  - Notepad++



Exercise-SSH and SCP

```
ifconfig | grep 'inet addr'

ping <their IP address>

ssh visg_user@<IP address>

scp visg_user@<IP address>:/VISG/00.raw/reference.fasta ~
```

- Get your IP address,swap with a partner
- Check you can reach their IP address
- SSH to each others machine as VISG_USER
- Copy a file to your home dir

>Feeling Overwhelmed Yet?

http://blogs.scientificamerican.com/guest-blog/2012/10/15/the-1000-genome-is-here-are-we-ready/

To keep sane, use approaches that are

- Scalable & shareable
- Open source
- Reproducible
- Documented
- Identifiable
- Disciplined

Reproducibility Questions

Where did these files come from?

What commands and options did I use?

What was I thinking?

How can I re-use this pipeline?

http://reproducibleresearch.net

http://cran.r-project.org/web/views/ReproducibleResearch.html

Community Support: Social Coding and Version Control

Code Sharing

- Sourceforge
- Github
- Google Code
- MyExperiment
- Galaxy Toolshed

Version Control

- SVN
- Git
- *Both* supported in RStudio
- Worth learning about soon!

Reproducibility -A Simple Approach

- Use one directory per atomic step, with an informative name
- Prefix directory names with numeric order
- Keep filenames consistent and informative
- Paste step commands into an executable shell script file that will enable re-creation
- Document stuff in
  - In-line comments *##some comment*
  - Plain text README , with formatting in Markdown if desired
- Version control using git

Where are My Tools: Scripts and Executables

```
echo $PATH ##where to look for executables

which bwa ##where is the bwa prog?

ls -l /usr/bin/bwa ##note x in the permissions

cd 05.reference/  ##move into a dir with run.sh

ls -l run.sh     ##note x in the permissions

cat run.sh   ##note shebang, denoting sh(bash) as
interpreter
```

Important script interpreters

- sh (normally bash)
- Rscript (R)
- Python
- Perl

Exercise-Make a Shell Script

- `cd ~`
- `mkdir test`
- `cd test`
- `cat > hello_unix.sh`

```
#!/bin/sh
echo “hello ”
whoami
echo “number of lines in file listing is:”
ls -l .. | wc -l
<ctrl-d>
```

- `ls -l`

- Move to HOME
- Make a dir
- Move into it
- Redirect to file (or use editor)
- Enter each line, then return
- Ctrl-d to finish

- Check you have created a file, and its permissions

Exercise-Run/edit a Shell Script

```
cat hello_unix.sh
sh hello_unix.sh
./hello_unix.sh
chmod +x hello_unix.sh
./hello_unix.sh
cat >> hello_unix.sh
echo “another command”
<ctrl-d>
nano hello_unix.sh
cd ..
rm -r test
```

- View the contents
- Run using sh
- Wont work
- Make it executable
- Should work
- Append to the file

- edit using nano
- Move up a level
- Delete the directory

Getting Programmes & Scripts

- sudo apt-get install bwa #on Debian/Ubuntu Linux
- wget curl <URL to file>
- ftp ftp://somewhere.org/file.tgz
- Check out from repository
  - svn checkout <URI>
  - git clone <URI>
- May require
  - Unzipping tar, gzip, GUI archive manager
  - Compilation *configure/make/make install*
  - Putting in your PATH or system PATH (as admin)

BWA Aligner

BWA= Burrows-Wheeler Aligner

Produces gapped alignment to reference

http://bio-bwa.sourceforge.net/

BWA-SW for reads > 200 bp

Need to index reference first

Produces output in SAM format
http://samtools.sourceforge.net/

(*I-am-not-an-expert-in*) SAMTOOLS

http://samtools.sourceforge.net/Program: samtools (Tools for alignments in the SAM format)

Version: 0.1.18 (r982:295)

```
Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         idxstats    BAM index stats (r595 or later)
         fixmate     fix mate information
         flagstat    simple stats
         calmd       recalculate MD/NM tags and '=' bases
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
         reheader    replace BAM header
         cat         concatenate BAMs
         targetcut   cut fosmid regions (for fosmid pool only)
         phase       phase heterozygotes
```

Manipulating 'Data Streams'

- Most bioinfo formats are 'streams' of columnar data, one line per element
- Can be manipulated and filtered in multiple ways
  - Unix tools
  - Bioinfo tools such as Samtools
  - Galaxy
  - If you must...spreadsheets

File and Stream Munging

There are many power tools for one-liner, or script-based file filtering and reformatting available in Unix

