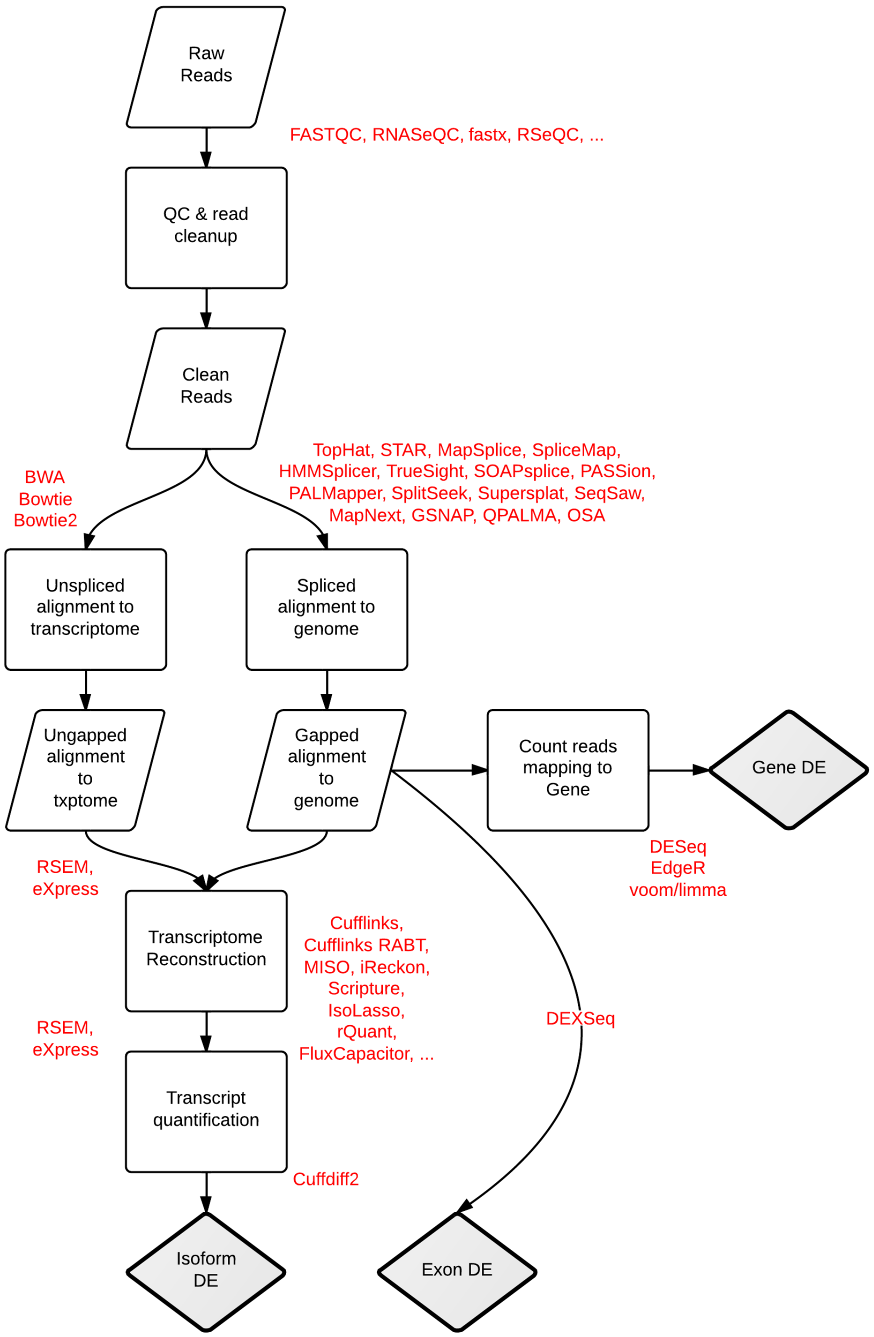
A table side chat on thinking about your NGS data statistically

NGS2013


Raw Reads
FASTQC, RNASeQC, fastx, RSeQC, ...
QC & read cleanup
Clean Reads
BWA
Bowtie
Bowtie2
TopHat, STAR, MapSplice, SpliceMap,
HMMSplicer, TrueSight, SOAPsplice, PASSion,
PALMapper, SplitSeek, Supersplat, SeqSaw,
MapNext, GSNAP, QPALMA, OSA
Unspliced alignment to transcriptome
Spliced alignment to genome
Ungapped alignment to txptome
Gapped alignment to genome
Count reads mapping to Gene
Gene DE
DESeq
EdgeR
voom/limma
RSEM,
eXpress
Transcriptome Reconstruction
Cufflinks,
Cufflinks RABT,
MISO, iReckon,
Scripture,
IsoLasso,
rQuant,
FluxCapacitor, ...
DEXSeq
RSEM,
eXpress
Transcript quantification
Cuffdiff2
Isoform DE
Exon DE

RNA-seq Workflows and Tools. Stephen Turner. Figshare. http://dx.doi.org/10.6084/m9.figshare.662782

Goals

I am not planning on trying to provide any sort of overview of statistical methods for genomic data. Instead I am going to provide a few short ideas to think about.

Statistics (like bioinformatics) is a rapidly developing area, in particular with respect to genomics. Rarely is it clear what the "right way" to analyze your data is.

Instead I hope to aid you in using some common sense when thinking about your experiments for using high throughput sequencing.

Useful references

Paul L. Auer and R.W. Doerge 2010. Statistical Design and Analysis of RNA-Seq Data. Genetics. 10.1534/genetics.110.114983
PMID: 20439781

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments BMC Bioinformatics, 11, 94. doi:10.1186/1471-2105-11-94

Designing your experiment before you start.

Sampling

Replication

Blocking

Randomization

Over all we are going to be thinking about how to **avoid Confounding** sources of variation in the data.

All of these are larger topics that are part of **Experimental Design**.

Sampling

Sampling

Replication

Blocking

Randomization

Sampling design is all about making sure that when you “pick” (sample) observations, you do so in a **random** and **unbiased** manner.

Proper sampling aims to control for unknown sources of variation that influence the outcome of your experiments.

This seems reasonable, and often intuitive to most experimental biologists, but it can be very insidious.
Whiteboard…

Sampling

Sampling

Replication

Blocking

Randomization

Replication

Sampling

R**eplication**

Blocking

Randomization

Imagine you have an experiment with one factor (sex), with two treatment levels (males and females).

You want to look for sex specific differences in the brains of your critters based on transcriptional profiling, so you decide to use RNA-seq.

Perhaps you have a limited budget so you decide to run one sample of male brains, and one sample of female brains, each in one lane of a flow cell.

What (useful) information can you get out of this?

Not much (but there may be some). Why?

Replication

Sampling

R**eplication**

Blocking

Randomization

Why?

No replication. How will you know if the differences you observe are due to differences in males and females, random (biological) differences between individuals, or technical variation due to RNA extraction, processing or running the samples on different lanes.

All of these sources of variation are confounded, and there are no particularly good ways of separating them out.

But there are lots of sources of variation, so how do we account for these?

