

Transcriptome Variation Under Dynamic Growth Conditions

Kevin Murray

Borevitz and Pogson Labs

September 24, 2013

Big Picture Question

- ▶ A set of broad hypotheses:
 - ▶ In nature, plants acclimate to repeated stresses
 - ▶ After acclimation, plants will be more tolerant to stresses
 - ▶ We can recreate this acclimation in lab using dynamic growth conditions

Terminology

- ▶ **Transcriptomics:** study of global gene expression
- ▶ **RNAseq:** transcriptome quantification by sequencing
- ▶ **Pipeline:** series of software which turns data into results
- ▶ **QTL Mapping:** technique to associate variation in genotype to phenotype variation

Investigating altered light intensity

- ▶ Hypothesised “hardening” of plants to harsher conditions
 - ▶ Increased steady state expression of stress genes
 - ▶ Decreased induction of stress genes after stress
- ▶ Hypothesised a relative order of “hardening”
 1. Fluctuating light
 2. Excess light
 3. Sufficient light
 4. Standard growth conditions

Aims

1. Design & implement dynamic growth conditions
2. Develop a pipeline of software to analyse RNAseq datasets
3. Generate phenomic and transcriptomic QTL mapping datasets from plants grown under dynamic light conditions
4. Determine effect of light intensity on transcriptome under dynamic light conditions

The growth condition dilemma

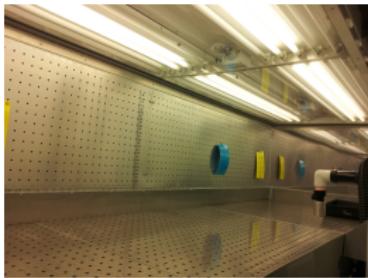


- ▶ Plants grow in nature (mostly)

The growth condition dilemma



- ▶ Plants grow in nature (mostly)

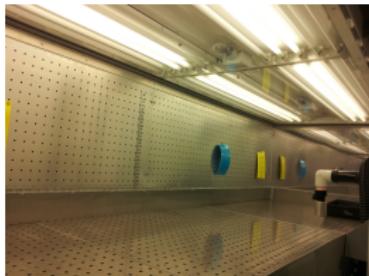


- ▶ Scientists work in labs (mostly)

The growth condition dilemma



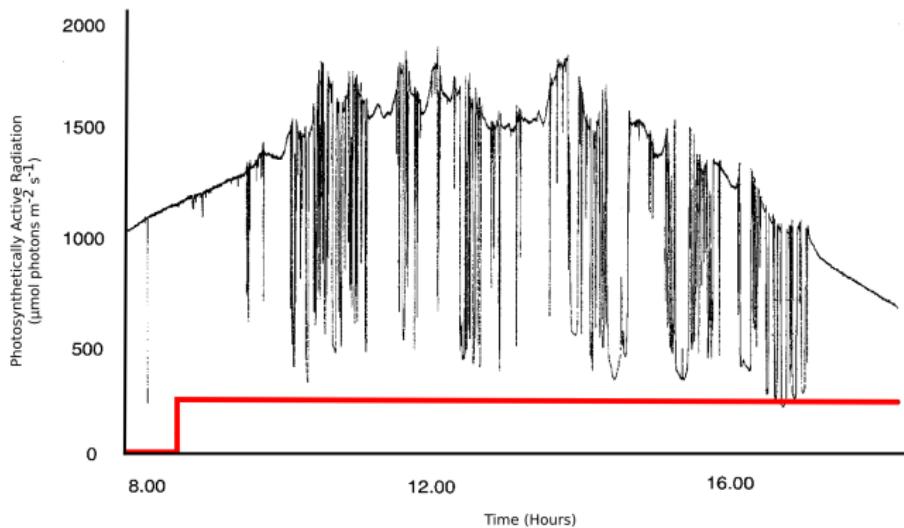
- ▶ Plants grow in nature (mostly)



- ▶ Scientists work in labs (mostly)

- ▶ Aim to “merge” elements of these two scenarios

Introducing the SpectralPhenoClimatron



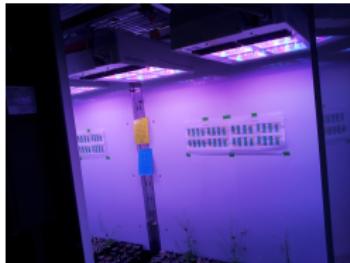
- ▶ Simulate regional climates
- ▶ Model diurnal and circannual trends of climate

(Külheim, Ågren, and Jansson 2002)

