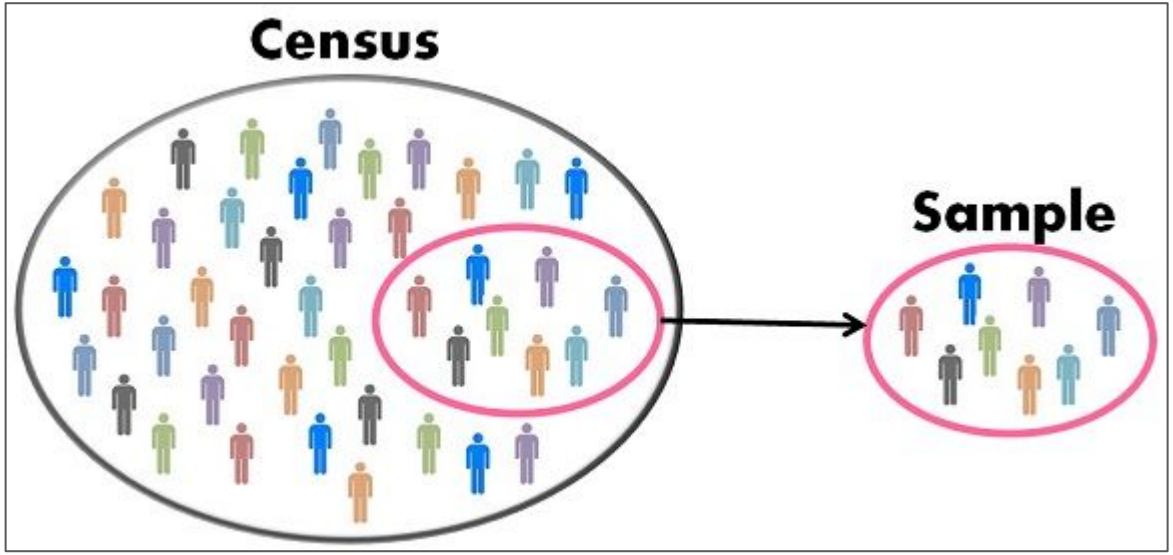
Experimental Design for Phylotranscriptomics