Replication

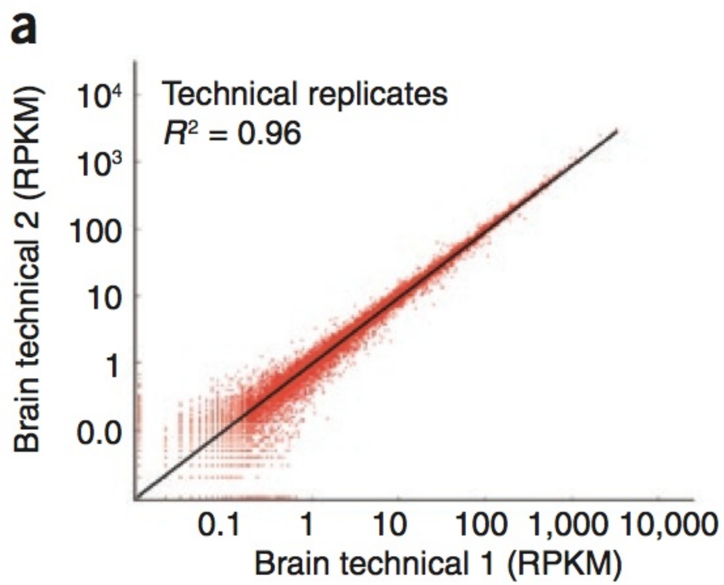
To date, several studies have suggested that “technical” replicates for RNA-seq show very little variation/ high c

Sampling

Replication

Blocking

Randomization



Mortazavi et al. 2008

How might such a statement be misleading about variation?

Replication

Sampling

Replication

Blocking

Randomization

This study looked at a single source of technical variation.

Running exactly the same sample on two different lanes on a flow cell.

This completely ignores other sources of "technical variation"
variation due to RNA purification
variation due to fragmentation, labeling, etc..
lane to lane variation
flow cell to flow cell variation

All of these may be important (although unlikely interesting) sources of variation…

However…..

Replication

Sampling

Replication

Blocking

Randomization

Many studies have ignored the BIOLOGICAL SOURCES of VARIATION between replicates. In most cases biological variation between samples (from the same treatment) are generally far more variable than technical sources of variation.

While it would be nice to be able to partition various sources of technical variation (such as labeling, RNA extraction), it often too expensive to perform such a design (see white board).

IF you have limited resources, it is generally far better to have biological replication (independent biological samples for a given treatment) than technical replication.

Does these lead to confounded sources of variation?

Blocking

Sampling

Replication

Blocking

Randomization

Blocks in experimental design represent some factor (usually something not of major interest) that can strongly influence your outcomes. More importantly it is a factor which you can use to group other factors that you are interested in.

For instance in agriculture there is often plot to plot variation. You may not be interested in the plot themselves but in the variety of crops you are growing.

But what would happen if you grew all of strain 1 on plot 1 and all of strain 2 on plot 2?

Whiteboard.

These plots would represent blocking levels

Blocking

Sampling

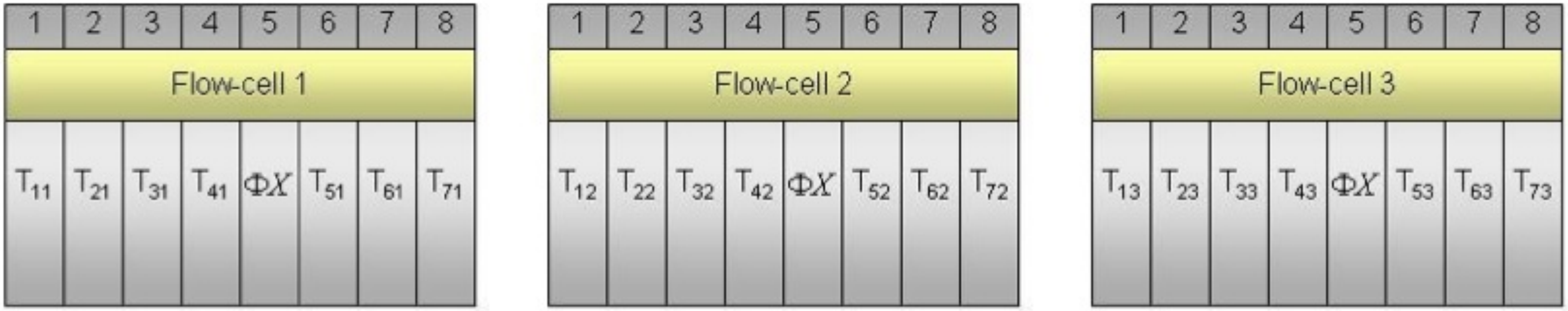
Replication

Blocking

Randomization

In genomic studies the major blocking levels are often the slide/chip for microarrays (i.e. two samples /slide for 2 color arrays, 16 arrays/slide for Illumina arrays).

For GAII/HiSeq RNA-seq data the major blocking effect is the flow cell itself, or lanes within the flow cell.



1	2	3	4	5	6	7	8
Flow-cell 1							
T_{11}	T_{21}	T_{31}	T_{41}	ΦX	T_{51}	T_{61}	T_{71}

1	2	3	4	5	6	7	8
Flow-cell 2							
T_{12}	T_{22}	T_{32}	T_{42}	ΦX	T_{52}	T_{62}	T_{72}

1	2	3	4	5	6	7	8
Flow-cell 3							
T_{13}	T_{23}	T_{33}	T_{43}	ΦX	T_{53}	T_{63}	T_{73}

Auer and Doerge 2010

Blocking

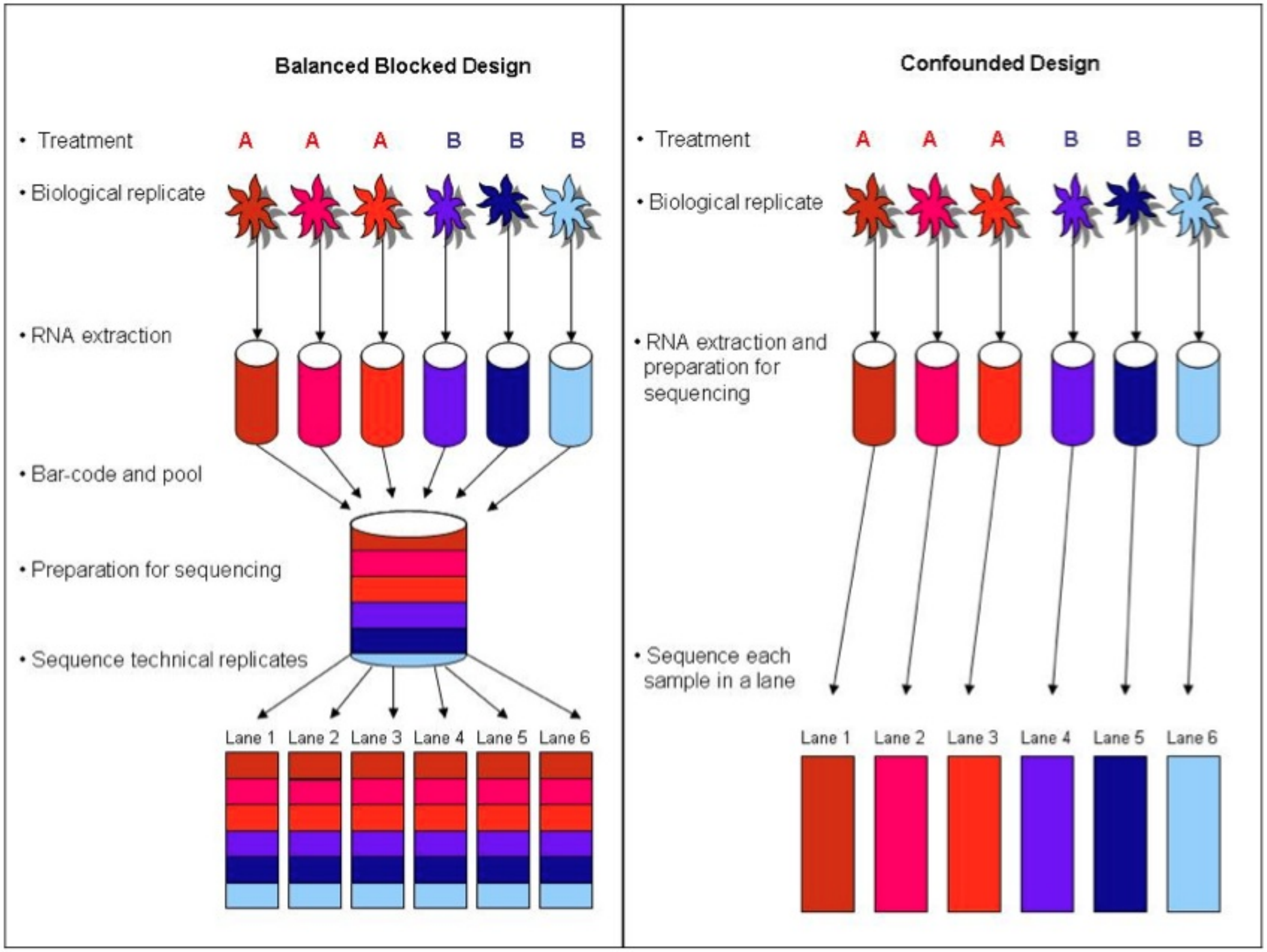
Incorporating lanes as a blocking effect