Controlling the SpectralPhenoClimatron

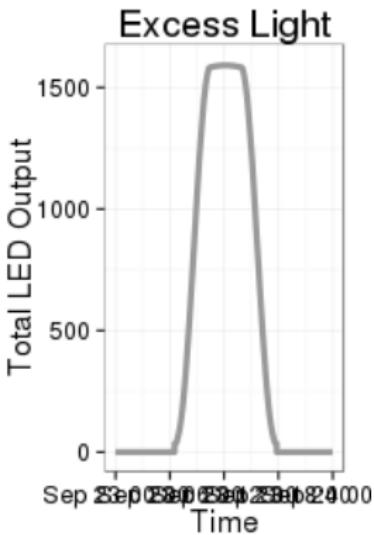
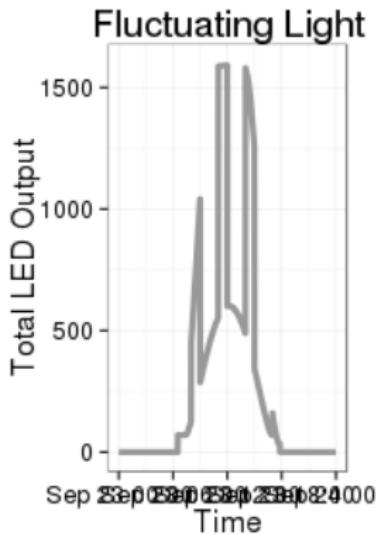
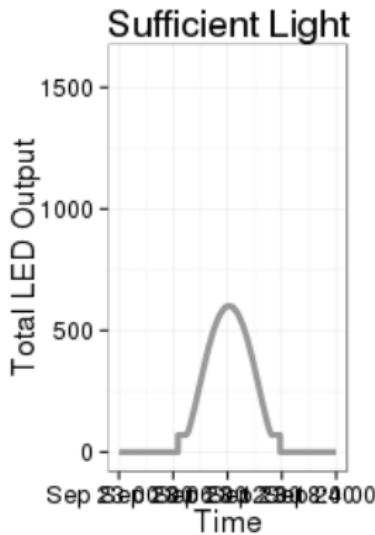
- ▶ Use model simulation to drive actual growth cabinet conditions
- ▶ Need software “glue” to stick bits together
- ▶ Wrote spcControl
 - ▶ 750 lines
 - ▶ 134 minor, 16 major versions

Movie:



Investigating altered light intensity

- ▶ Within a simulated climate, modify light intensity
- ▶ Create 3 new conditions:
 - ▶ Sufficient light
 - ▶ Fluctuating light
 - ▶ Excess light



Investigating altered light intensity

- ▶ Hypothesised “hardening” of plants to harsher conditions
 - ▶ Increased steady state expression of stress genes
 - ▶ Decreased induction of stress genes after stress
- ▶ Hypothesised a relative order of “hardening”
 1. Fluctuating light
 2. Excess light
 3. Sufficient light
 4. Standard growth conditions

Aims

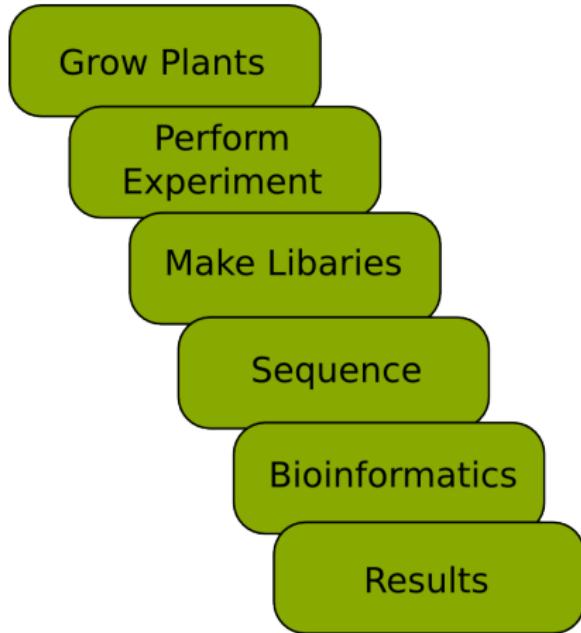
- ▶ Design & implement dynamic growth conditions
- ▶ Develop a pipeline of software to analyse RNAseq datasets
- ▶ Generate phenomic and transcriptomic QTL mapping datasets from plants grown under dynamic light conditions
- ▶ Determine effect of light intensity on transcriptome under dynamic light conditions

What's a pipeline?

- ▶ A collection of software to turn raw data into results

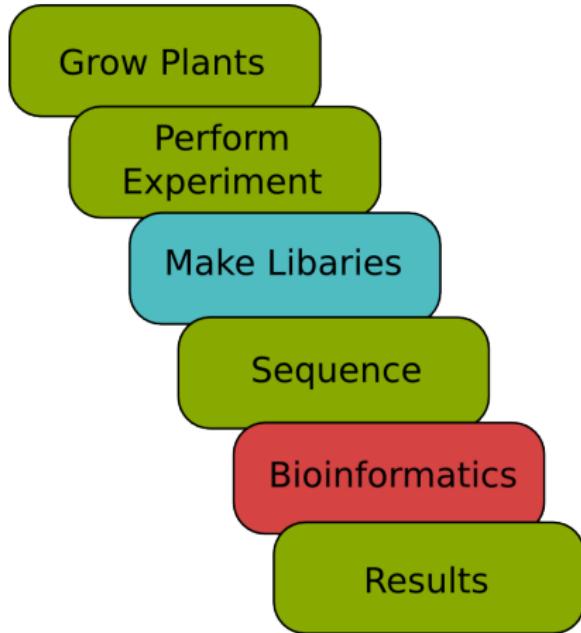


How does RNAseq work?