Goals

1. Clear some misconceptions
2. Explain what can be done
3. Talk about basics of how to do it

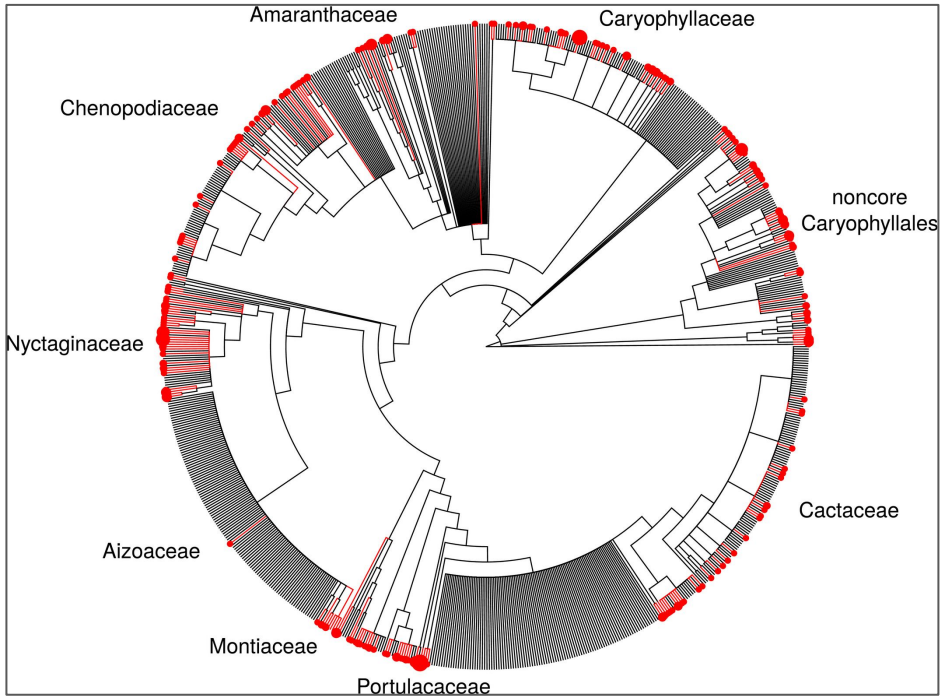
Outline

1. **Sampling scheme**
2. Tissue choice and collection
3. Sequencing



Choosing which lineages to sample

1. Species relationships

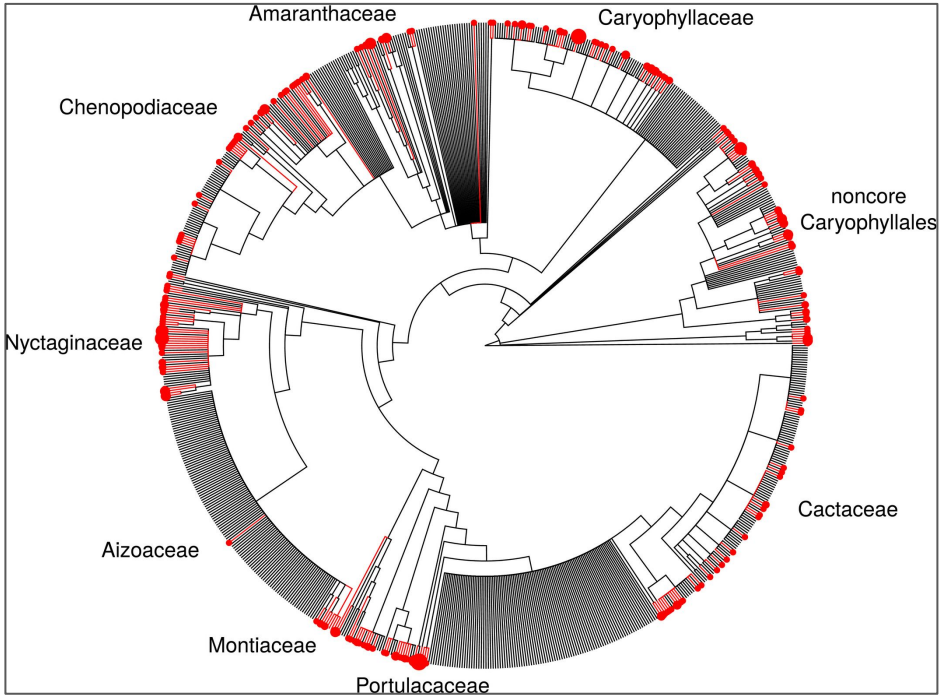


Walker et al. 2018

Choosing which lineages to sample

1. Species relationships

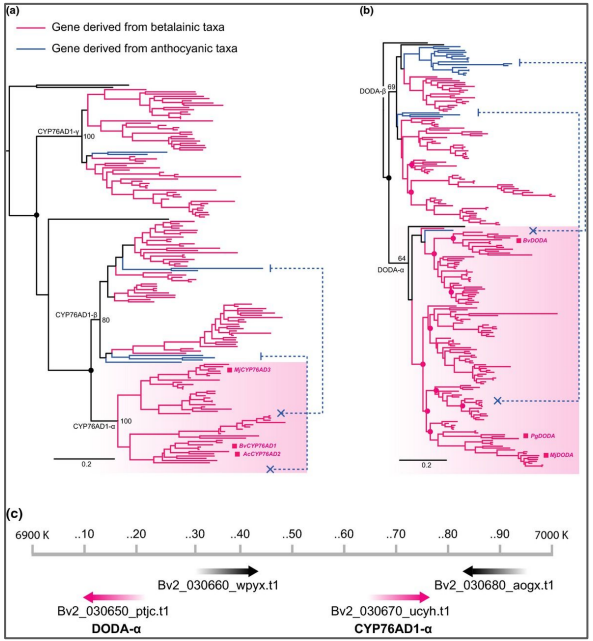
   (Don’t forget outgroups!)