Sampling

Replication

Blocking

Randomization



Auer and Doerge 2010

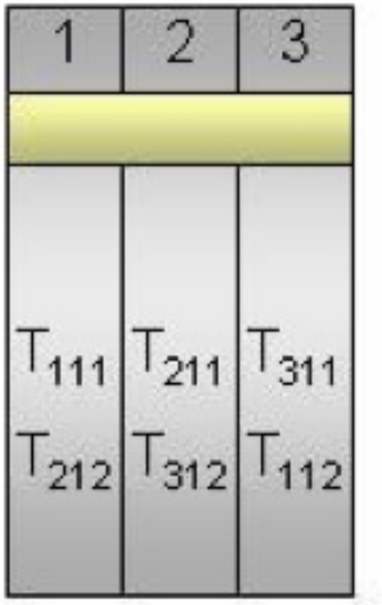
Blocking designs

Sampling

Replication

Blocking

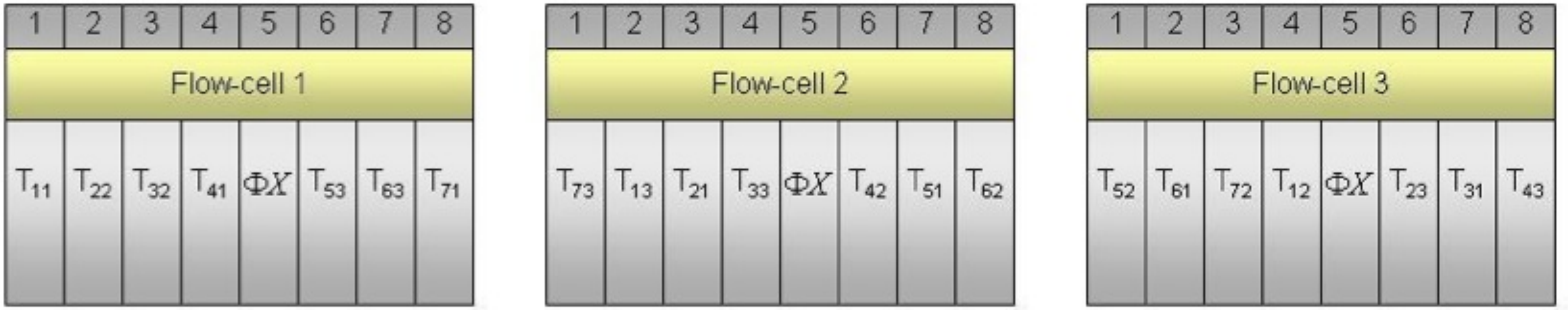
Randomization



1	2	3
T_{111}	T_{211}	T_{311}
T_{212}	T_{312}	T_{112}

Balanced **I**ncomplete **B**locking **D**esign (BIBD)

Let’s dissect these subscripts.