- ▶ Assay **ALL** expression in your tissue
- ▶ Unbiased, as quantitative as qPCR
- ▶ Becoming cheaper and easier

How does RNAseq work?



- ▶ Will focus on two areas of improvement
 - ▶ Making RNAseq library prep. cheaper & higher throughput
 - ▶ Making RNAseq data analysis easier & faster

The Cazzonelli Button

*"Can't there just be a
'do my bioinformatics'
button?"*

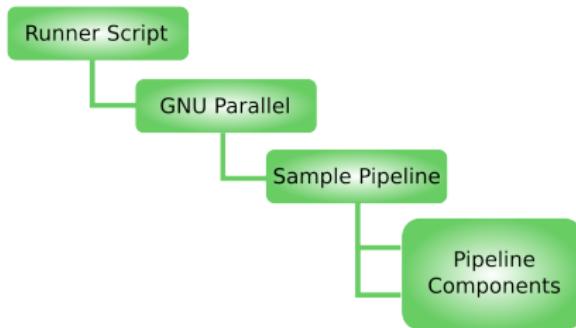
*Well, we laughed at you
Chris, but its now
almost that easy!*

FID	USER	FRI	MI	SINT	RES	SHR	CPR%	REDS	TIMES	Command
1	[REDACTED]						29.6%	7	[REDACTED]	
2	[REDACTED]						12.4%	8	[REDACTED]	
3	[REDACTED]						47.0%	9	[REDACTED]	
4	[REDACTED]						34.6%	10	[REDACTED]	
5	[REDACTED]						35.5%	11	[REDACTED]	
6	[REDACTED]						40.7%	12	[REDACTED]	
Max	[REDACTED]									
Sum	[REDACTED]									
58847	kevin	20	0	11664	108000	836	R 81.6	0.8	30:09.05	subread-align -i /home/kevin/wd
7181	kevin	20	0	11664	108000	836	R 81.6	0.8	34:12.92	subread-align -i /home/kevin/wd
7306	kevin	20	0	11664	108000	836	R 81.6	0.8	34:17.36	subread-align -i /home/kevin/wd
7426	kevin	20	0	11664	108000	836	R 80.0	0.8	33:37.09	subread-align -i /home/kevin/wd
7367	kevin	20	0	11664	108000	836	R 78.0	0.8	33:56.25	subread-align -i /home/kevin/wd
7237	kevin	20	0	11664	108000	836	R 76.0	0.8	35:06.97	subread-align -i /home/kevin/wd
7421	kevin	20	0	11664	108000	836	R 74.0	0.8	35:16.15	subread-align -i /home/kevin/wd
6827	kevin	20	0	11664	108000	836	R 72.0	0.8	35:26.26	subread-align -i /home/kevin/wd
5890	kevin	20	0	11664	108000	836	R 72.0	0.8	36:40.15	subread-align -i /home/kevin/wd
7704	kevin	20	0	11664	108000	836	R 43.0	0.8	35:47.01	subread-align -i /home/kevin/wd
7064	kevin	20	0	11664	108000	836	R 34.0	0.8	35:50.09	subread-align -i /home/kevin/wd
7250	kevin	20	0	49452	7408	2824	S 26.0	0.8	11:39.82	ssh -c blowfish gduserv rsync
3867	kevin	20	0	18620	1344	856	D 18.0	0.0	0:12.12	samtools view -5 u-align/Samp
7249	kevin	20	0	18972	1592	884	D 7.0	0.0	3:39.56	rsync -ihrwl --progress -e ssh
5913	kevin	20	0	25560	2556	1180	S 2.0	0.0	0:29.35	httpd
5886	kevin	20	0	25878	3536	1424	S 2.0	0.0	0:30.00	httpd
1	[REDACTED]									1/12 [2]
6002	root	20	0	21630	304	384	S 0.0	0.0	0:00.35	udevd --daemon
801	root	20	0	19468	292	288	S 0.0	0.0	0:00.00	/sbin/getty 38400 ttty2
1297	pete	20	0	52896	88	88	S 0.0	0.0	0:00.00	/usr/bin/gnome-keyring-daemon
2277	root	20	0	19124	288	292	S 0.0	0.0	0:06.41	/sbin/rpcbind -w
2309	statd	20	0	23484	316	312	S 0.0	0.0	0:00.00	/sbin/rpc.statd
2324	root	20	0	25432	0	0	S 0.0	0.0	0:00.00	/usr/sbin/rpc.idmapd
3652	kevin	9 -1	4341	1468	776	5	6.0	0.0	0:01.33	/usr/bin/pulseaudio --start
2661	root	20	0	2491	1808	812	S 0.0	0.0	0:07.00	/usr/sbin/rsyslogd -c5
2721	root	20	0	20493	1600	500	S 0.0	0.0	0:01.50	/usr/bin/avahi-daemon --system
2799	root	20	0	4251	196	196	S 0.0	0.0	0:00.71	/usr/sbin/acpid
2850	daemon	20	0	16812	164	136	S 0.0	0.0	0:00.06	/usr/sbin/rtld
3003	root	20	0	20620	304	296	S 0.0	0.0	0:11.08	/usr/sbin/cron
3024	avahi	20	0	36900	2368	680	S 0.0	0.0	0:45:40.52	avahi-daemon: running [K88Hiw0n]
3025	avahi	20	0	34176	68	36	S 0.0	0.0	0:00.00	avahi-daemon: chroot helper
3331	root	20	0	78928	1580	888	S 0.0	0.0	0:32.59	/usr/sbin/cupsd -c /etc/cups/cupsd.conf
3352	root	20	0	1581	3384	1340	S 0.0	0.0	1:54.20	/usr/sbin/NetworkManager
3387	root	20	0	1464	904	738	S 0.0	0.0	0:03.43	/usr/sbin/gpm3
3410	root	20	0	15141	1520	520	S 0.0	0.0	0:00.00	/usr/bin/avahi-daemon --local -l -i -p -c /etc/avahi/avahi-daemon.conf
3412	pete	20	0	53900	284	284	S 0.0	0.0	0:00.00	/usr/bin/gnome-keyring-daemon
3438	debian-ex	20	0	51569	368	312	S 0.0	0.0	0:01.51	/usr/bin/exim4 -bd -o3@#
3462	root	20	0	2164	2600	1084	S 0.0	0.0	12:51.96	/usr/lib/polkit-1/polkitd --no-roots
3468	colorad	20	0	4699	332	332	S 0.0	0.0	0:02.73	/usr/lib/x86_64-linux-gnu/colorad
3488	root	20	0	2164	2600	1084	S 0.0	0.0	12:51.96	/usr/lib/polkit-1/polkitd --no-roots