Walker et al. 2018

Choosing which lineages to sample

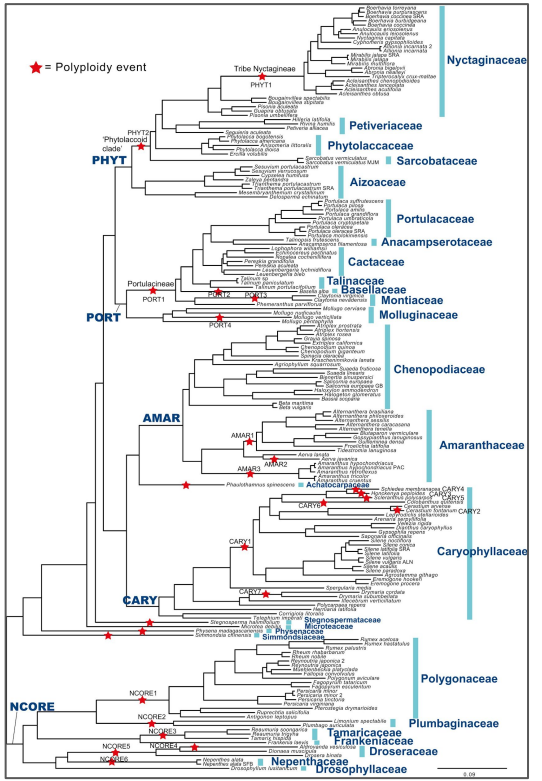
1. Species relationships
2. Gene duplication



Brockington et al. 2015

Choosing which lineages to sample

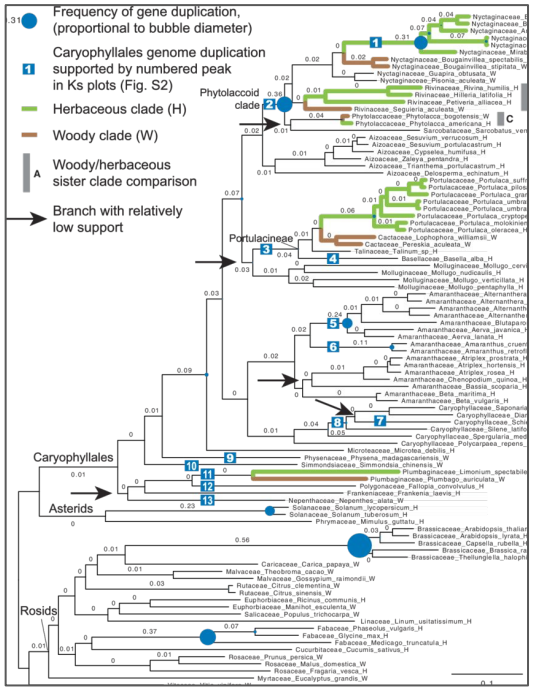
1. Species relationships
2. Gene duplication
3. Genome duplication events



Yang et al. 2018

Choosing which lineages to sample

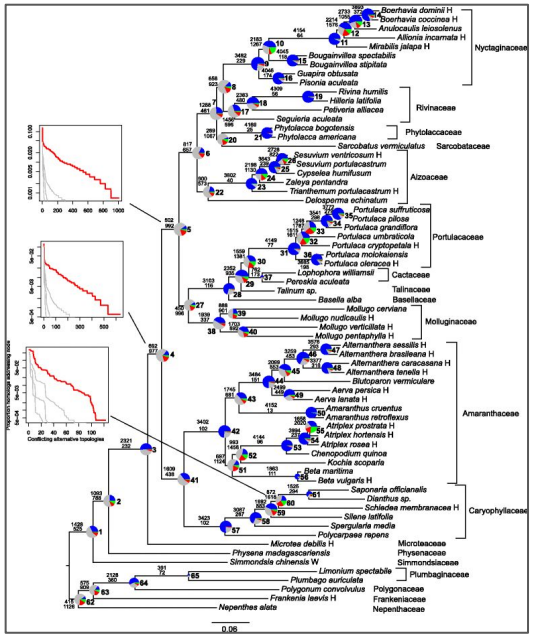
1. Species relationships
2. Gene duplication
3. Genome duplication events
4. Transcriptome wide molecular patterns



Yang et al. 2015

Choosing which lineages to sample

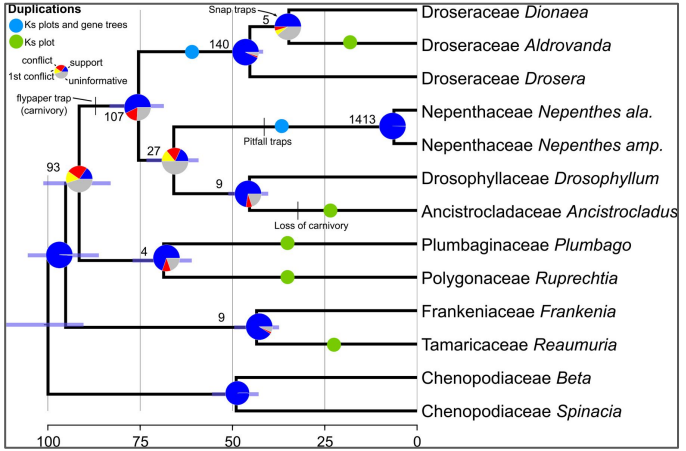
1. Species relationships
2. Gene duplication
3. Genome duplication events
4. Transcriptome wide molecular patterns
5. Gene tree conflict



