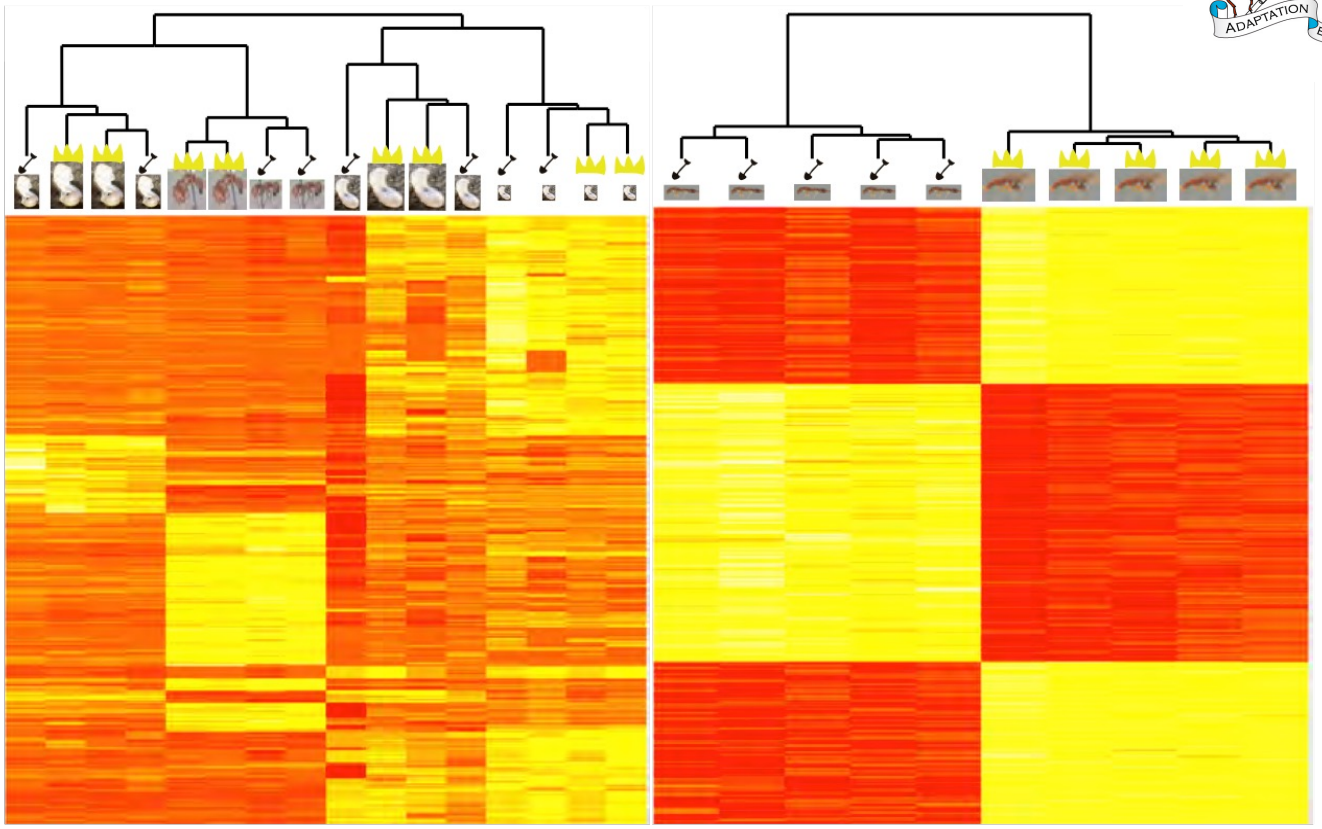
Transcriptomics

Prof. Dr. Jürgen Gadau



I have…

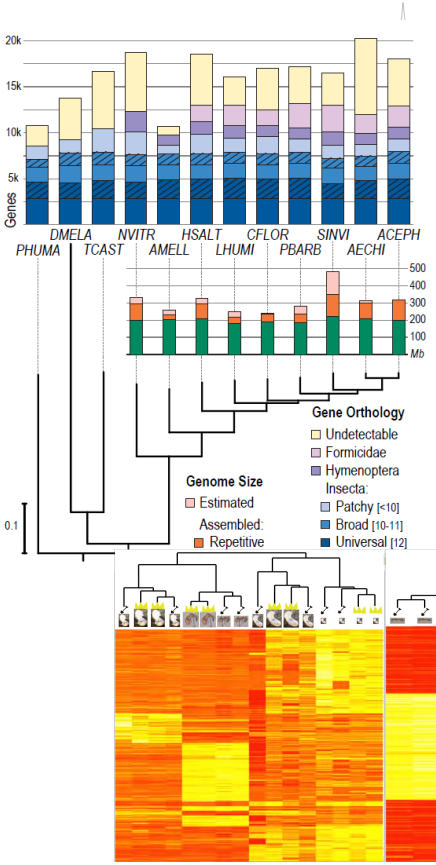
I have …

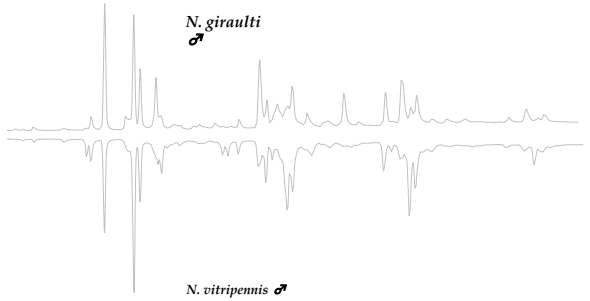
- heard about transcriptome but can`t explain it.
- heard about transcriptomics and can explain what it is.
- read an article where transcriptomics was used.
- I already worked with a transcriptomic technique.

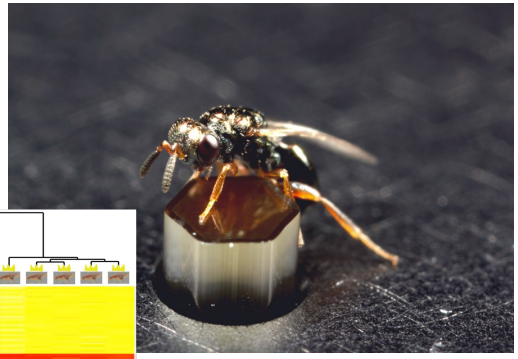
What is a transcriptome?

- The sum of all transcribed components (RNA) of a genome (DNA) is referred to as its transcriptom.

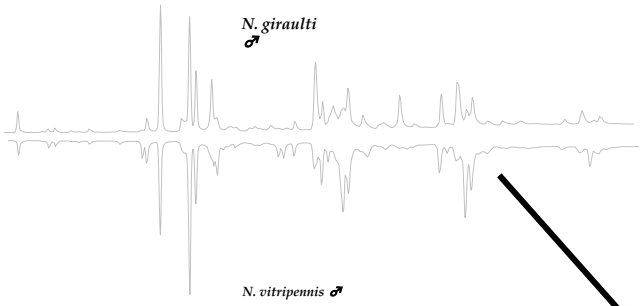
What does my laboratory do?





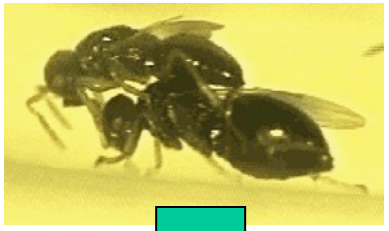


Functional and Comparative Genomics of speciation and species differences

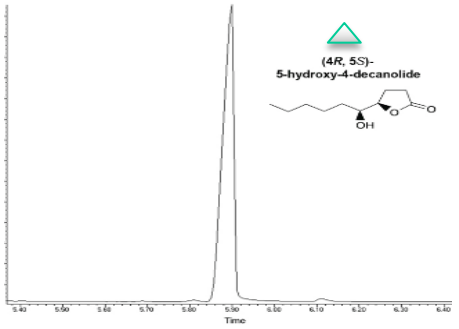


Cuticular hydrocarbon pattern – species and sex recognition

Male Courtship behavior



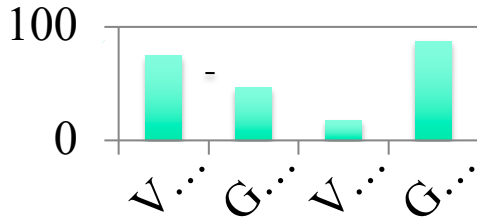
Male sex pheromone evolution



Prezygotic

Postzygotic

F2 Hybrid breakdown

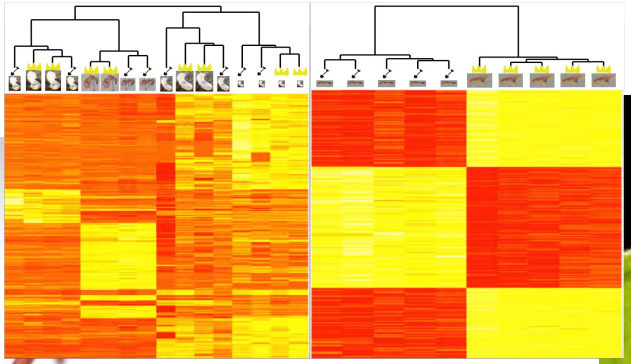


Evolution of Gene Regulation – Phenotypic Plasticity

Aggression

COOPERATION AND CONFLICT

Reproductive division of labor



non-reproductive division of lab

Worker Subcastes

What is Functional Genomics?

…and how does transcriptomics factor into it?

Some Definitions…

- **Generally…Functional genomics** is a field of molecular biology that attempts to make use of the vast wealth of data produced by **genomic** and transcriptomic projects (such as **genome** sequencing projects and RNA-seq- RNA sequencing) to describe gene (and protein) functions and interactions.

- **More mechanistically….functional genomics** uses genomic data to study gene and protein expression and function on a global scale (genome-wide or system-wide), focusing on gene transcription, translation and protein-protein interactions, and often involving high-throughput methods.

- **RNAseq…** (RNA sequencing) is a sequencing technique which uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment,

Why is it not enough to have the genome sequence of an organism/individual , if one wants to understand the funktion of a gene?

Genotype to phenotype

- Classic Mendelian genetics and how we teach Genetics/Evolution is usually looking at a single locus/gene that we can map and has a definitive phenotype.