How To Run a Pipeline

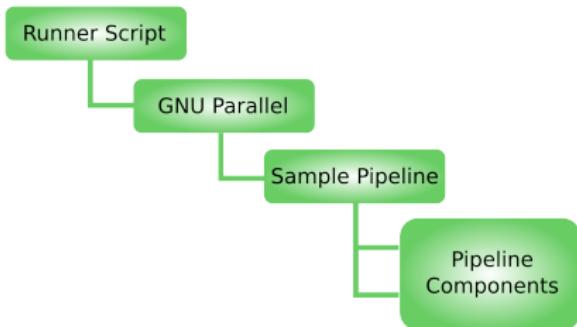
Command

```
bash runner.sh keyfile.key
```



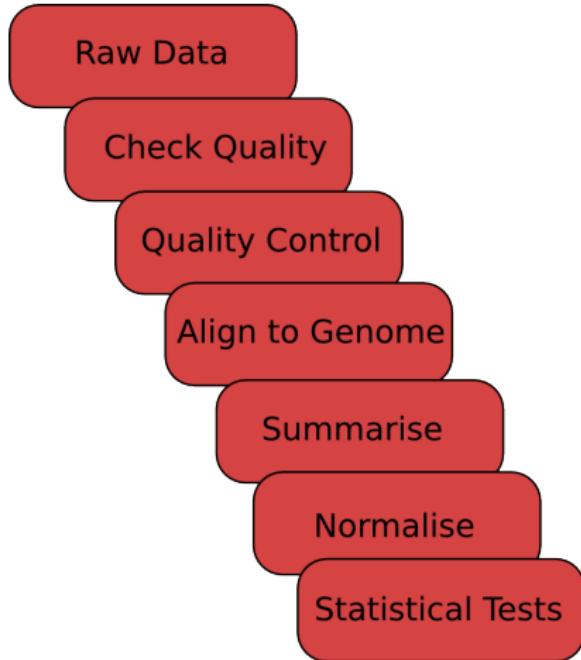
- ▶ Let's dissect that:
 - ▶ bash runner.sh
Call the runner script
 - ▶ keyfile.key
Give it the “keyfile”
- ▶ No need to run each component separately

How does that work?



- ▶ More than 1300 lines of code
- ▶ Written in bash, python and R
- ▶ 144 minor versions
- ▶ Code is on github.com
- ▶ Licensed under GNU GPL v3

RNAseq Pipeline Components



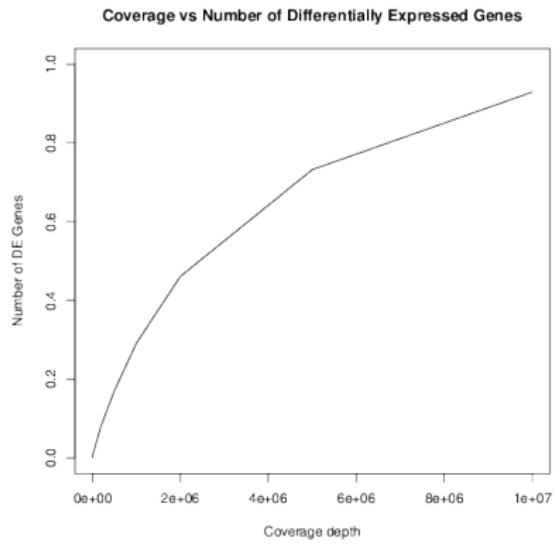
1. `fastqc`
2. `scythe`, `seqtk`
3. `subread (tophat, subjunc)`
4. `featureCounts`
5. `edgeR (DESeq, cuffdiff)`