Most use <u>regular expressions to define patterns</u>

sed 's/SOC1/VISG/' Pool2_Pop2.sam | head #*change all instances of SOC1 to VISG*

awk '!/^@/ && $5 > 20 {print $1, $5}' Pool2_Pop2.sam | head #*print read name and MAP quality if filter MAPQ > 20*

cut -f5 Pool2_Pop2.sam | sort -n > temp # *write numeric sorted MAP quality values into file*

grep SOC1_Pool2_contig00003 Pool2_Pop2.sam | wc -l #*count the number of SOC1 alignments in the file*

Perl Very powerful <u>regular expressions</u>

Popoolation

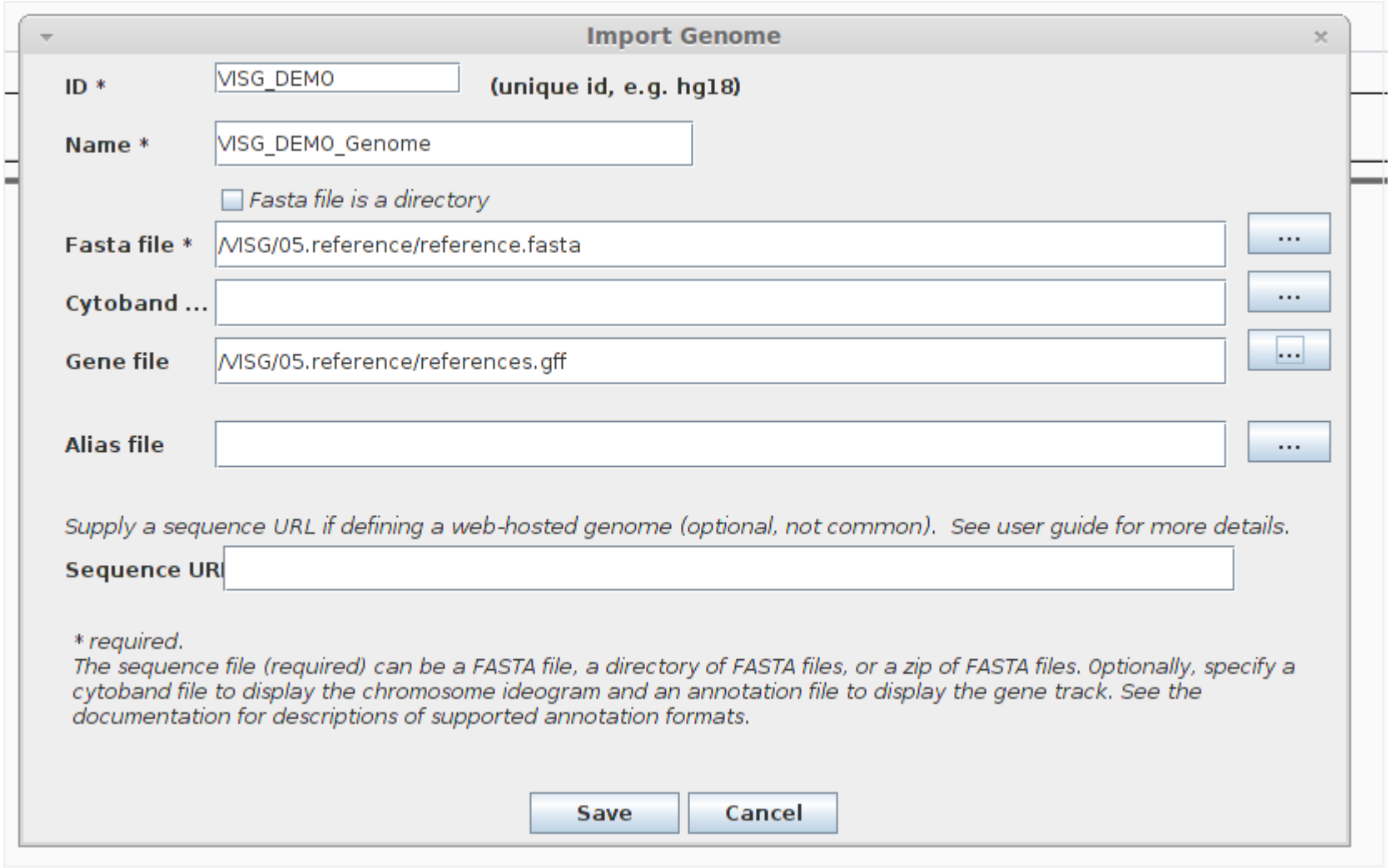
- a collection of tools to facilitate population genetic studies of next generation sequencing data from pooled individuals
- Popoolation
  - A pipeline for analyzing pooled next generation sequencing data for single populations.
  - Tajima's Pi, Watterson's Theta and Tajima's D
  - http://code.google.com/p/popoolation/
- Popoolation2
  - Allows analyzing the population frequencies of SNPs from two or more populations.
  - F_{st}, Fisher's exact test, Cochran-Mantel-Haenszel test
  - http://code.google.com/p/popoolation2/