- **One gene (*white*) → 1 character**

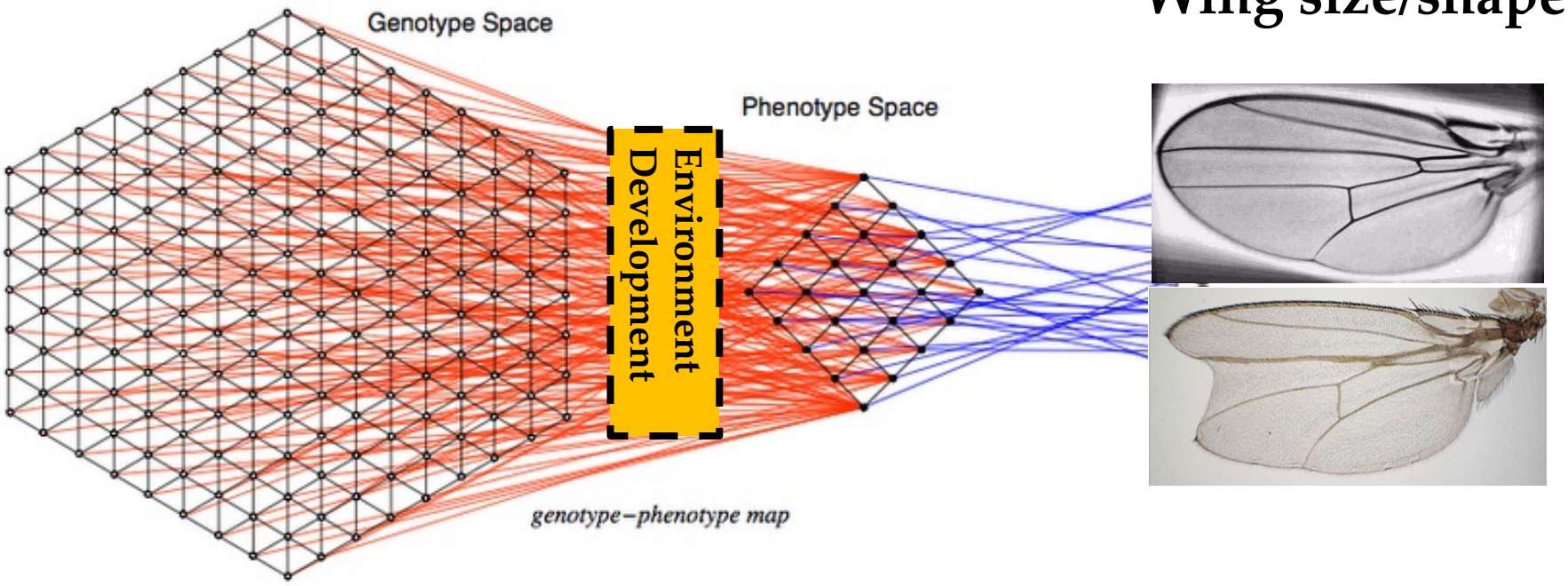
White-eyed mutant*

- However, organisms or the genetic architecture of complex/simple traits are much more complex.

***In January 1910, more than a century ago, Thomas Hunt Morgan *discovered* his *first Drosophila mutant*, a white-eyed male (Morgan 1910).**

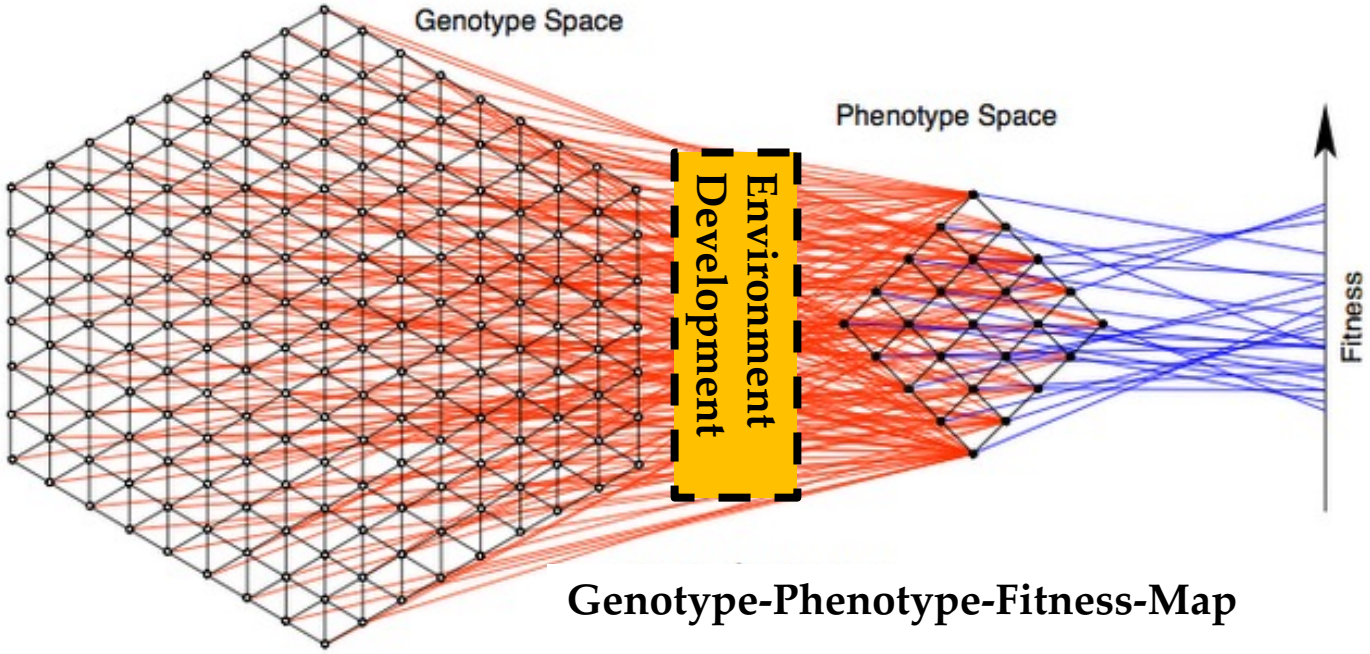
From Genotype to Phenotype…a map!

Wing size/shape



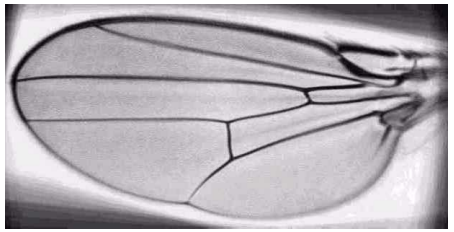
Genome - Transcriptome - Phenotype 2
(Phenotype 1)

From Genotype to Phenotype…fitness effect (Evolution)



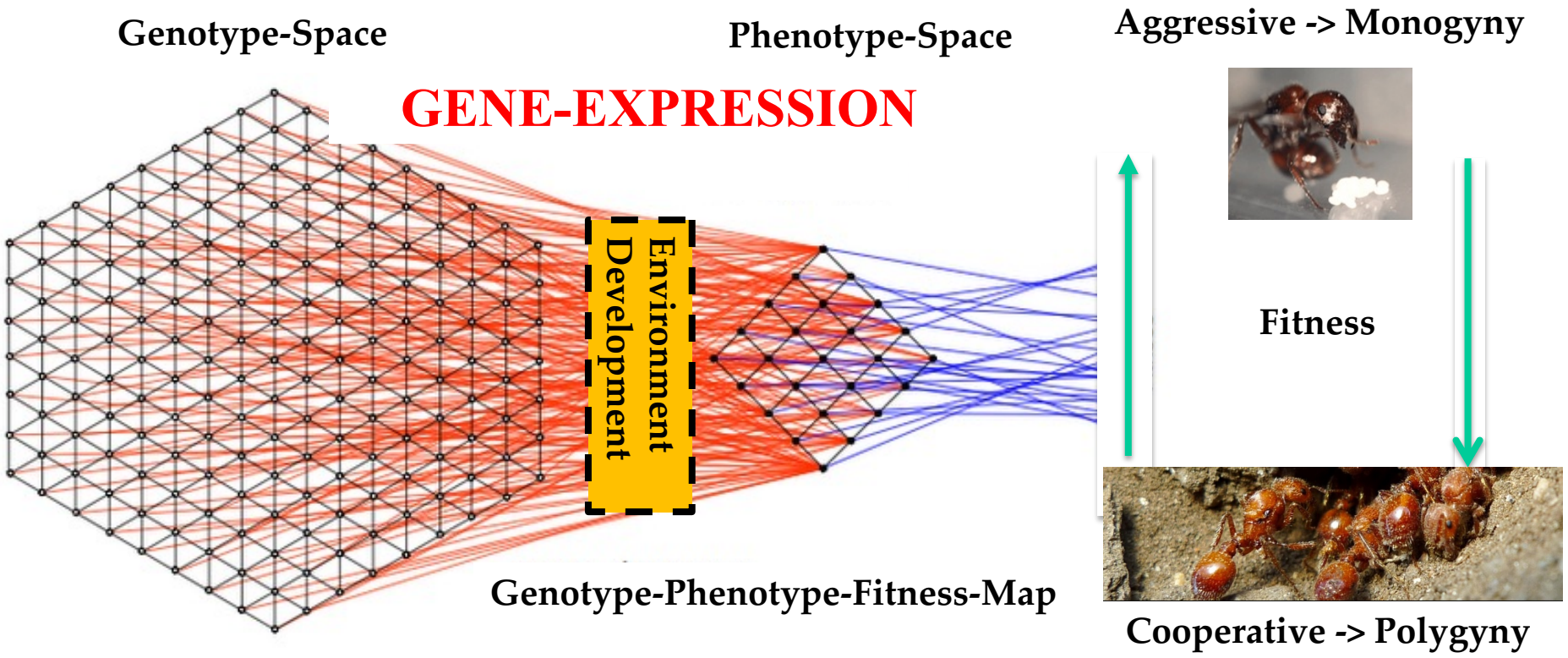
Genotype-Phenotype-Fitness-Map

Good flier



Bad flier

…or behavior

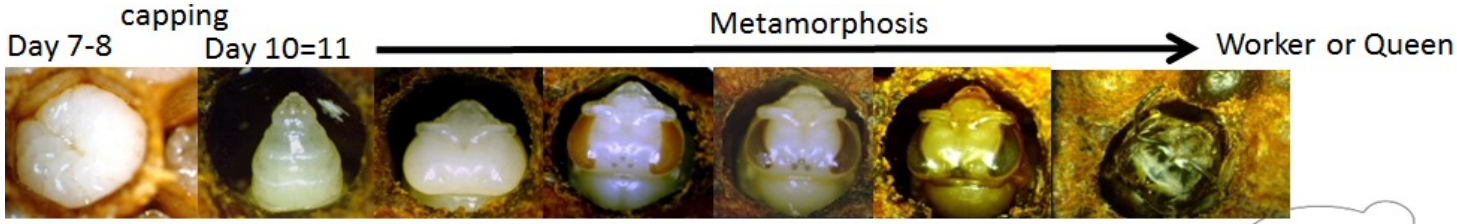

Genotype-Space
Phenotype-Space
GENE-EXPRESSION
Environment
Development
Genotype-Phenotype-Fitness-Map
Aggressive -> Monogyny
Fitness
Cooperative -> Polygyny

…one more complication or something to study

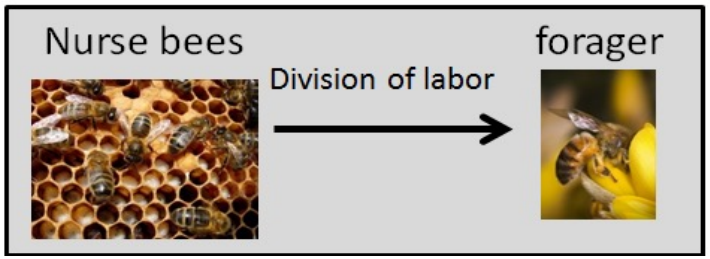
Fixed Genetic Differences

versus

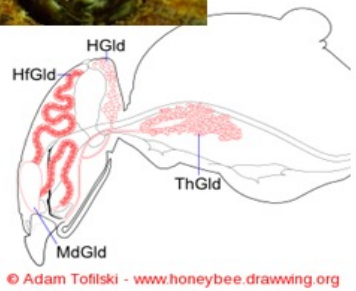
Phenotypic Plasticity



?