Effect of Coverage

- ▶ Coverage depth a limiting factor in RNAseq
- ▶ Aimed to empirically determine “cutoff” depth for our study system
- ▶ Hypothesised that 2 million reads gives 20% reduction in power
- ▶ 2 million reads gives \approx 50% reduction
- ▶ Conclusion: use $> 2M$ reads per lane

Effect of Coverage

- ▶ Coverage depth a limiting factor in RNAseq
- ▶ Aimed to empirically determine “cutoff” depth for our study system
- ▶ Hypothesised that 2 million reads gives 20% reduction in power
- ▶ 2 million reads gives \approx 50% reduction
- ▶ Conclusion: use $> 2M$ reads per lane



Aims

- ▶ Design & implement dynamic growth conditions
- ▶ Develop a pipeline of software to analyse RNAseq datasets
- ▶ Generate phenomic and transcriptomic QTL mapping datasets from plants grown under dynamic light conditions
- ▶ Determine effect of light intensity on transcriptome under dynamic light conditions

QTL mapping aims

- ▶ Look for loci which control transcriptional response to abiotic stress

Growth of plants

- ▶ A QTL mapping set, parental lines and genetic controls
- ▶ 3 replicates
- ▶ Grown for 2 weeks, then dynamic growth conditions
- ▶ Assay expression before and after high light pulse treatment at 5 weeks

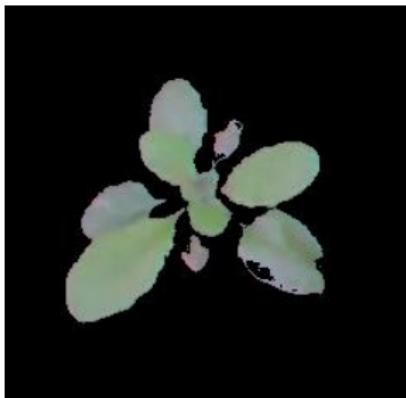
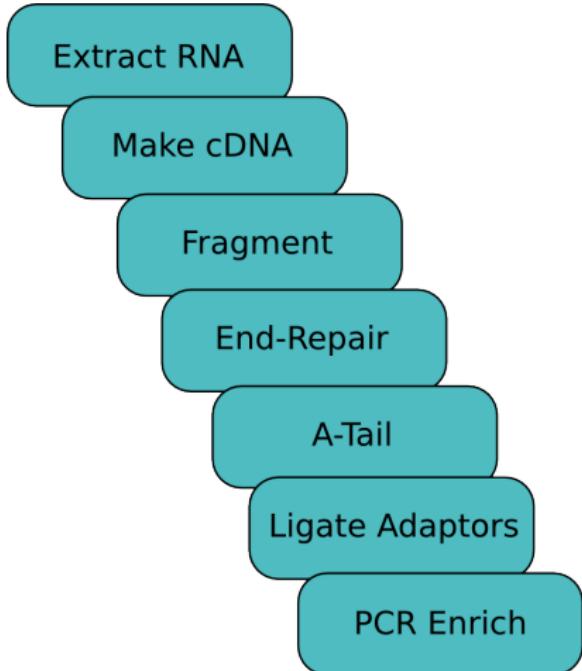


Figure: Sufficient Light



Figure: Excess Light

Cheaper, Higher Throughput Method Needed

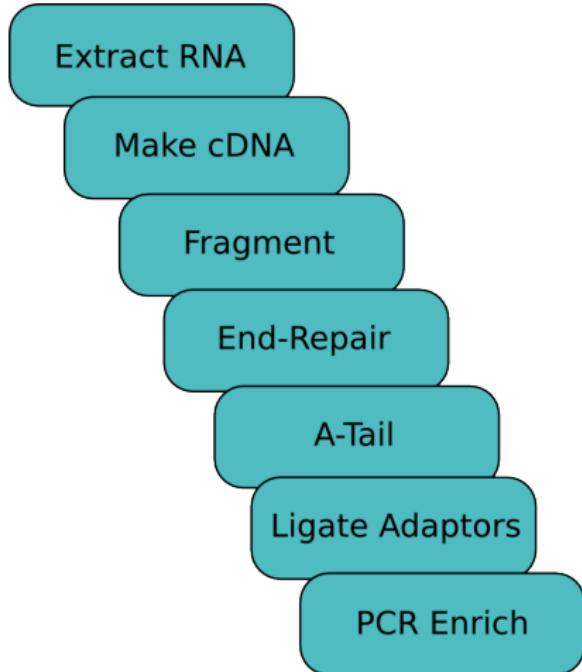


- ▶ Adapted from Kumar et al. (2012)
- ▶ On-bead SPRI protocol
- ▶ Performed in 96 well plate
- ▶ $\approx \$50$ per sample, 96 samples per lane
- ▶ Successful until final step
- ▶ Sidelined due to lack of time