Getting PoPoolation

- Browse, download, unpack with archive manager
- or....from CL
- wget http://popoolation2.googlecode.com/files/popoolation2_1201.zip
- unzip popoolation2_1201.zip
- or.....from CL
  - apt-get update ##update package information
  - apt-get install svn ##install SVN
  - svn checkout http://popoolation2.googlecode.com/svn/trunk/popoolation2 ##check out copy

Start IGV and Set Up a Genome

```
igv & ##start up IGV in the background
```



IGV

- Layer up *file->load from file*
  - Vcf
  - gff
  - igv
  - Bam
- May need to sort and index
  - *file->run igvtools → command-> sort/index*

Getting & Compiling Software-Github

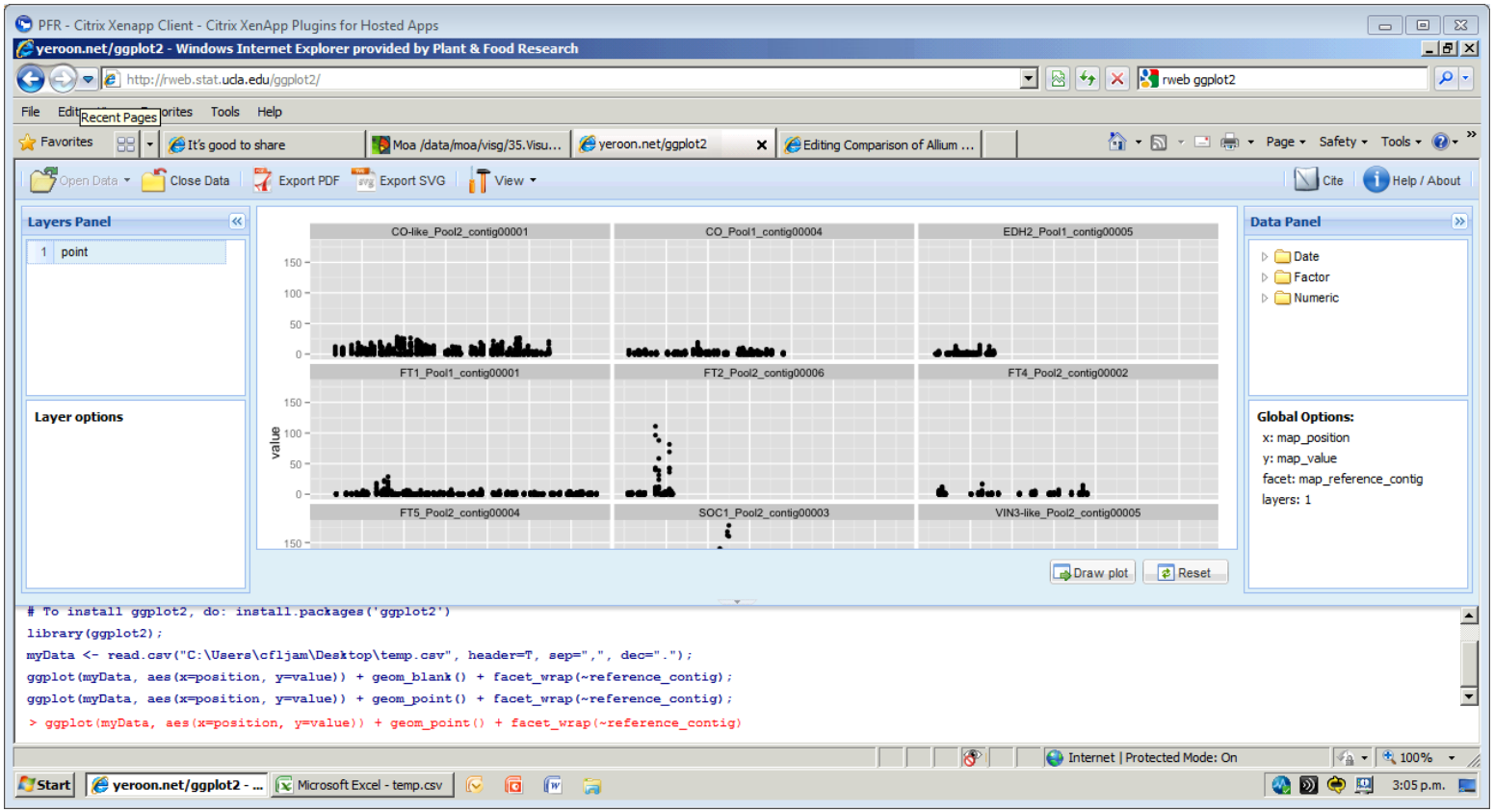
```
cd ~/Downloads/
git clone https://github.com/lh3/wgsim.git
cd wgsim
less README ##read the instructions
gcc -g -O2 -Wall -o wgsim wgsim.c -lz -lm
#compile
echo $PATH ##check your path
cp wgsim /usr/local/sbin ##copy to PATH
wgsim  ##check it works, read help
```