HGld : head salivary gland
THGld : thorax salivary gland
MDGld : mandibular gland
HfGld : hypopharyngeal gland



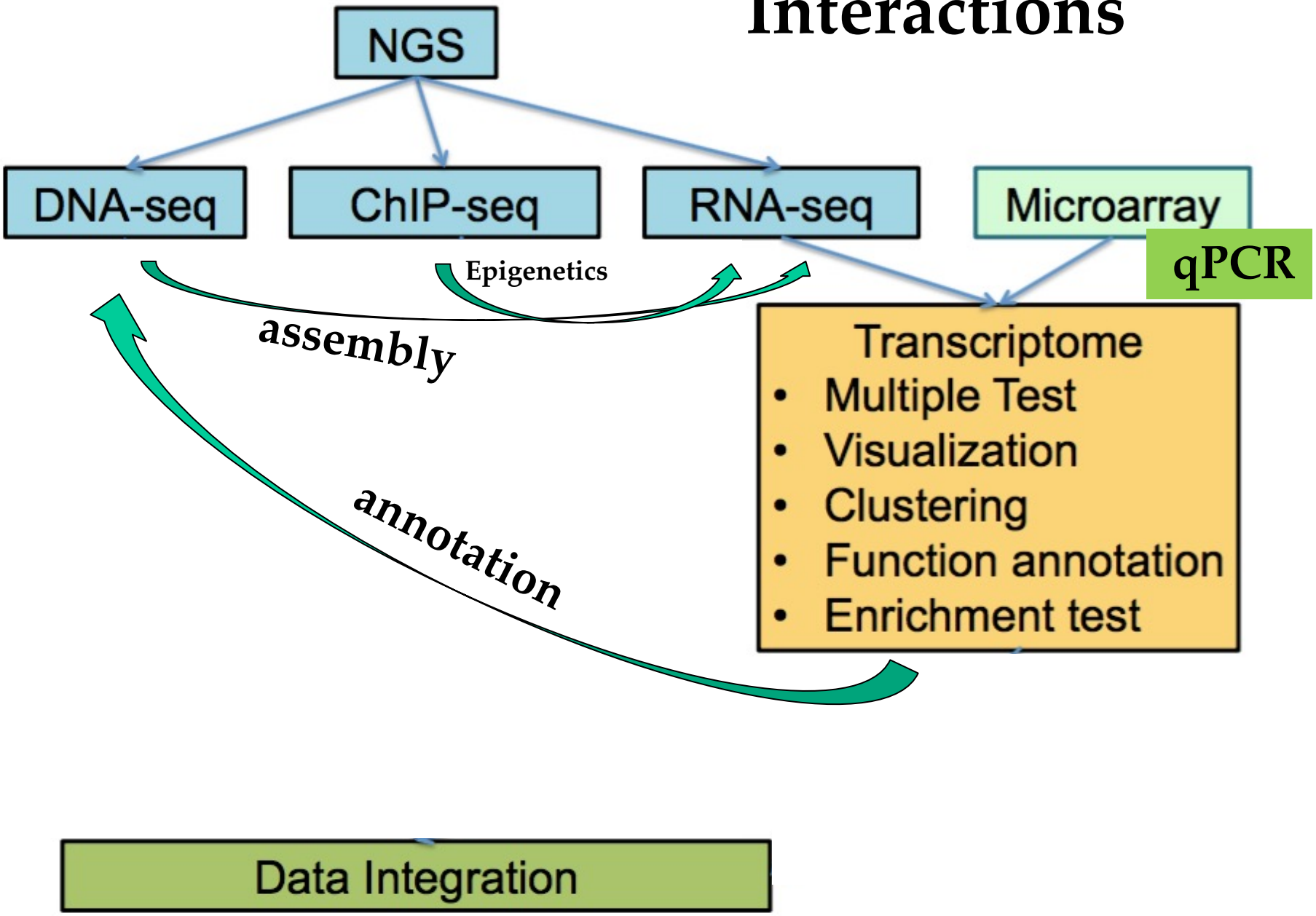
Any Questions?

Phenotypic plasticity is…

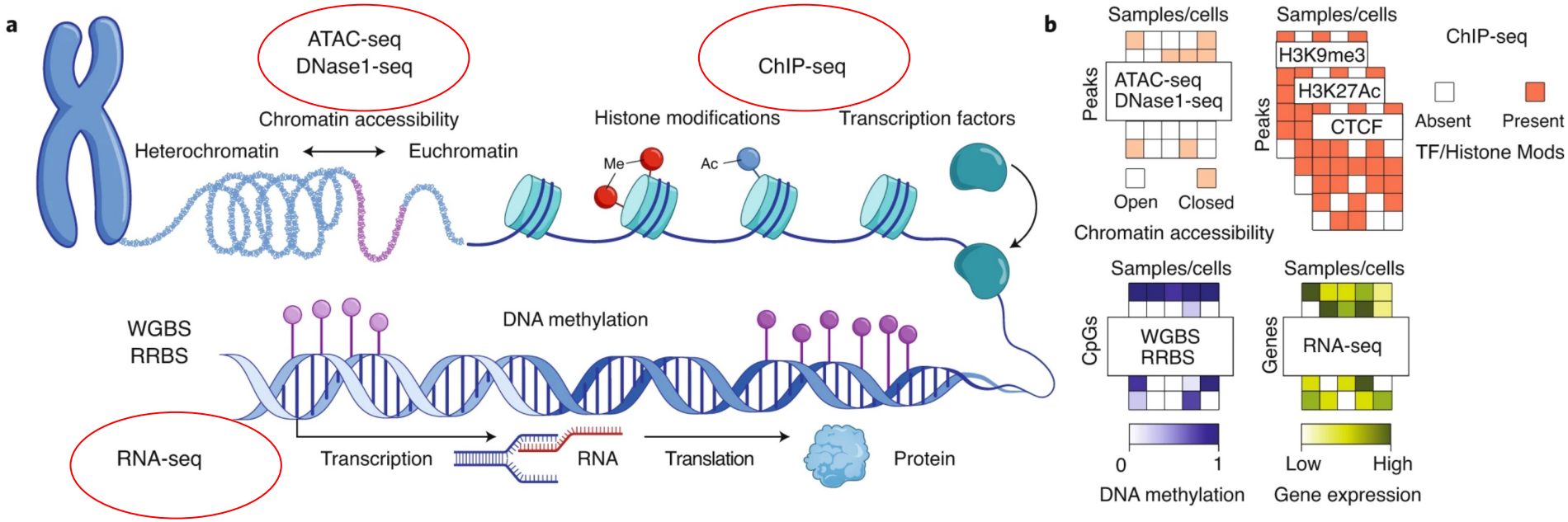
Phenotypic Plasticity is …

- based on random variation in individuals
- is based on gene regulation
- is based on genetic variation in the coding sequence of a gene
- generating different phenotypes from the same genotype, depending on the environment
- the result of differential gene expression

Interactions



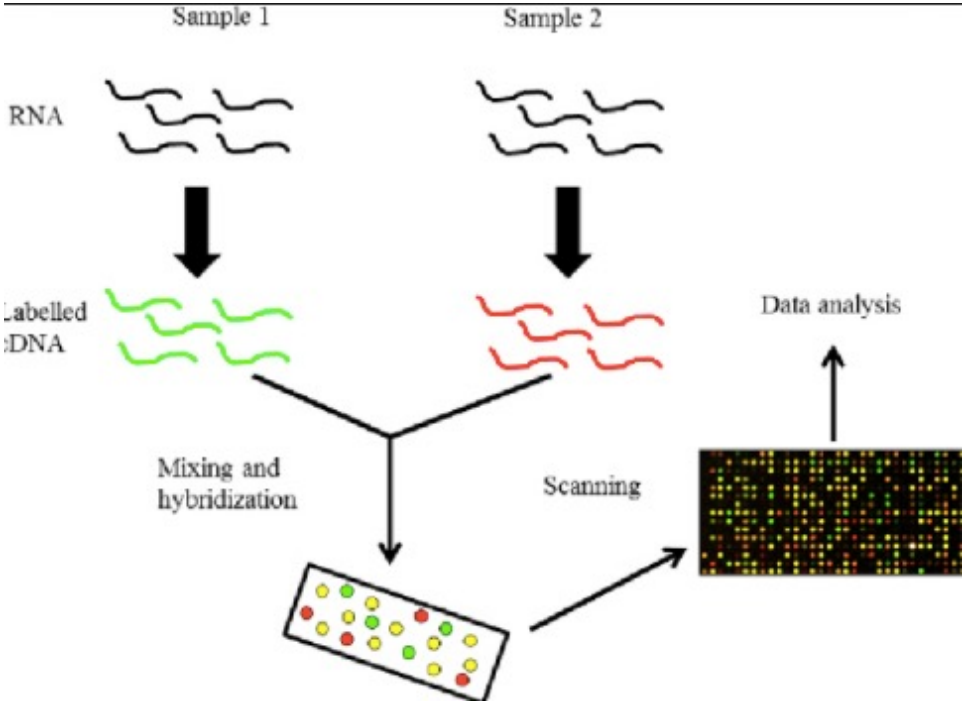
Methods to measure/estimate Generegulation



MICROARRAY –DNA-DNA HYBRIDISATION BASED

RNA -> labeled DNA

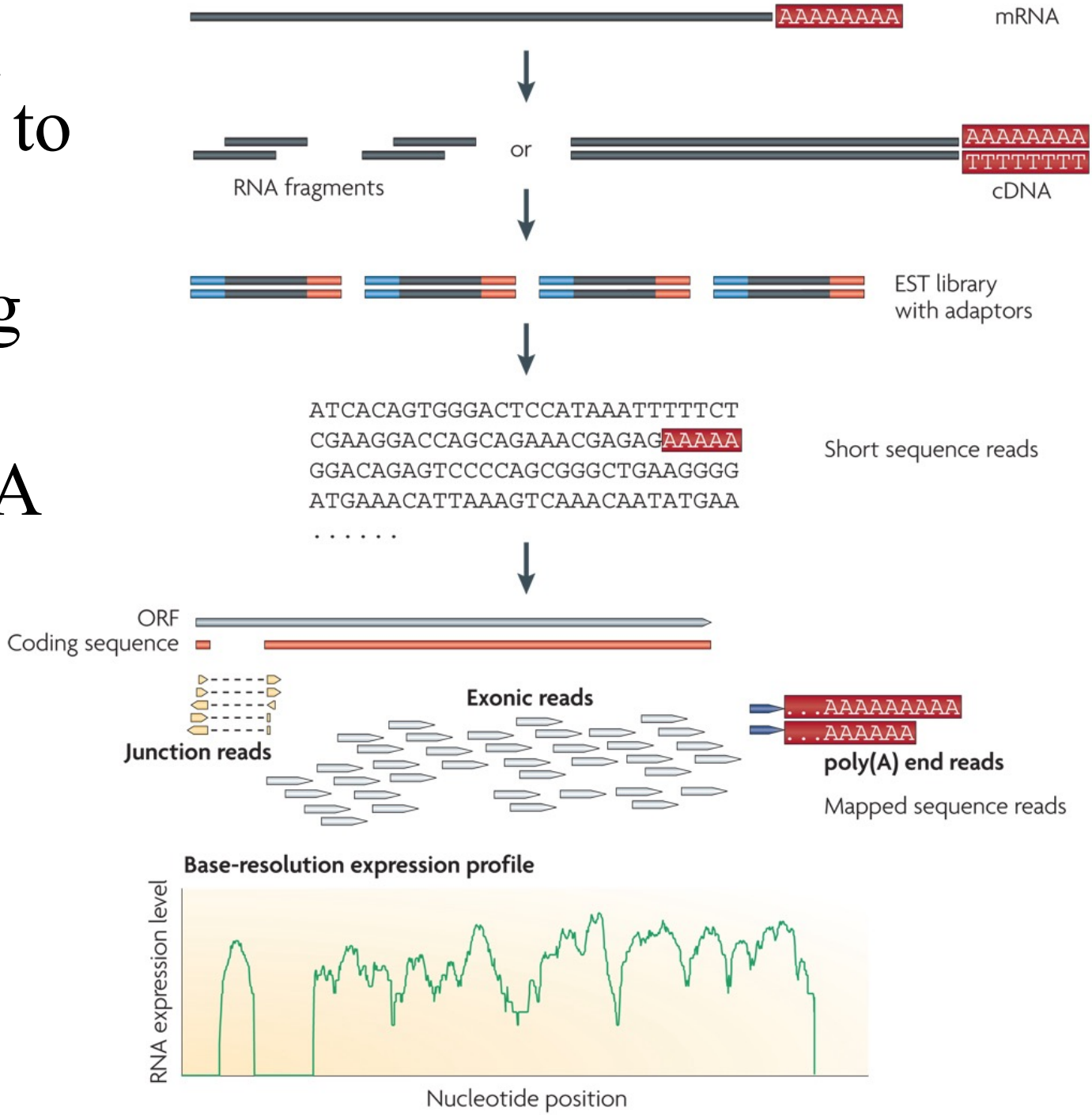
Microarray



Premade Chip with short Oligos

RNA-Seq very similar to Genome-sequencing

RNA-cDNA



What are the advantages of RNA-Seq vs Microarray?