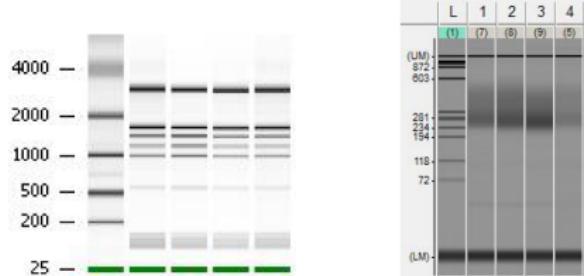
Aims

- ▶ Design & implement dynamic growth conditions
- ▶ Develop a pipeline of software to analyse RNAseq datasets
- ▶ Generate phenomic and transcriptomic QTL mapping datasets from plants grown under dynamic light conditions
- ▶ Determine effect of light intensity on transcriptome under dynamic light conditions

Illumina RNAseq Library Prep.

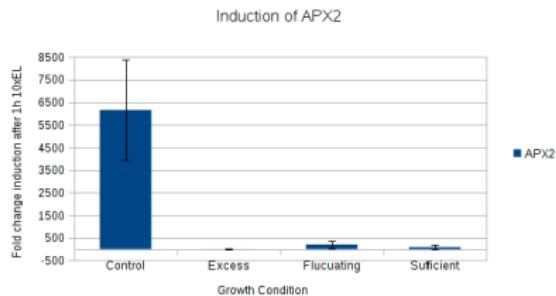


- ▶ 3-5 day protocol
- ▶ Up to 12 samples per lane
- ▶ \$240-400 per sample



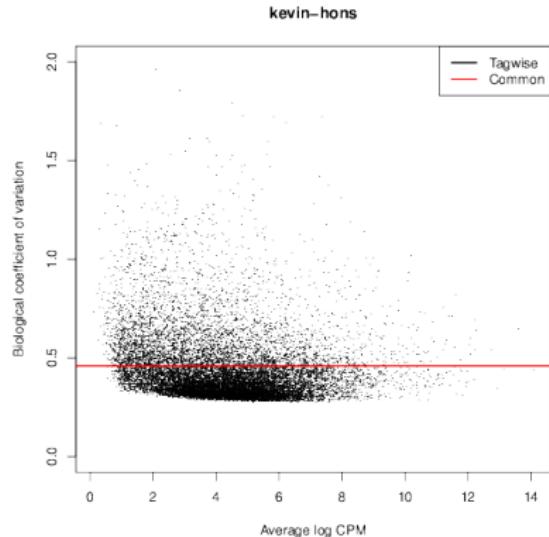
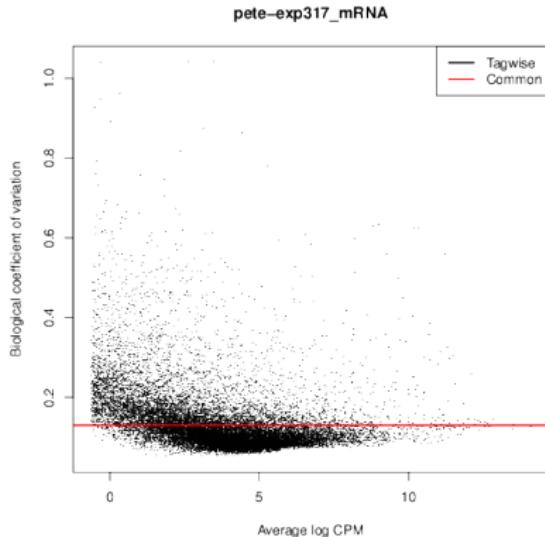
qPCR analysis tells a small story

- ▶ Examine expression of known excess light responsive genes
- ▶ Show reduced induction and increased steady state expression
- ▶ Hypotheses appear mostly correct