Flow-cell 1

1	2	3	4	5	6	7	8
T_{11}	T_{22}	T_{32}	T_{41}	ΦX	T_{53}	T_{63}	T_{71}

Flow-cell 2

1	2	3	4	5	6	7	8
T_{73}	T_{13}	T_{21}	T_{33}	ΦX	T_{42}	T_{51}	T_{62}

Flow-cell 3

1	2	3	4	5	6	7	8
T_{52}	T_{61}	T_{72}	T_{12}	ΦX	T_{23}	T_{31}	T_{43}

Balanced for treatments across flow cells.. Randomized for location

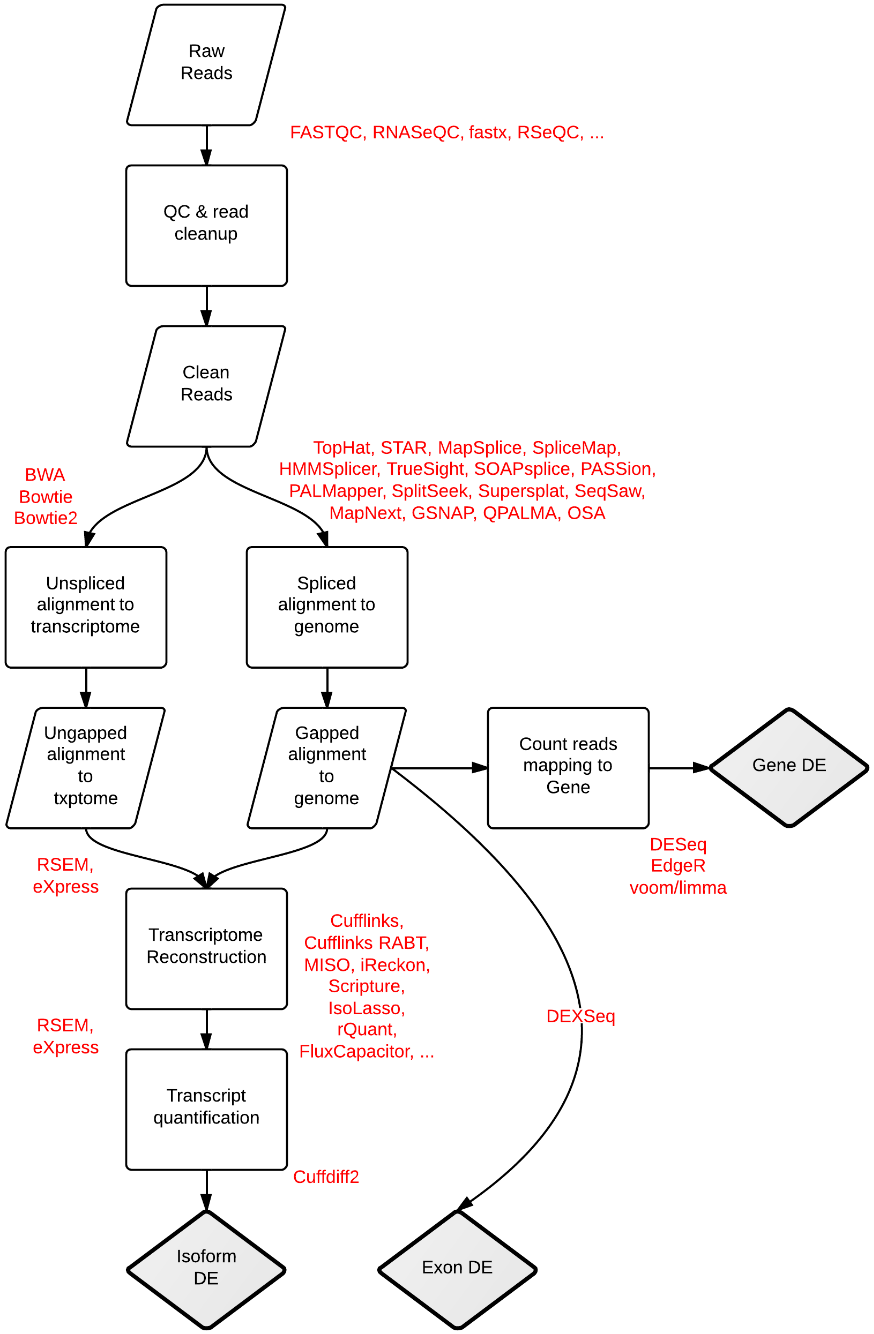
Auer and Doerge 2010

You have designed and run the experiment… now what?

First a couple of quotes from a great statistician:

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.
John Tukey

Numerical quantities focus on expected values, graphical summaries on unexpected values.
John Tukey



Raw Reads

FASTQC, RNASeQC, fastx, RSeQC, ...

QC & read cleanup

Clean Reads

BWA
Bowtie
Bowtie2

TopHat, STAR, MapSplice, SpliceMap,
HMMSplicer, TrueSight, SOAPsplice, PASSion,
PALMapper, SplitSeek, Supersplat, SeqSaw,
MapNext, GSNAP, QPALMA, OSA

Unspliced alignment to transcriptome

Spliced alignment to genome

Ungapped alignment to txptome

Gapped alignment to genome

Count reads mapping to Gene

Gene DE

DESeq
EdgeR
voom/limma

RSEM,
eXpress

Transcriptome Reconstruction

Cufflinks,
Cufflinks RABT,
MISO, iReckon,
Scripture,
IsoLasso,
rQuant,
FluxCapacitor, ...

DEXSeq

RSEM,
eXpress

Transcript quantification

Cuffdiff2

Isoform DE

Exon DE

RNA-seq Workflows and Tools. Stephen Turner. Figshare. http://dx.doi.org/10.6084/m9.figshare.662782

Visually examine your data at every step of the analysis!!!!!

By far the single most important thing you should be thinking about with your data, at every stage of the analysis (raw, filtered, aligned, normalized, modeled) is how to present the data graphically.

Plots are a very good way to pick out if something wacky is going on with your data.

Even if your data is of the highest quality, different software can produce very different results (in unattended ways). Plenty of Bugs and features.

R

Exploratory Data Analysis

- What is EDA?
- An approach to data analysis, largely using graphical techniques to help refine hypotheses and aid in the model building process.
- Some advocates suggest it can be used for hypothesis generation (we will discuss when and where this might be sensible).

Objectives of EDA

- Propose/refine models to explain the observed patterns of the data.
- ***Assess assumptions on which statistical modeling is based***
- ***Provide a context for further data collection.***
- ***Help in determining the appropriate form of statistical modeling (LS, MLE< Bayesian, resampling…)***.

Useful online tutorial on EDA

http://www.itl.nist.gov/div898/handbook/eda/eda.htm

EDA

- EDA emphasizes “robust” and non-parametric approaches to examining the data.

Critiques of EDA

- “Data-Dredging”. Using it as “free lunch” with respect to *a posteriori* hypothesis testing.

- Observing patterns that are not real.

“ Under torture, the data readily yields false confessions”

(MainDonald & Braun 2003)

suggestions

Bolker (2007) suggests an honest approach: prior to examining your data you write down a list of the patterns you are looking for so that you can distinguish between:

1) Patterns you were initially looking for
2) Unanticipated patterns that answer the same question in different ways.
3) Novel (but likely spurious) patterns.

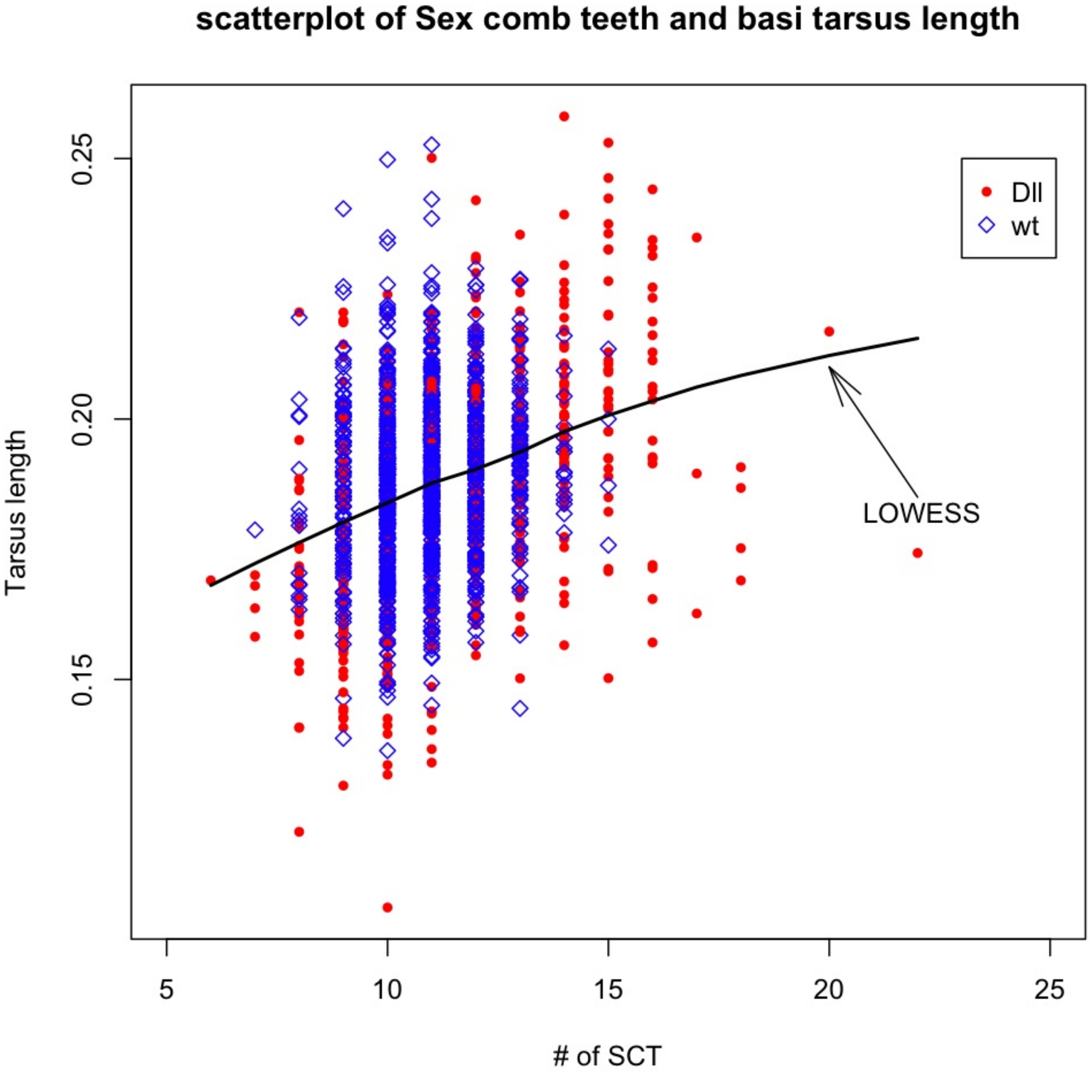
Cross-validation

Subset your data (cross-validation):