- Prior sequence (genome or transcriptome) not required
- Microarrays are costly and time consuming to design and produce and are just for one species
- Greater dynamic range and sensitivity than microarrays
- Improved ability to discriminate regions of high sequence identity
- Greater multiplexing of samples

Typical RNA-Seq Strategy

Library Construction

~300 bp insert size

Sequencing

Illumina HiSeq 2000
•75-10 bp paired end reads

Assembly
or Alignment

ABySS

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

De Novo Contigs (shotgun assembly)
<u>or</u>
Alignment to genome

Transcripts found by RNA-Seq

- **Sequencing cDNA (Reverse transcription)**
- **mRNA vs unprocessed RNA (depends on library prep but you want to get rid of the ribosomal RNAs (PolyA selection)**
- **Annotate coding SNPS**
- **Transcript isoforms**
- **Splice junctions**
- **Count abundance of transcripts**

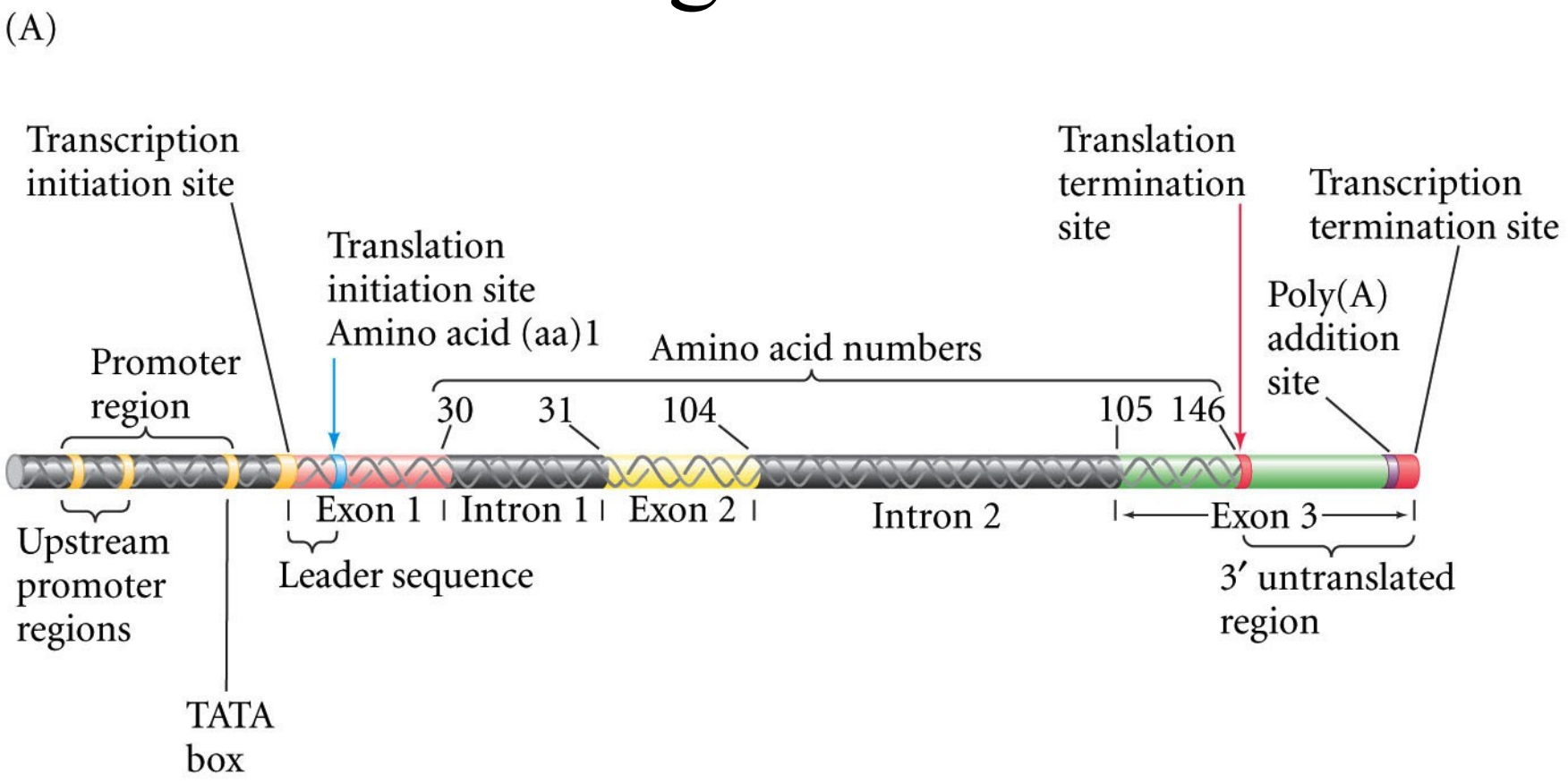
cDNA generation and what it is good for?

(but this could be another source for errors)

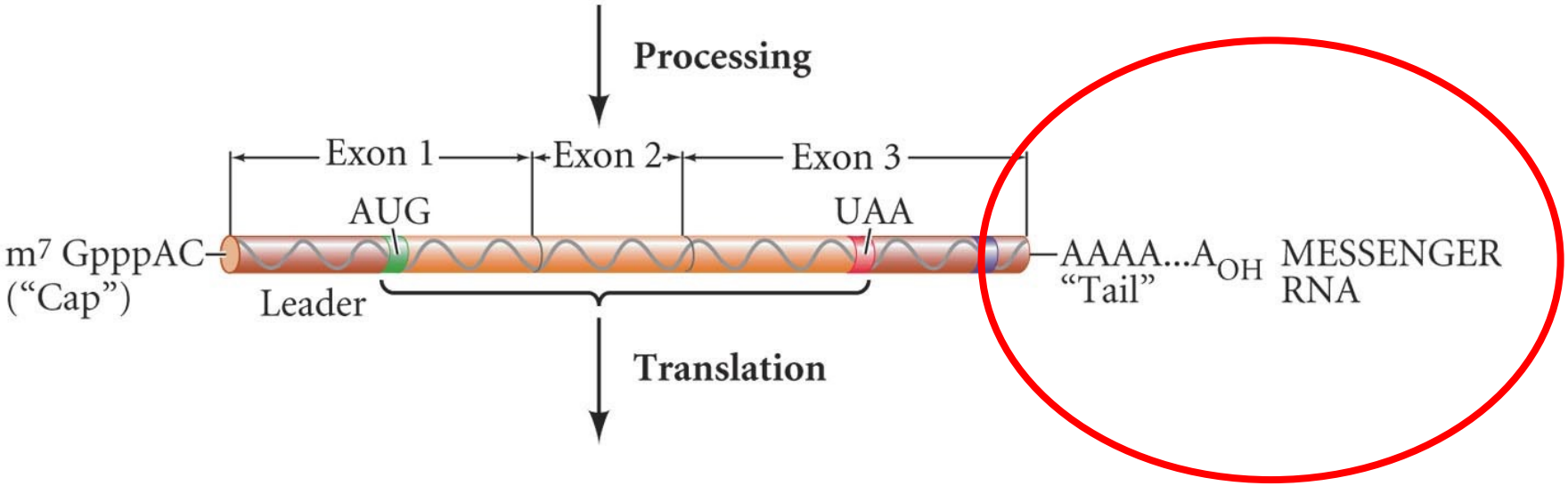
- The enzyme **reverse transcriptase** (**RT**) is used to generate complementary DNA (cDNA) from an RNA template.

https://en.wikipedia.org/wiki/Reverse_transcriptase

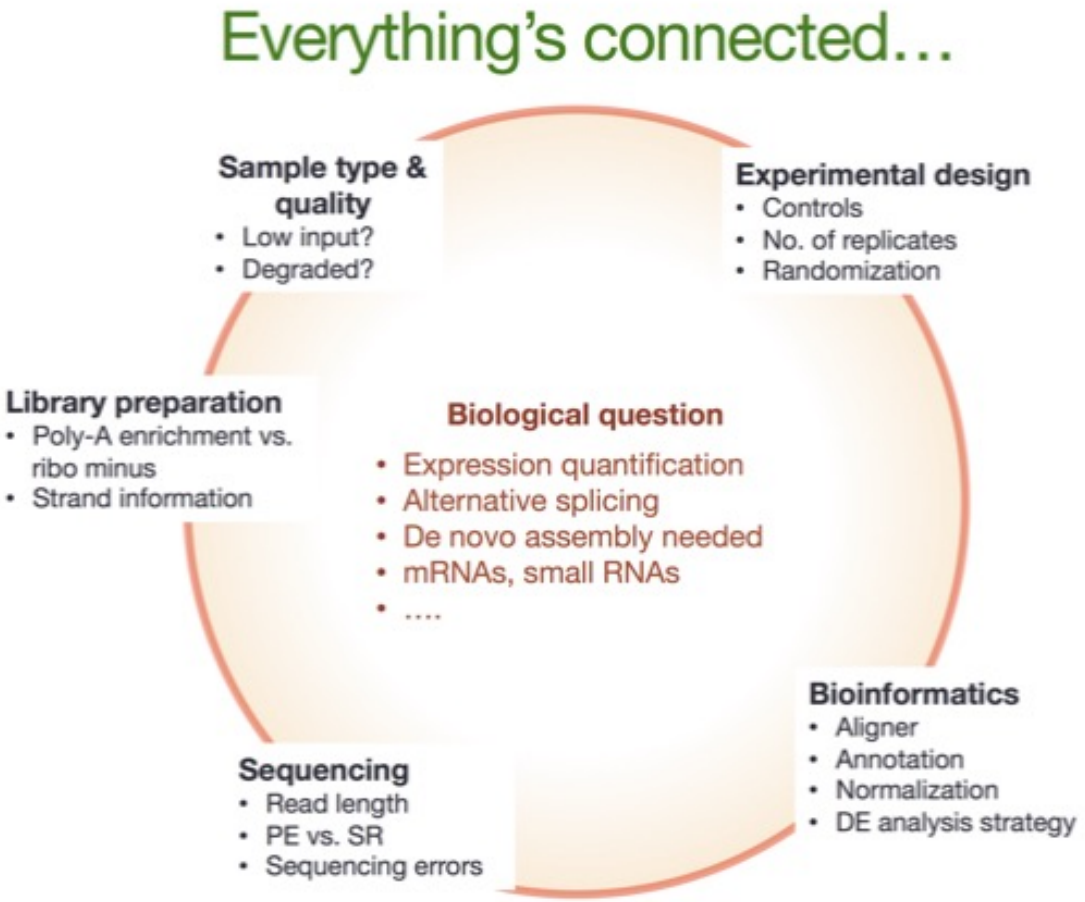
Review: Elements of protein coding genes