RNAseq shows the whole picture

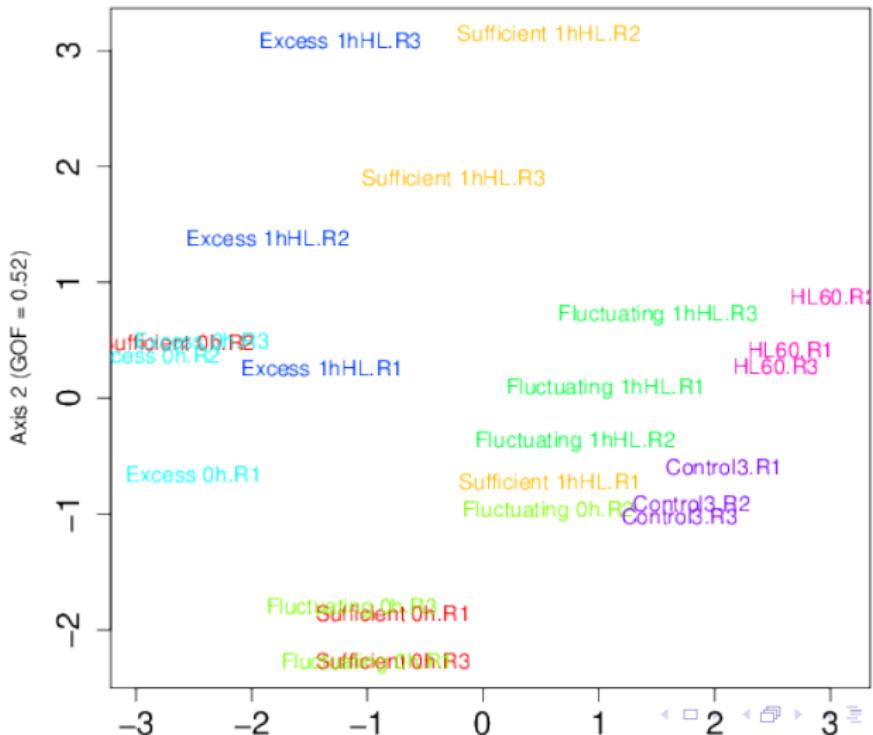
- ▶ Overall, high variation amongst samples



RNAseq shows the whole picture

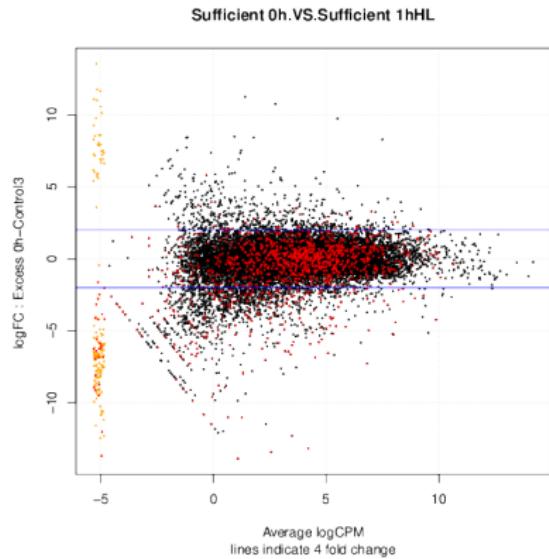
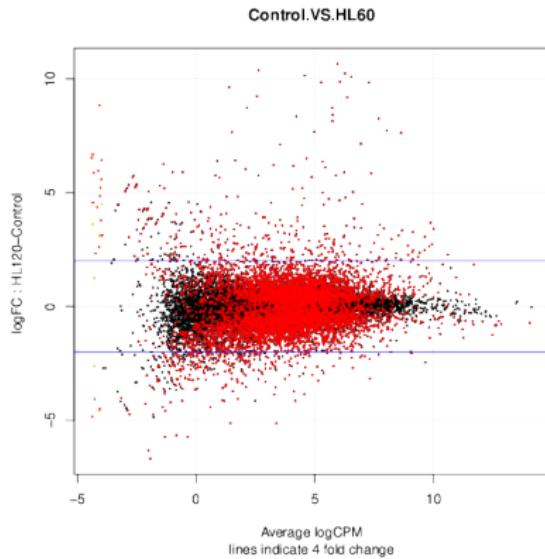
- Amongst a noisy response, a pattern emerges

Multiple-dimensional Scaling of Samples



RNAseq shows the whole picture

- Less differential expression than expected



DE table

- ▶ This is how many genes are DE
- ▶ This is what they have in common
- ▶ Analysis/interpretation is ongoing - (GSEA)

Conclusions

- ▶ Dynamic growth conditions give interesting biology
- ▶ Patterns of differential expression seen, more replicates needed
- ▶ RNAseq is here, and easier than you might think

Future Work

- ▶ Optimise 96-well RNAseq protocol
- ▶ Analyse QTL mapping set:
 - ▶ RNAseq to map QTLs for expression of stress responsive genes
 - ▶ Find phenomic traits e.g. anthocyanin accumulation
- ▶ Repeat entire experiment with improved sampling techniques - increase statistical power

Old/Dead slides

Project Overview

- ▶ What I've spent my year doing:

Project Overview

- ▶ What I've spent my year doing:
- ▶ Create & implement “dynamic growth conditions”

Project Overview

- ▶ What I've spent my year doing:
- ▶ Create & implement "dynamic growth conditions"
- ▶ Study plant growth under these conditions

Project Overview

- ▶ What I've spent my year doing:
- ▶ Create & implement "dynamic growth conditions"
- ▶ Study plant growth under these conditions
- ▶ Design & implement data analysis pipeline for RNAseq

Light Quantity

- ▶ Plants need a “happy medium” of light
- ▶ Too little = sub-optimal growth
- ▶ Too much = photooxidative damage
- ▶ Interesting model system for examining dynamic growth conditions

Light Response & Transcriptomics

- ▶ Excess light invokes known transcriptional response
- ▶ Many genes induced in excess light constitutively expressed in field
- ▶ Transcriptomics sensitive to subtle or fast changes

History of Light Transcriptomics

- ▶ Pogson, Nagashi, Karpinski, Jannsen

Aims

1. Develop a pipeline of software to analyse RNAseq datasets

Aims

1. Develop a pipeline of software to analyse RNAseq datasets
2. Determine the response of *Arabidopsis thaliana* to altered light intensity under dynamic growth conditions

Making a pipeline?