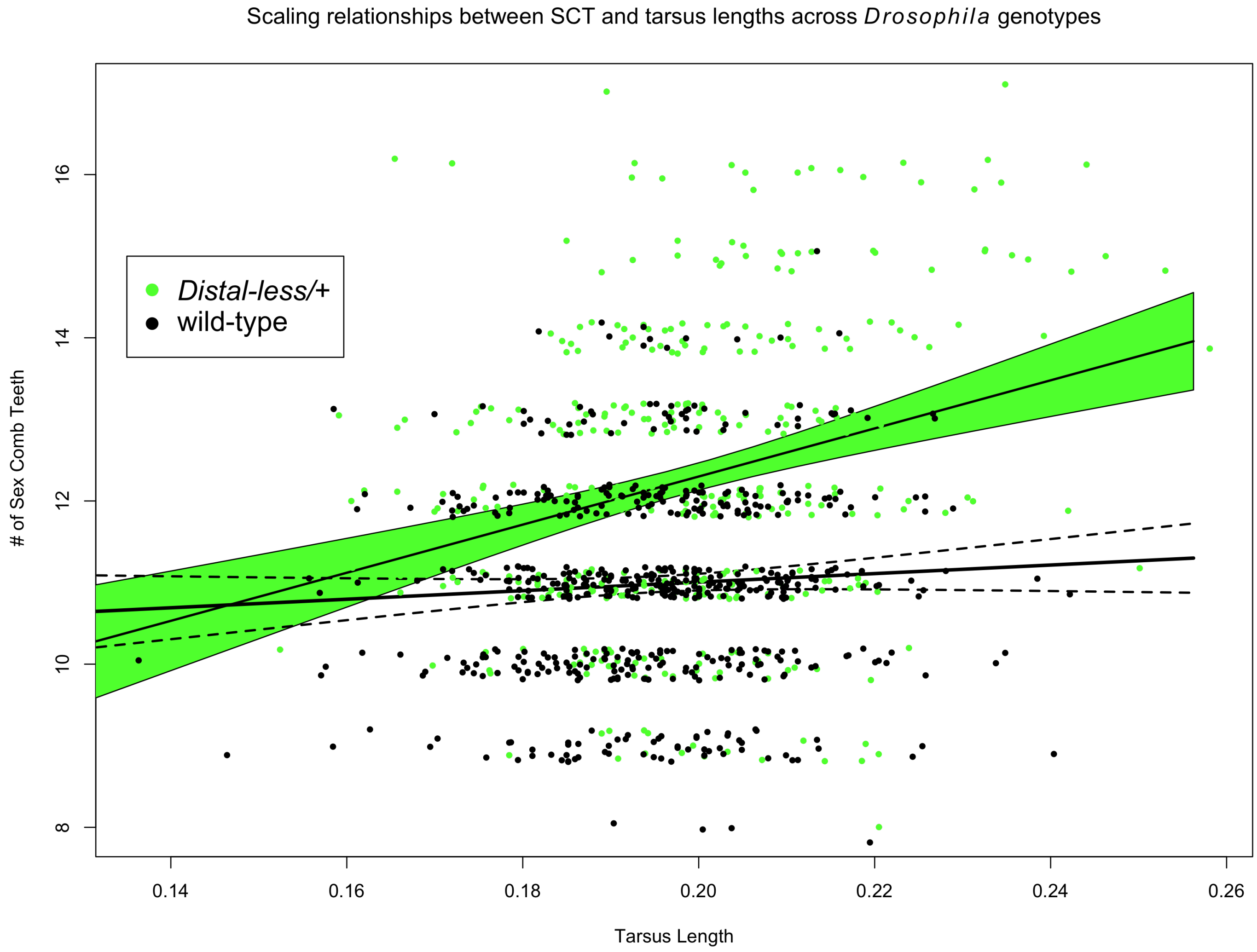
The other useful approach is to use a random subset of all of your data (No more than 40% of it) for data exploration, and then you can perform the model fitting based on the entire data set (or better yet the remaining 60%).

Of course, this only works if you have enough data to do so!

Bi-variate scatterplot with loess. Is this data dredging?

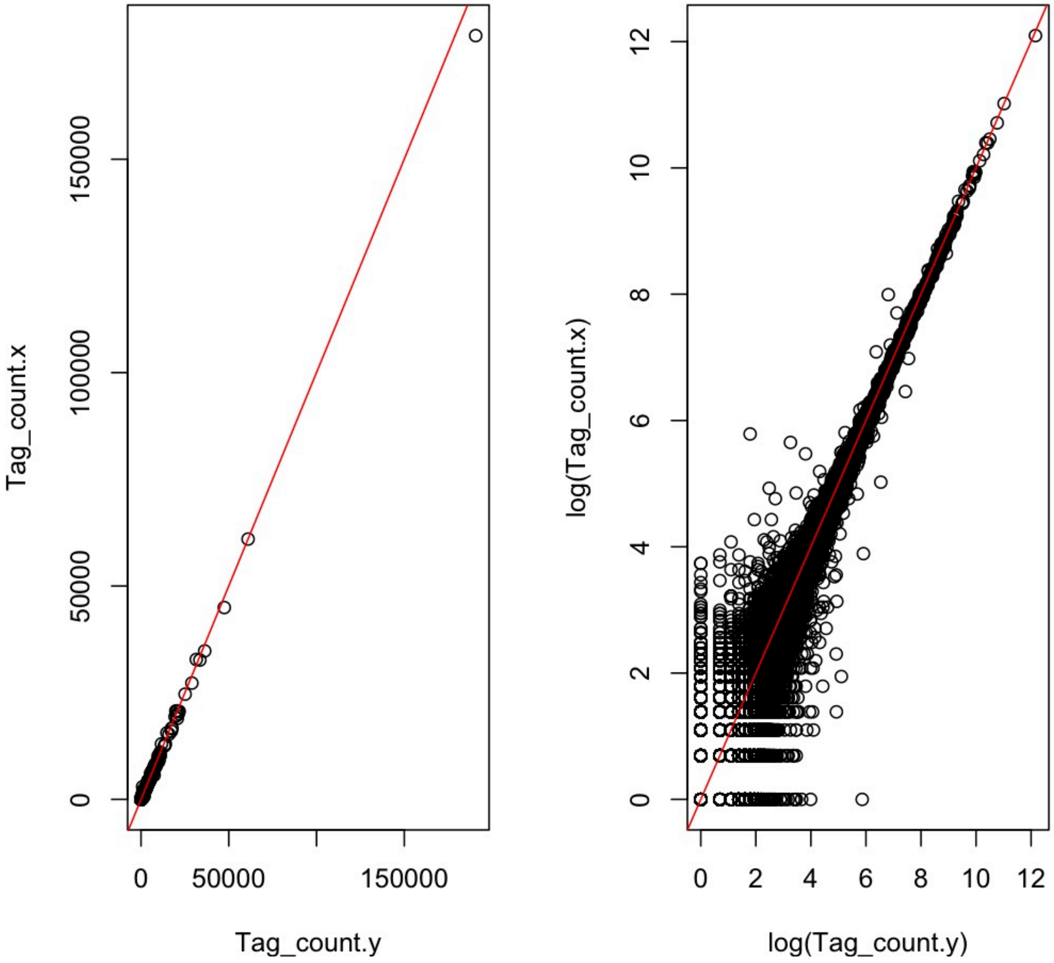


This would not be considered EDA.. Why not?