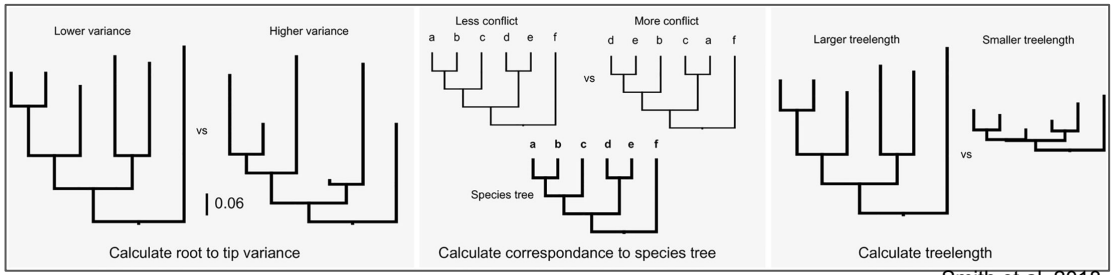
Smith et al. 2015

Choosing which lineages to sample

1. Species relationships
2. Gene duplication
3. Genome duplication events
4. Transcriptome wide molecular patterns
5. Gene tree conflict
6. Molecular dating



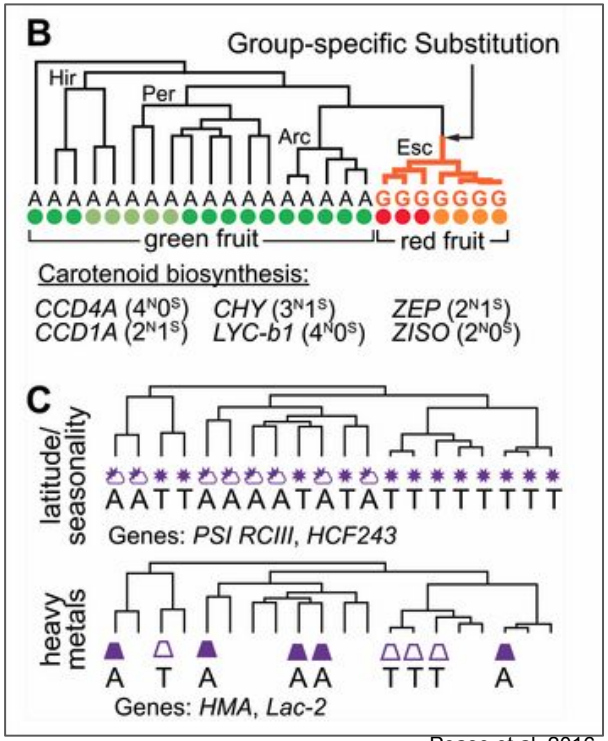
Walker et al. 2017



Smith et al. 2018

Choosing which lineages to sample

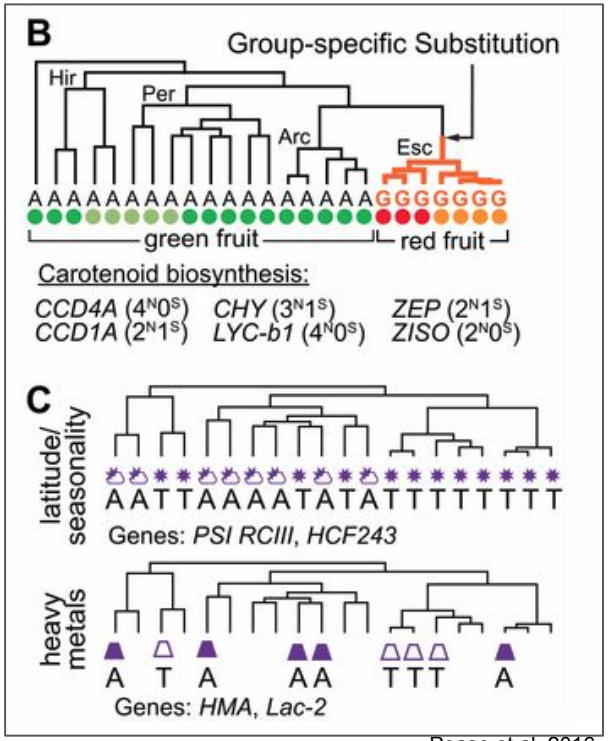
1. Species relationships
2. Gene duplication
3. Genome duplication events
4. Transcriptome wide molecular patterns
5. Gene tree conflict
6. Molecular dating
7. PhyloGWAS