+



Making a pipeline?



+



Aim 1

Develop a pipeline of software to analyse RNAseq datasets

Aim 1

Develop a pipeline of software to analyse RNAseq datasets

- ▶ Define & overcome shortcomings in existing analysis pipelines
- ▶ Experimentally define limitations of RNAseq experimental designs

Steps in an RNAseq Analysis

- ▶ Raw sequence data needs to be:
 - ▶ Filtered for quality
 - ▶ Aligned to the genome

Steps in an RNAseq Analysis

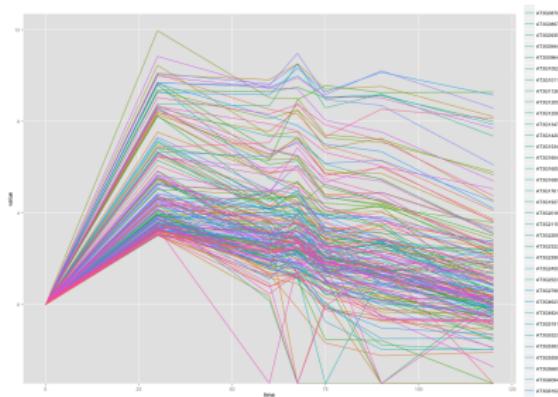
- ▶ Raw sequence data needs to be:
 - ▶ Filtered for quality
 - ▶ Aligned to the genome
- ▶ Once aligned, one needs to:
 - ▶ Quantify gene expression
 - ▶ Normalise counts
 - ▶ Perform statistical tests

Existing “Best Practice” Pipeline

- ▶ Settings optimised for non plants
- ▶ Not optimised for large/dramatic changes
- ▶ Result in hypothesised weird artefacts
- ▶ Can result in false positives or false negatives

Existing “Best Practice” Pipeline

- ▶ Settings optimised for non plants
- ▶ Not optimised for large/dramatic changes
- ▶ Result in hypothesised weird artefacts
- ▶ Can result in false positives or false negatives

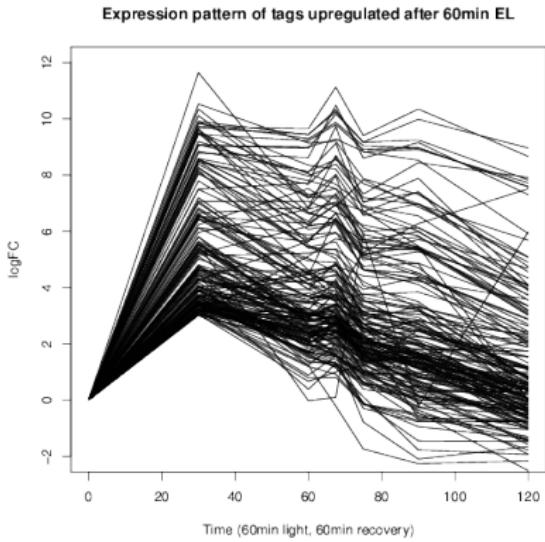


“Improved” plot

- ▶ Using edgeR, state of the art statistics
- ▶ Hypothesis was artefacts would be removed

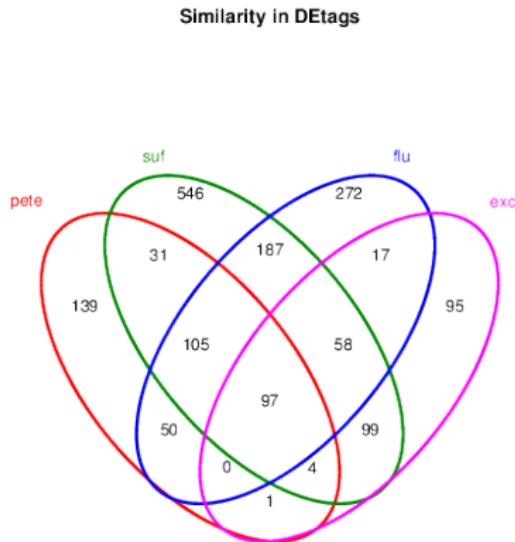
“Improved” plot

- ▶ Using edgeR, state of the art statistics
- ▶ Hypothesis was artefacts would be removed



Venn Diagrams

- ▶ Hard to see, but there is most overlap between Sufficient and Control, and Sufficient and Fluctuating



Unique objects: All = 1701; S1 = 427; S2 = 1127; S3 = 786; S4 = 371