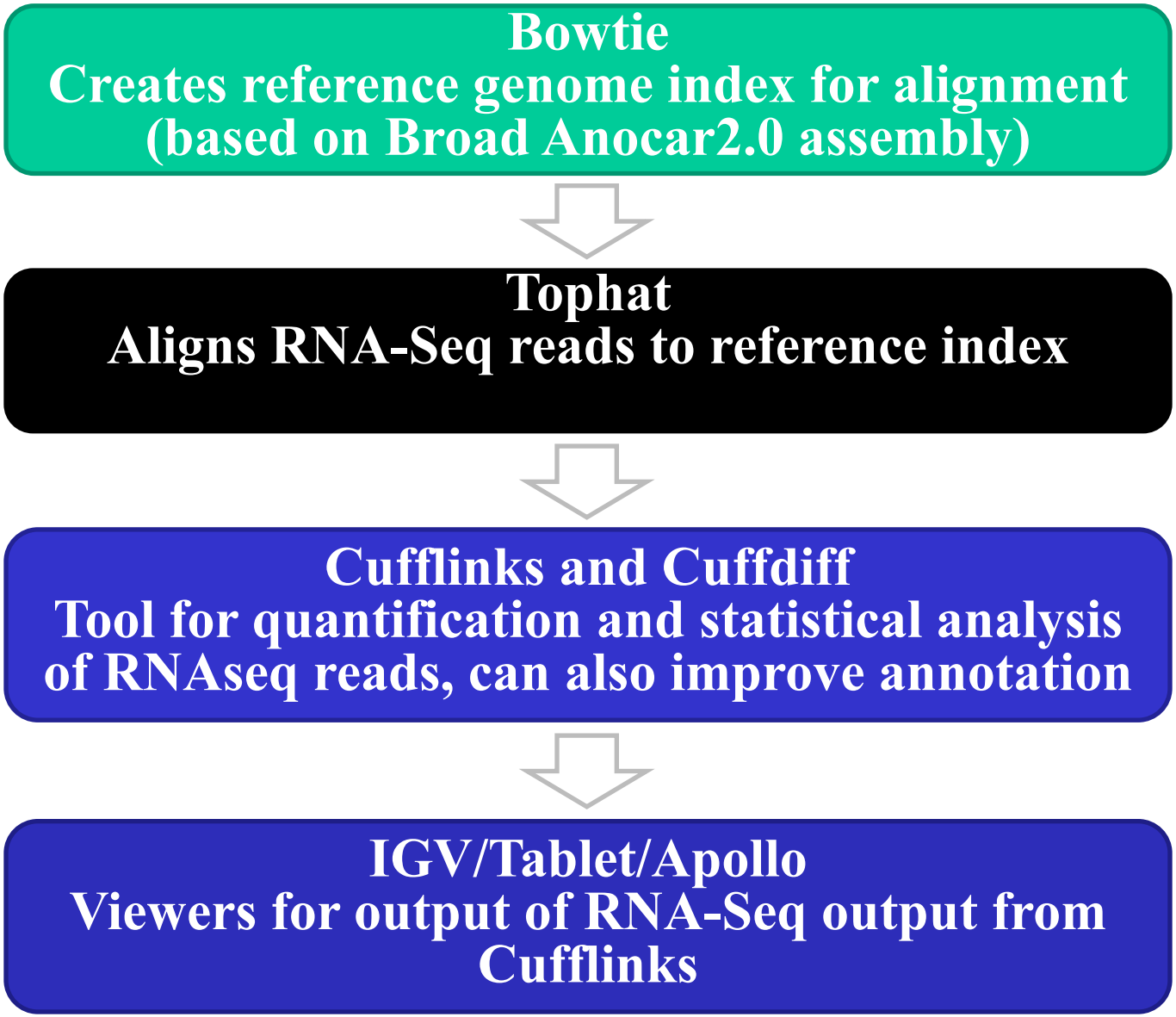
Review: Processed transcript



RNAseq - an introduction

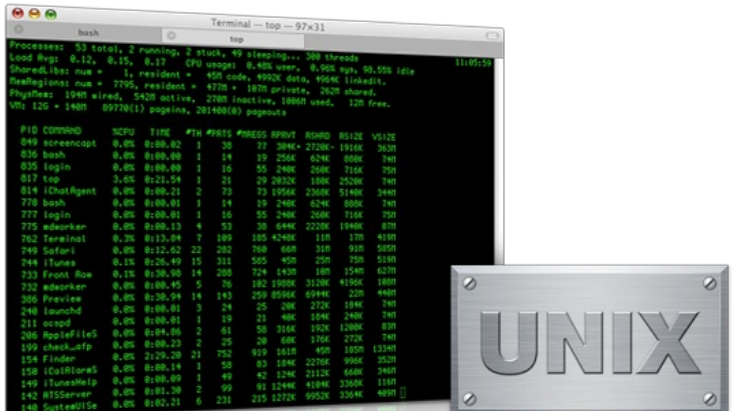


https://galaxyproject.org/tutorials/rb_rnaseq/


Bowtie
Creates reference genome index for alignment (based on Broad Anocar2.0 assembly)
Tophat
Aligns RNA-Seq reads to reference index
Cufflinks and Cuffdiff
Tool for quantification and statistical analysis of RNAseq reads, can also improve annotation
IGV/Tablet/Apollo
Viewers for output of RNA-Seq output from Cufflinks

Tools for RNA-Seq Analysis (mostly UNIX/Linux)

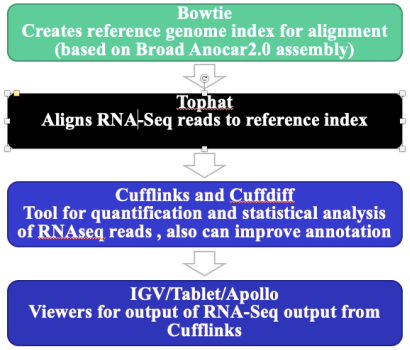
- Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
- Tophat: http://tophat.cbcb.umd.edu/
- Samtools: http://samtools.sourceforge.net/
- Picard tools: http://picard.sourceforge.net/
- Cufflinks: http://cufflinks.cbcb.umd.edu/
- Apollo: http://apollo.berkeleybop.org/current/index.html
- IGV: http://www.broadinstitute.org/igv/
- ABySS: http://www.bcgsc.ca/platform/bioinfo/software/abyss



Resources Needed

- Genomic sequence in fasta format
  - For alignment of RNA sequence
- Annotations in GTF (General Feature Format) format
  - For binning RNAseq reads into transcripts for quantitative analysis, as well as to help with annotation
- RNAseq reads in FASTQ format (sequenced and quality score)

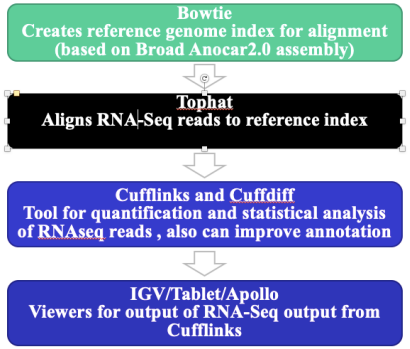
Bowtie



Creates reference genome index for alignment

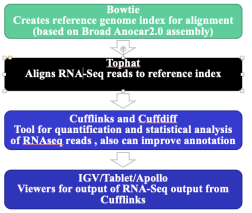
- Commands needed:
  - Bowtie-build –f (for fasta file) reference_in.fa out_file_name
  - Runtime: ~4 hrs for vertebrate genome
- Output:
  - 6 files that will be used by tophat

Tophat



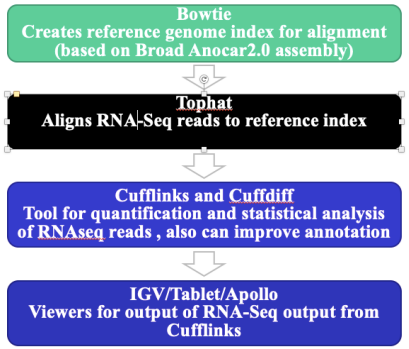
Aligns RNAseq reads to reference index

- Commands needed
  - tophat –r (average gap distance for paired reads) –o (output directory) reference_input (name you gave output in bowtie) RNAseq_reads_1.fastq RNAseq_reads_2.fastq
  - Runtime: multiple days
- Output
  - Read allignments in .bam or .sam format
  - Junctions, insertion and deletion as .bed format



Tophat aligns reads to reference genome

Cufflinks



Tool for quantification and statistical analysis of RNAseq reads

- Commands used:
  - Cufflinks –G (.GTF annotation file) input_file.sam
    - Basic command that gives **FPKM** values and errors for each transcript in a tab delimited file
    - Runtime: several hours (goes for all cuff programs)
  - Cuffcompare reference_annotation.gtf input_file_tissue1.gtf (from cufflinks) input_file_tissue2.gtf
    - Secondary analysis using output from cufflinks. Gives data on pooled transcript levels, a .GTF file for annotation based on all tissues, major isoforms found (if annotated in reference .GTF file)

Why can`t we just simply compare raw read numbers, or asked the other way around whar factors besides expression influence read counts?

What we need to control for!

- Sequencing depth
- Gene length

RPKM, FPKM or TPM

- Reads/Fragments per kilobase million – difference: R is for single reads and F for paired end reads
- Transcripts per million – makes samples with different reading depth more comparable

- For a usefull discussion see - https://btep.ccr.cancer.gov/question/faq/what-is-the-difference-between-rpkm-fpkm-and-tpm/

RPKM

(FPKM just avoids counting fragments twice in paired end sequencing)