Some examples: Digital Gene Expression (Sequence tags) for RNA quantification

A comparison of two lanes of DGE sequence tags.

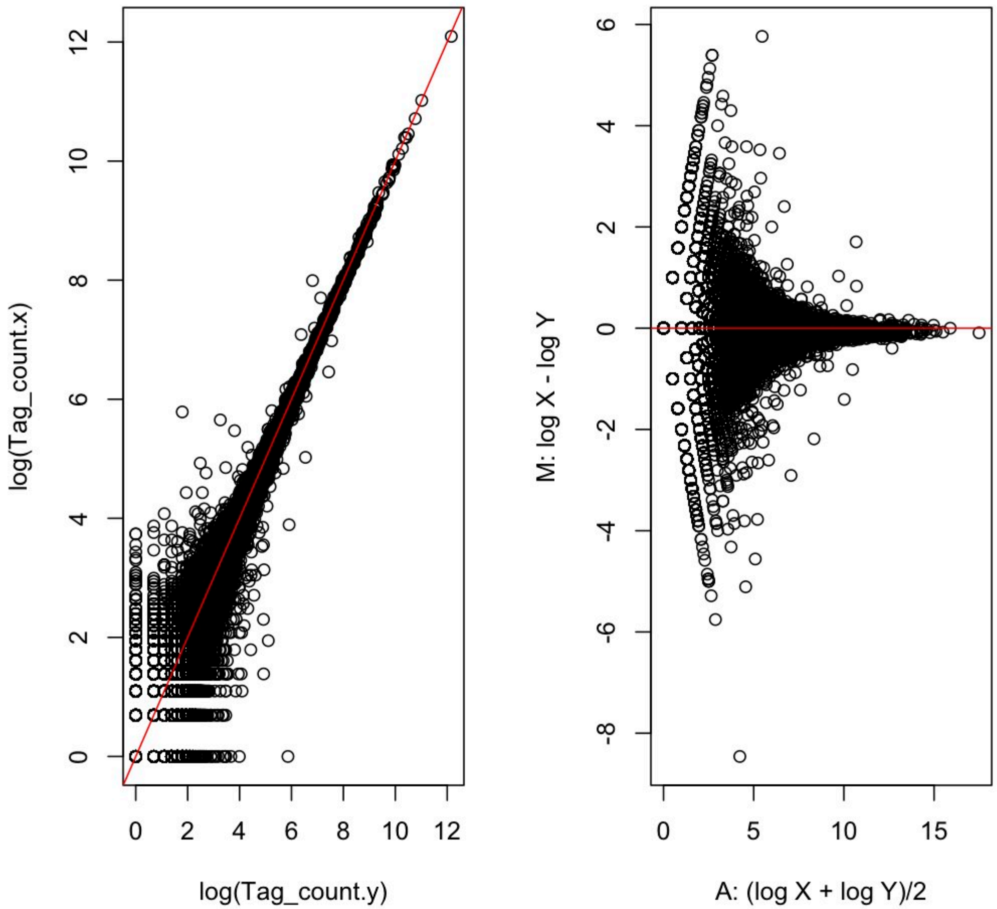


What's the difference between these plots?

Some examples: Digital Gene Expression (Sequence tags) for RNA quantification

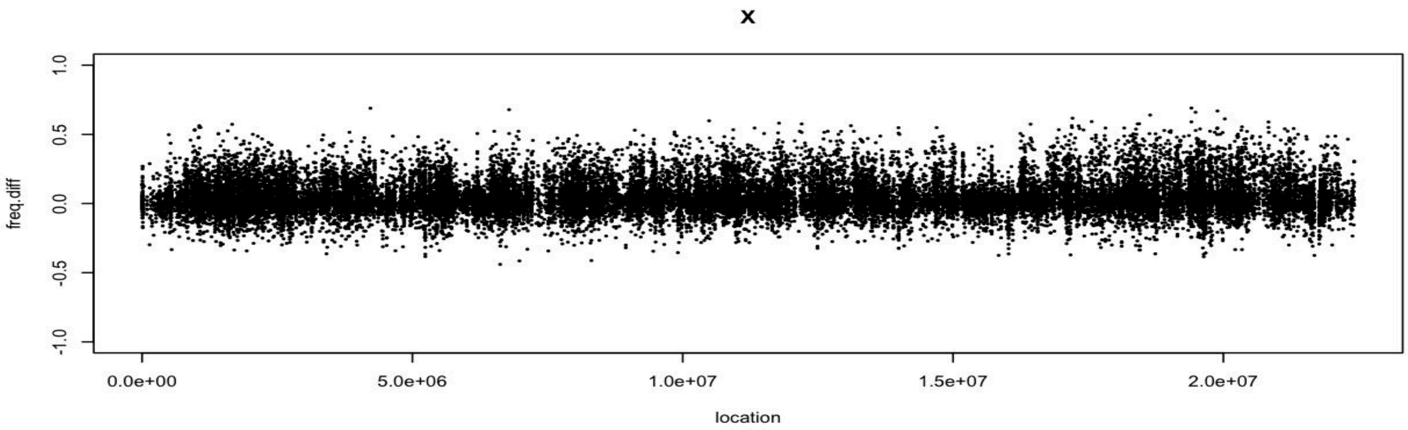
A comparison of two lanes of DGE sequence tags.

What’s the difference between these plots?

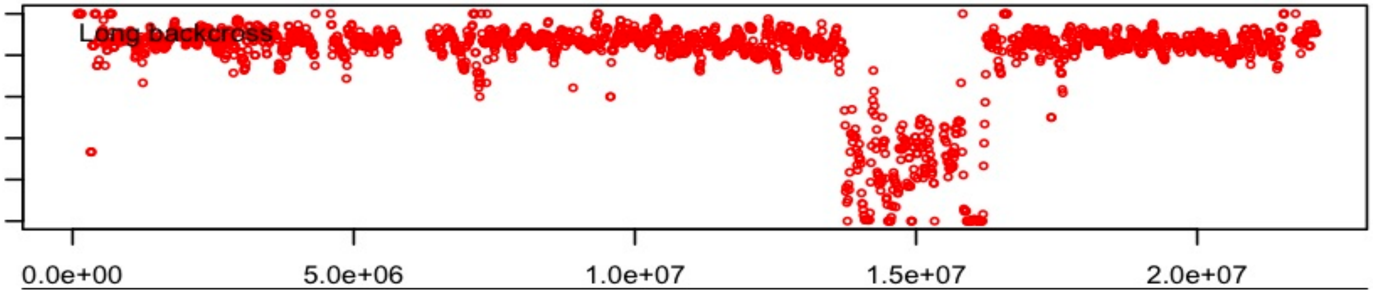


MA plot

Whole genome analysis can be messy, how do we deal with this.