Pease et al. 2016

Choosing which lineages to sample

1. Species relationships
2. Gene duplication
3. Genome duplication events
4. Transcriptome wide molecular patterns
5. Gene tree conflict
6. Molecular dating
7. PhyloGWAS
8. And much more!



Pease et al. 2016

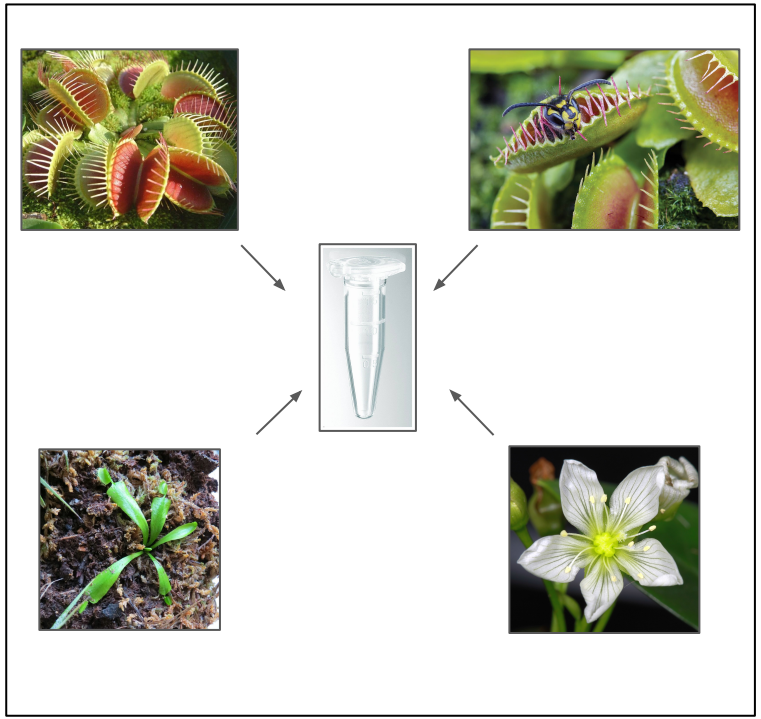
Outline

1. Choosing sampling method
2. **Tissue choice and collection**
3. Sequencing

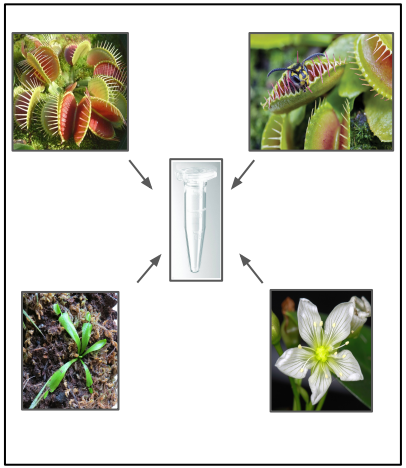
Traditional differential gene expression

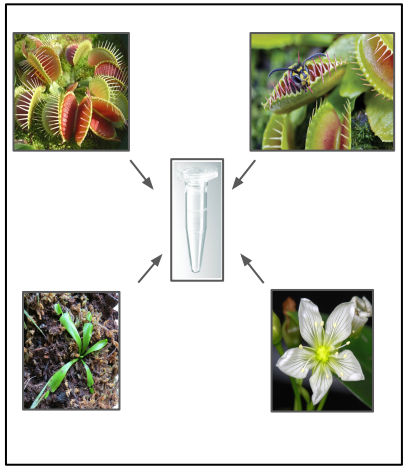
Traditional differential gene expression