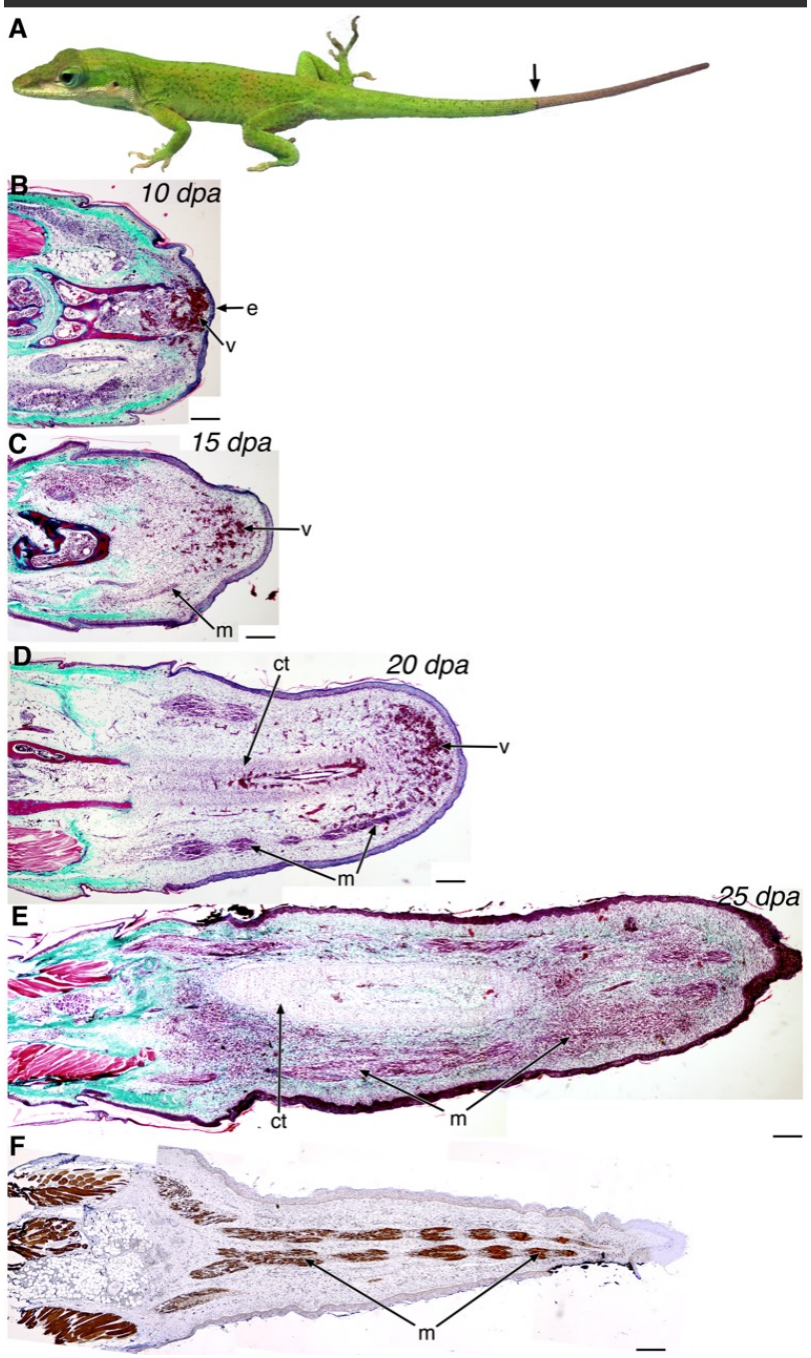
- **Count up** the total reads of your sample and divide that number by 1,000,000 – this is our "per million" **scaling factor** (standardizes for differences in read depth)
- **Divide 1** the read counts for your gene of interest by the "per million" **scaling factor**, giving you reads per million (RPM) for your gene.
- **Divide 2** the RPM values by the length of the gene, in kilobases. This gives you RPKM.

A little math (adding and dividing) …

- FPKM of 10 means :
- = *200.000.000* reads per sample /1.000.000

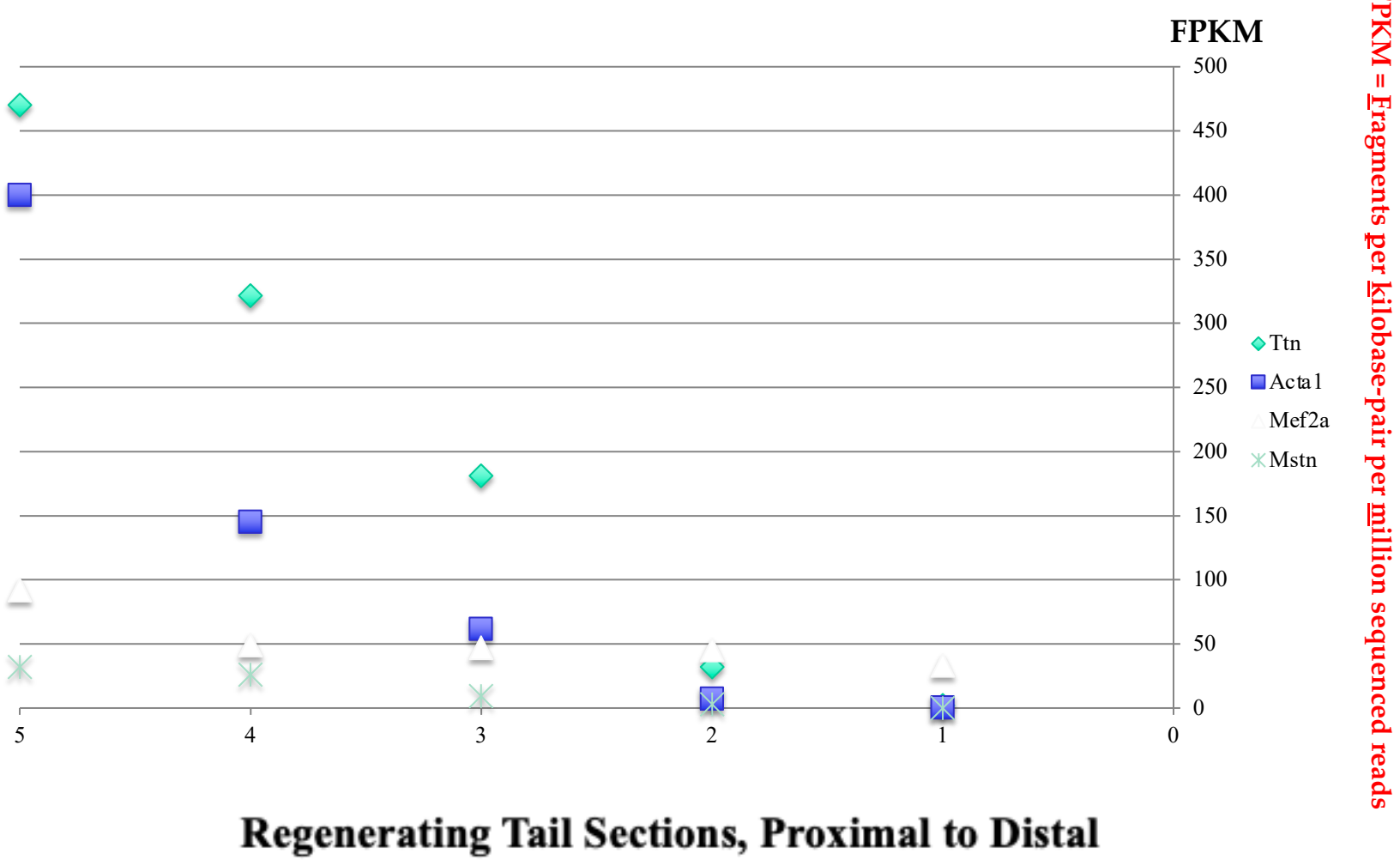
  Read count scaling factor = *200*
- Read count for gene A = *2000*
- Rpm for gene A =2000/200 = 10
- The gene is *1 kb* in length – RPKM for Gene A is 10/1 = 10

An Example:
Anolis carolensis
Tail-Regeneration



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105004

Cufflinks quantifies expression level based on FPKM



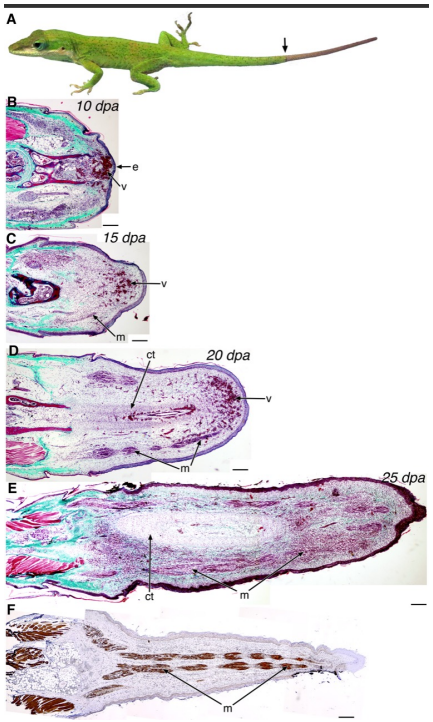
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105004

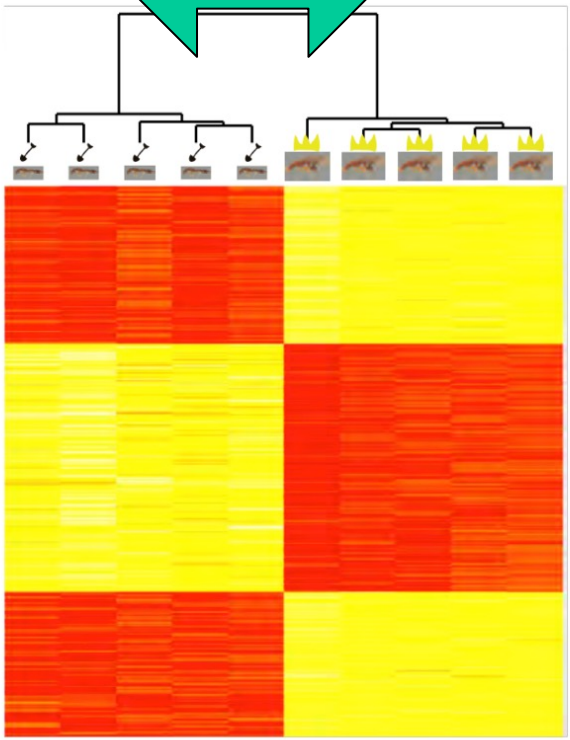
WHAT TO DO WITH TRANSCRIPTOMIC DATA?