Every SNP



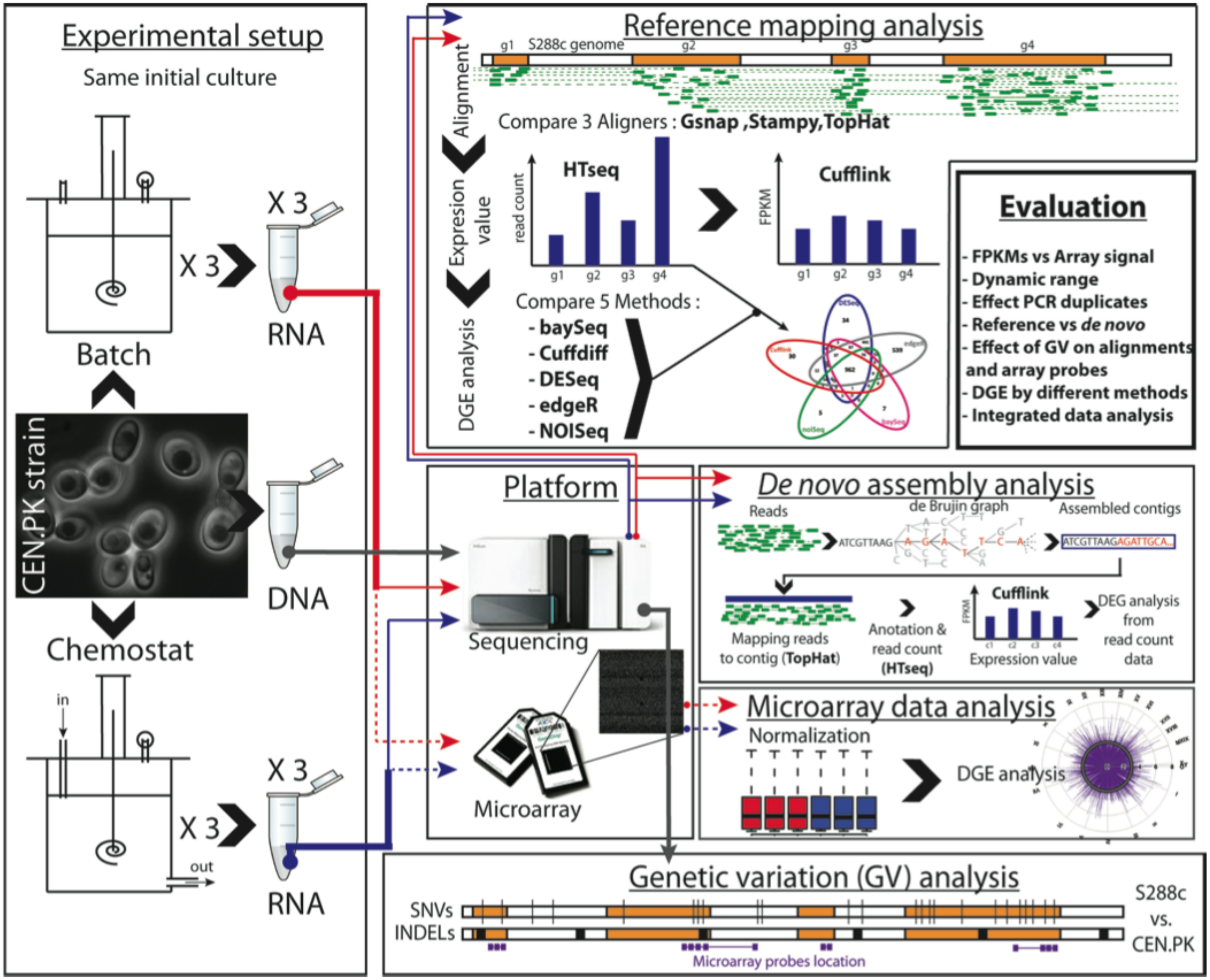
Sliding window& binning

Genome level Correlations are not a good measure of repeatability

- One BIG mistake people make with BIG genomics data sets is treat each gene (or genomic interval) as an independent data point. In particular for correlation analysis.

- This is almost never the case…

chalkboard.

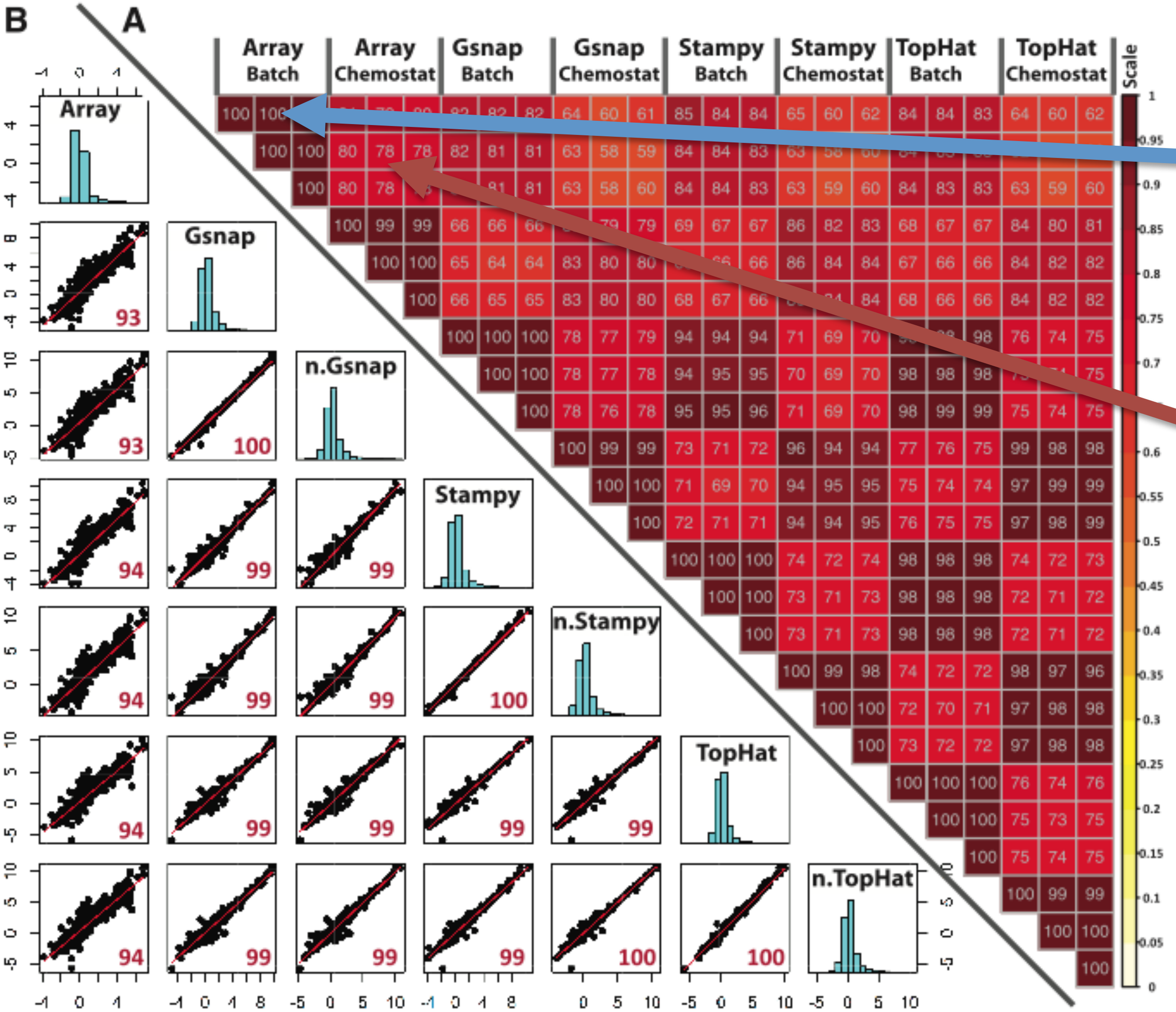


Evaluation

- FPKMs vs Array signal
- Dynamic range
- Effect PCR duplicates
- Reference vs *de novo*
- Effect of GV on alignments and array probes
- DGE by different methods
- Integrated data analysis

Nookaew et al 2102 NAR

What does this tell us?