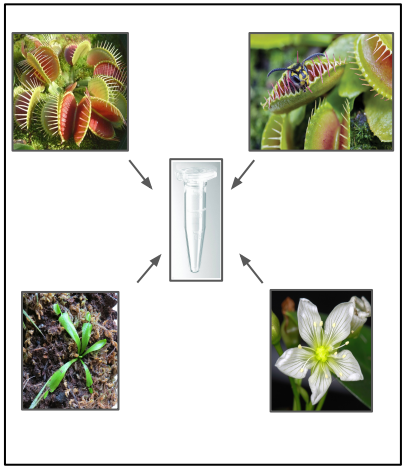
Phylotranscriptomic: Max gene coverage



Replication: ideal, expensive and question dependent







Tissue collections

Find a great systematist or botanic garden manager to go collecting with

Collect tissue for extraction and collect something for a herbarium voucher

Ideally collect multiple samples to prepare for the worse

After collecting place in liquid nitrogen or RNA later ASAP

Store in a -80

An efficient field and laboratory workflow for plant phylotranscriptomic projects*

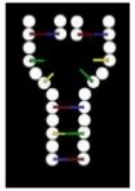
Outline

1. Choosing sampling method
2. Tissue choice and collection
3. **Sequencing**

SANGERS Vs. NGS

Features	Sanger	NGS
Sequencing Samples	Clones, PCR	DNA Libraries
Preparation Steps	Few, Sequencing reactions clean up	Many, Complex procedures
Data Collection	Samples in plates : 96, 384	Samples on slides 1-16+
Data	1 Read/ Sample	Thousands & Millions of Reads/ Samples.

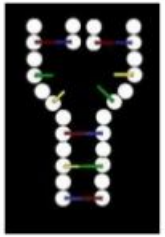




Next generation sequencing is necessary

SANGERS Vs. NGS

Features	Sanger	NGS
Sequencing Samples	Clones, PCR	DNA Libraries
Preparation Steps	Few, Sequencing reactions clean up	Many, Complex procedures
Data Collection	Samples in plates : 96, 384	Samples on slides 1-16+
Data	1 Read/ Sample	Thousands & Millions of Reads/ Samples.



Short reads are usually fine

Platform	Instrument	Unit	Reads / Unit (single reads)	Ave. or Max Read Length (bp)	Read Type	Error Type	Highlight	Reference
Illumina	HiSeq X Ten	Lane	375,000,000	300	PE	substitution	Greatest throughput and number of reads compared to all other instruments. Currently only available for whole human genome sequencing.	1
Illumina	NextSeq 500 High-Output	Lane	400,000,000	300	SR & PE	substitution	Highest output of any desktop sequencer, this instrument is ideally suited for exome, transcriptomics, whole genome and targeted resequencing.	2
Illumina	NextSeq 500 Mid-Output	Lane	130,000,000	300	PE	substitution		2
Illumina	HiSeq High-Output v4	Lane	250,000,000	250	SR & PE	substitution	Well suited for de novo and resequencing of small and large genomes, this Illumina instrument is today's sequencing workhorse. Latest v4 chemistry now allows more reads per lane and a slightly longer read length, 2x125 bp	3

We've had good luck with ~25,000,000 reads

Platform	Instrument	Unit	Reads / Unit (single reads)	Ave. or Max Read Length (bp)	Read Type	Error Type	Highlight	Reference
Illumina	HiSeq X Ten	Lane	375,000,000	300	PE	substitution	Greatest throughput and number of reads compared to all other instruments. Currently only available for whole human genome sequencing.	1
Illumina	NextSeq 500 High-Output	Lane	400,000,000	300	SR & PE	substitution	Highest output of any desktop sequencer, this instrument is ideally suited for exome, transcriptomics, whole genome and targeted resequencing.	2
Illumina	NextSeq 500 Mid-Output	Lane	130,000,000	300	PE	substitution		2
Illumina	HiSeq High-Output v4	Lane	250,000,000	250	SR & PE	substitution	Well suited for de novo and resequencing of small and large genomes, this Illumina instrument is today's sequencing workhorse. Latest v4 chemistry now allows more reads per lane and a slightly longer read length, 2x125 bp	3

Avoiding contamination is key!

Final thoughts

This is a very rough intro into doing this but we’re happy to talk about our experience when you get closer to designing your experiment

Contamination is a costly and easy mistake to make

Sometimes your data won’t be able to address the question you want but lots of other cool things can be pursued

Thanks for your attention!