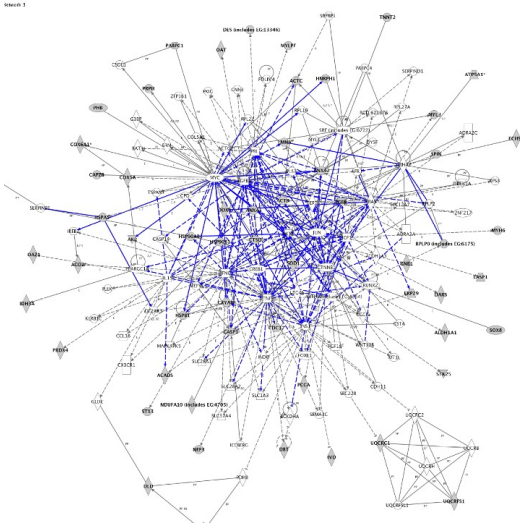
time

types

coexpression

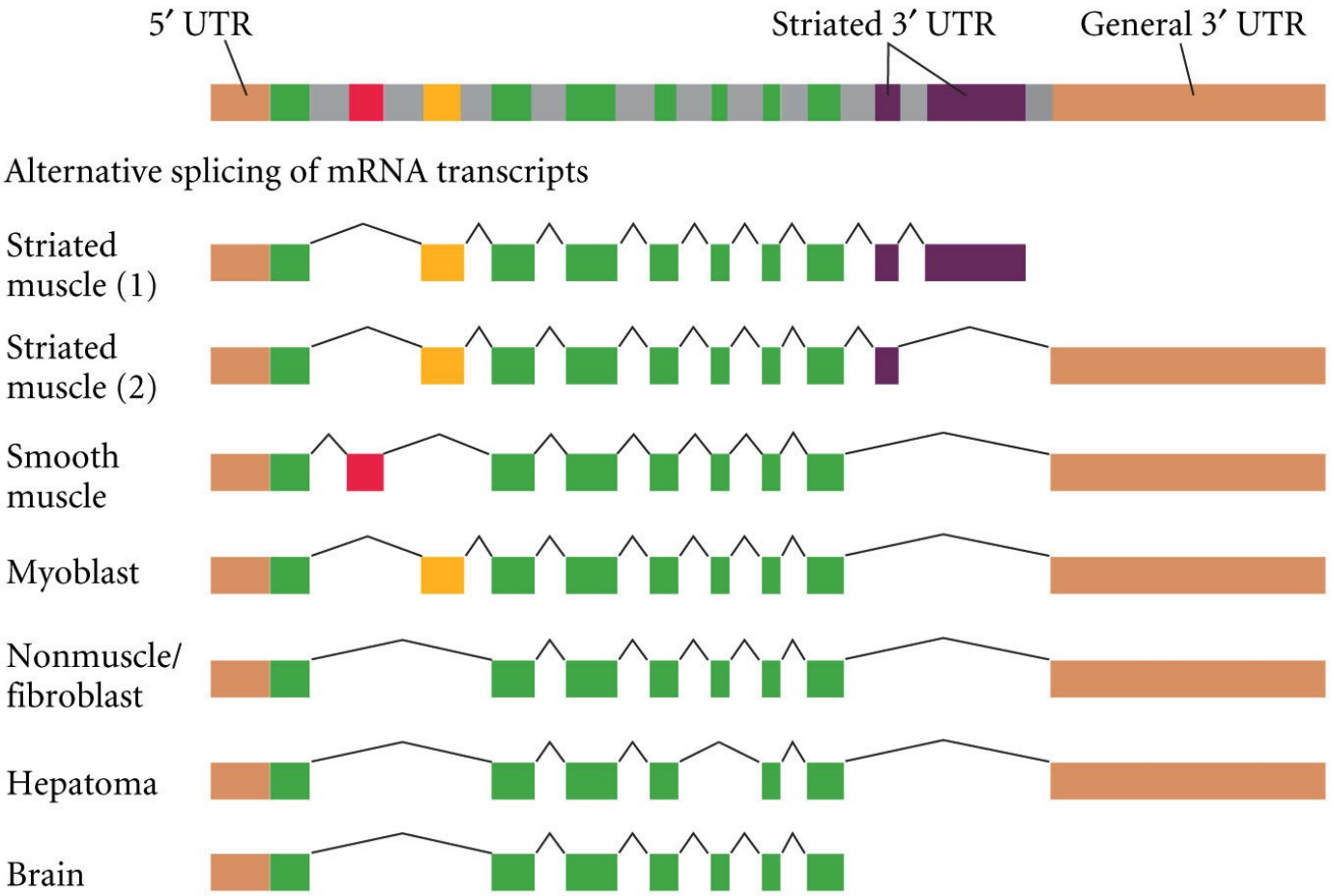






Gene expression/co-expression changes over time and between types!

Review: Alternate splicing

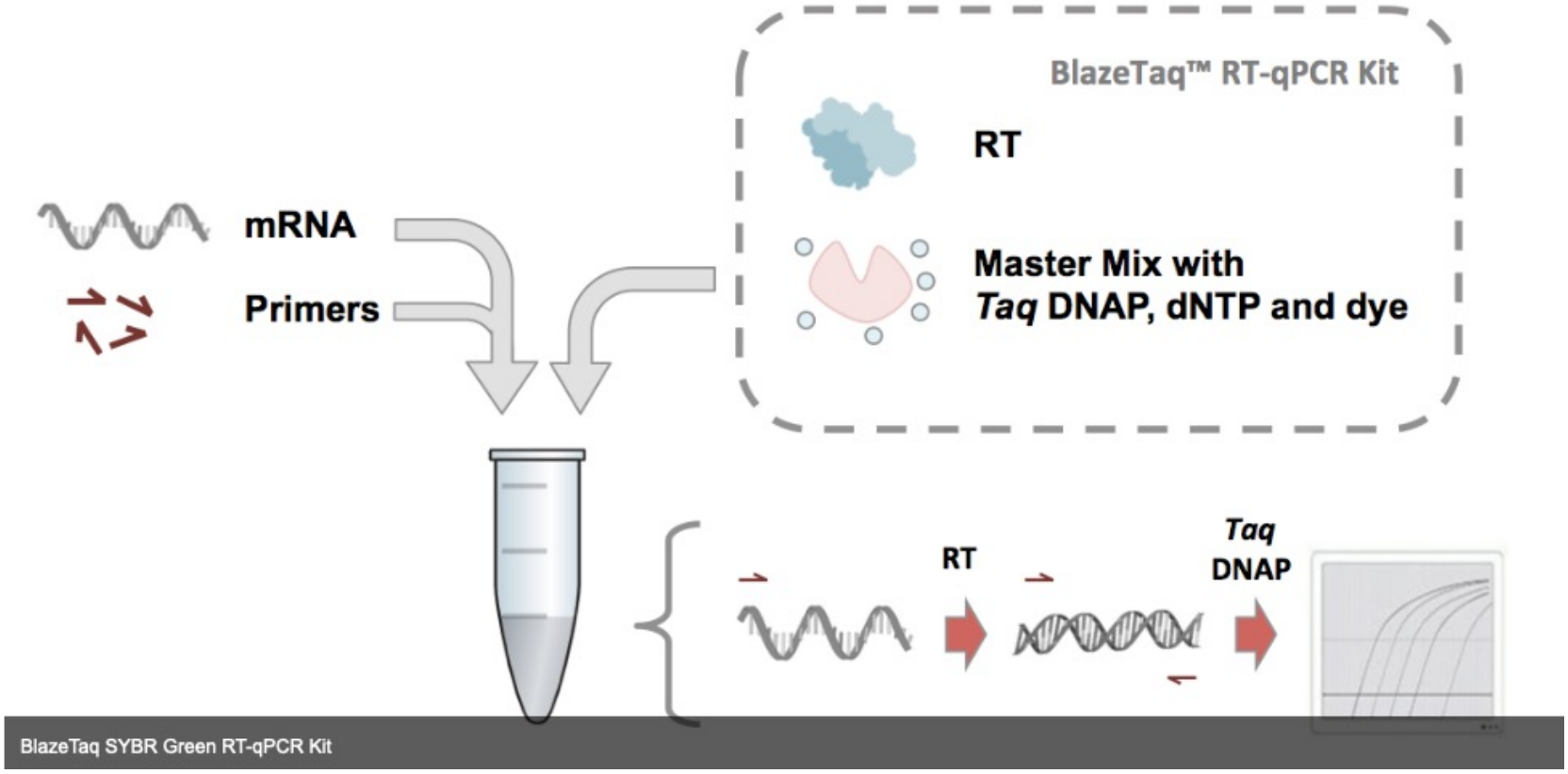


DEVELOPMENTAL BIOLOGY, Eighth Edition, **Figure 5.29** © 2006 Sinauer Associates, Inc.

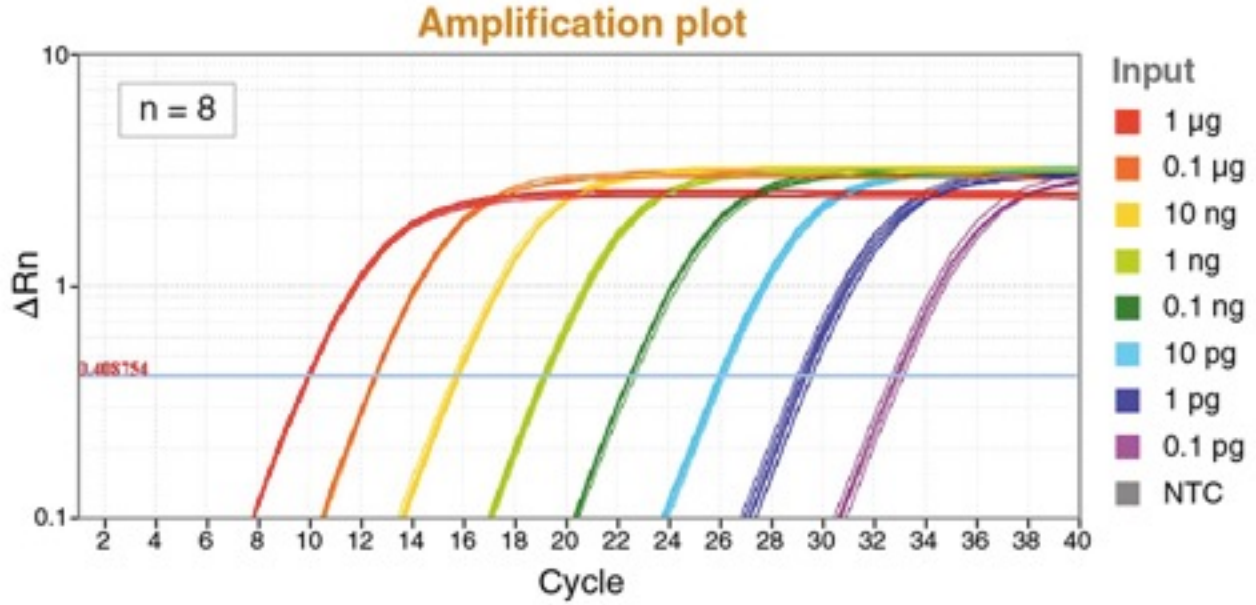
Validation of your RNASeq results: different methods

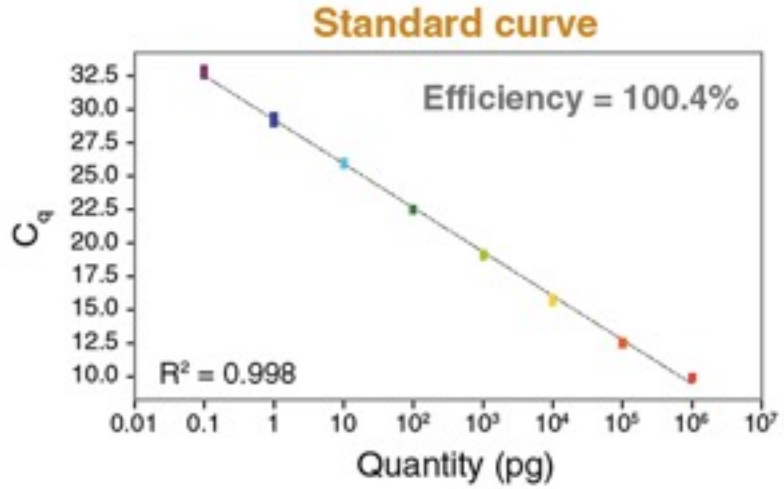
- Rerun the same sample on the same instrument
- Use multiple sequencing methods for the same sample.
  - Quantatative
    - qPCR
    - RNA seq
    - Microarrays
  - Qualitative
    - PCR
    - In situ
- Run multiple samples of same tissue type

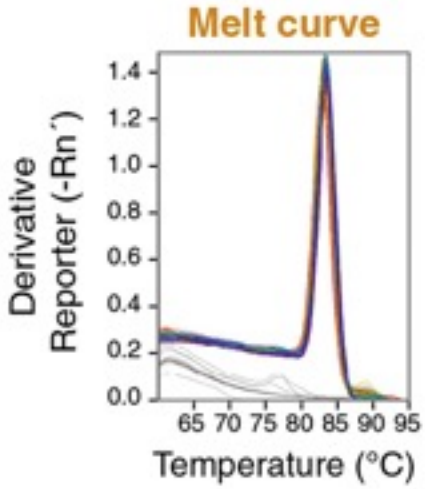
How does RT-qPCR work?