Array Batch			Array Chemostat			Gsnap Batch			Gsnap Chemostat			Stampy Batch			Stampy Chemostat			TopHat Batch			TopHat Chemostat		
100	100	[illegible]	[illegible]	[illegible]	[illegible]	82	82	82	64	60	61	85	84	84	65	60	62	84	84	83	64	60	62
	100	100	80	78	78	82	81	81	63	58	59	84	84	83	63	58	60	84	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		100	80	78	[illegible]	[illegible]	81	81	63	58	60	84	84	83	63	59	60	84	83	83	63	59	60
			100	99	99	66	66	66	[illegible]	79	79	69	67	67	86	82	83	68	67	67	84	80	81
				100	100	65	64	64	83	80	80	[illegible]	66	66	86	84	84	67	66	66	84	82	82
					100	66	65	65	83	80	80	68	67	66	[illegible]	[illegible]	84	68	66	66	84	82	82
						100	100	100	78	77	79	94	94	94	71	69	70	[illegible]	[illegible]	98	76	74	75
							100	100	78	77	78	94	95	95	70	69	70	98	98	98	[illegible]	[illegible]	75
								100	78	76	78	95	95	96	71	69	70	98	99	99	75	74	75
									100	99	99	73	71	72	96	94	94	77	76	75	99	98	98
										100	100	71	69	70	94	95	95	75	74	74	97	99	99
											100	72	71	71	94	94	95	76	75	75	97	98	99
												100	100	100	74	72	74	98	98	98	74	72	73
													100	100	73	71	73	98	98	98	72	71	72
														100	73	71	73	98	98	98	72	71	72
															100	99	98	74	72	72	98	97	96
																100	100	72	70	71	97	98	98
																	100	73	72	72	97	98	98
																		100	100	100	76	74	76
																			100	100	75	73	75
																				100	75	74	75
																					100	99	99
																						100	100
																							100

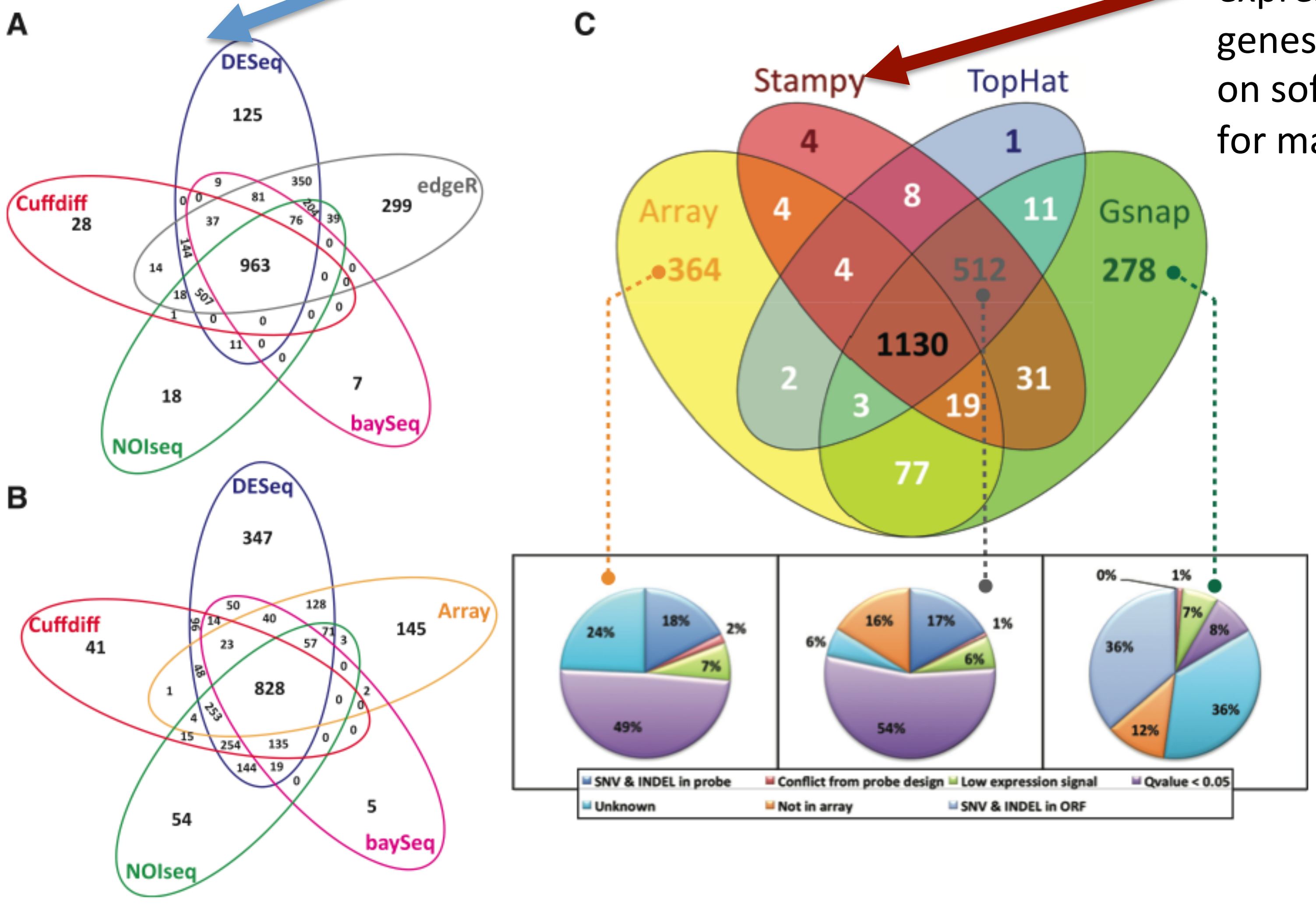
These cells are biological replicates (diagonals)

These cells are for different software

Nookaew et al 2102 NAR

Differentially expressed genes based on software for quantification

Differentially expressed genes based on software for mapping



Nookaew et al 2102 NAR

One right way?

- At this point it is safest to assume that there is no one single “right way” to analyze your NGS data (for RNAseq or anything else).
- While a number of studies have demonstrated that several pipelines give similar results, it is best to try several approaches.
- Even fitting what seems like the “same” model can give different results from different software.
- Let’s discuss why.

Is performing your analysis multiple ways enough?

- Just because you create multiple forks in your analysis (different read mappers, different tools for quantification), does not mean you are out of the woods.

- Always generate lots of plots to help you visualize your data in as many ways as possible.

Simulating your analysis

- One other important tool (that is very straightforward) is to use simple simulation or resampling approaches to “rig” the analysis.

- White board.

Readings for Monday and Tuesday

Vijay, N., Poelstra, J. W., Künstner, A., & Wolf, J. B. W. (2012). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular Ecology. doi:10.1111/mec.12014

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods, 8(6), 469–477. doi:10.1038/nmeth.1613

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Research, 40(20), 10084–10097. doi:10.1093/nar/gks804