reShape2

- Flexible rshaping of data
- Especially valuable for turning 'wide' into 'long' (stream-oriented') data
- http://had.co.nz/reshape/
- http://www.jstatsoft.org/v21/i12

ggplot2

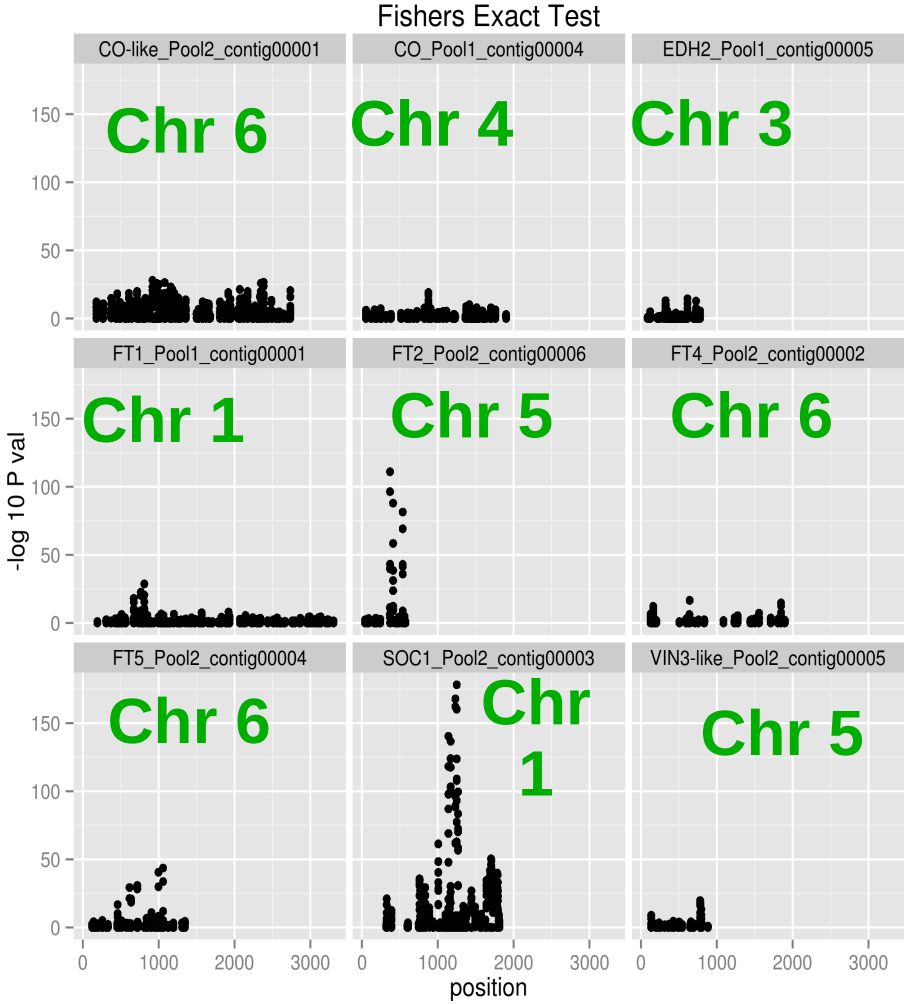
http://www.stat.ucla.edu/~jeroen/ggplot2/



http://had.co.nz/ggplot2/ http://docs.ggplot2.org/current/

Conclusions & Directions

- AcFT2 and AcSOC1 show strong differentiation among populations
- <u>Both</u> peaks are adjacent to non-synonymous SNPs
- Mapped most loci
- Physiological data suggests AcFT2 associated with bulbing response
- Bolting QTL on Chrom 1,3,6
- Trying to map AcSOC1 in relation to chrom 1 QTL



Acknowledgements

- Richard Macknight, Robyn Lough (Univ. Otago)
- NZGL
- Meeghan Pither-Joyce, Kathryn Wright, Martin Shaw, Roopa Revanna (PFR Onion Group)
- Mark Fiers (Plant & Food)
- MSI, MBI and MBIE