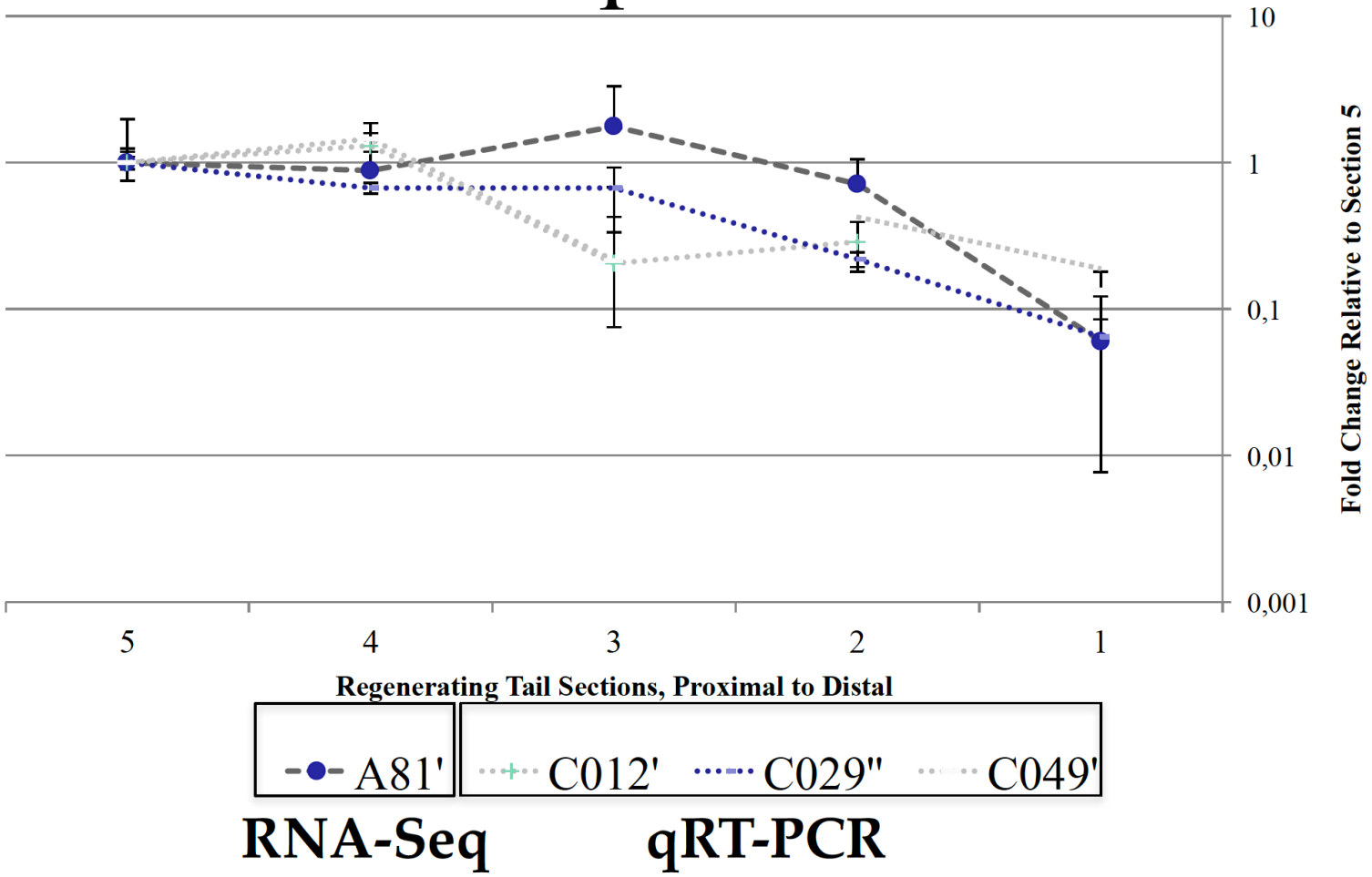
How does RT-qPCR work?







Validation of RNA-Seq results with qRT-PCR



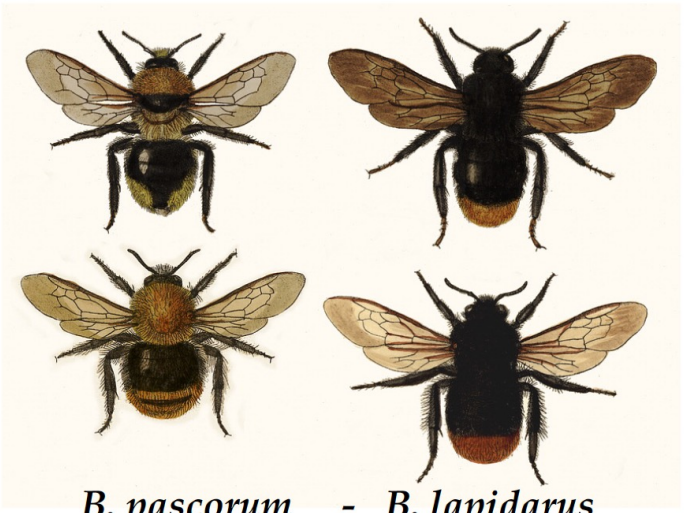
Anolis - lizard https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105004

Another example – Queens vs Workers

Social Parasitism

Wasps, Polistes
Polistes sulcifer, **is an obligate social parasite, whose only host is *P. dominula*.**

P. campestris - *P. rupestris*



B. pascorum - *B. lapidarus*

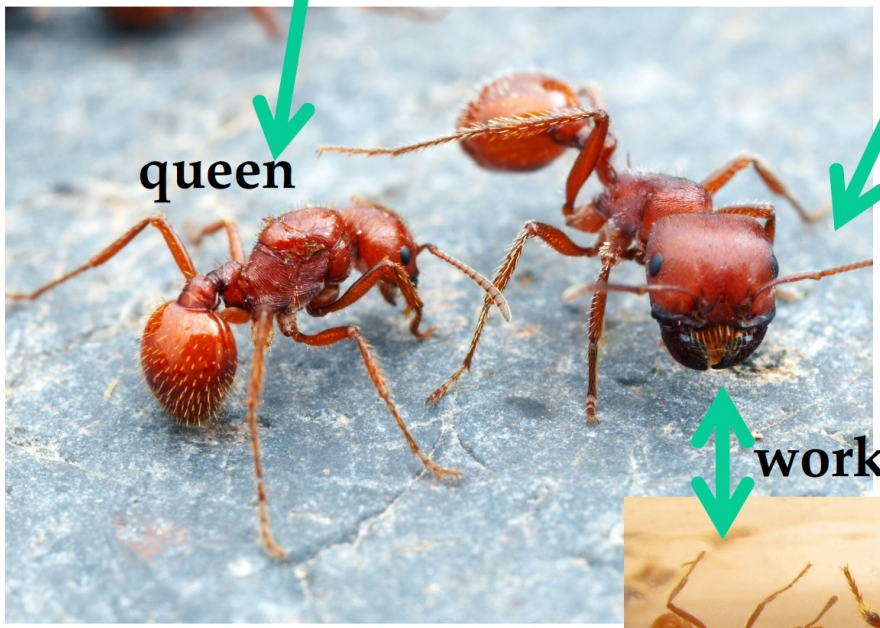
Bumble bees, *Bombus/Psiturus*

Ants, *Polyergus/Formica*

Social parasitism has evolved multiple times independently in the eusocial insects: 3 times in wasps at least 12 times in bees 7 times in Allodapinae 2 times in Halictidae and many times in the ants.

Pogonomyrmex anergism/colei
parasite

Pogonomyrmex barbatus/rugosus
host

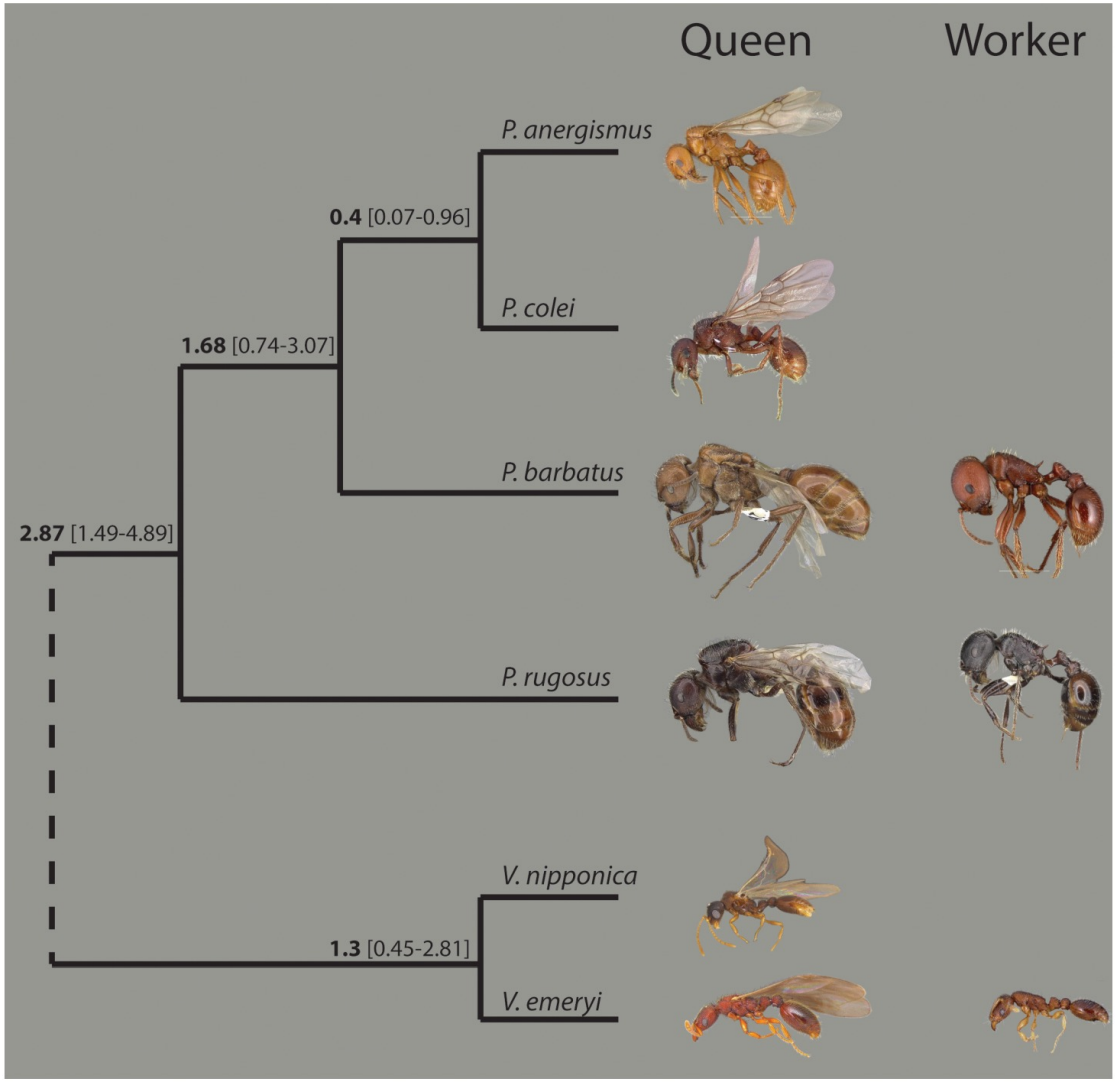


worker

queen

***Pogonomyrmex* inquilines**
P. colei* + *P. anergism

77



NEW

NEW

Smith et al. 2013

NEW

NEW

Reference Genome

Social Parasites --- Hosts

78

Summary of Genome Sequencing

Species	Fold coverage	# contigs/scaffold/# contigs/scaffolds covered 90%**	No. of genes*	GC content	annotation	Source
Pogonomyrmex barbatus	12 (Roche 454)	4646	17,177	33.9	denovo	Smith et. al 2011
P. rugosus	71	3861	17,093	33.9	mapped to P. barbatus	this study
P. colei	55	3467	16,910	33.9	mapped to P. barbatus	this study
P. anergismus	40	3743	16,968	33.9	mapped to P. barbatus	this study
Vollenhovia emeryi	11 (Roche 454)	46989	26,902	42.2	denovo	this study
V. nipponica	94	40233	25,102	42.2	mapped to V. emeryi	this study

* for mapped genomes number of genes is defined as genes that have at least a 90% coverage of the cds.

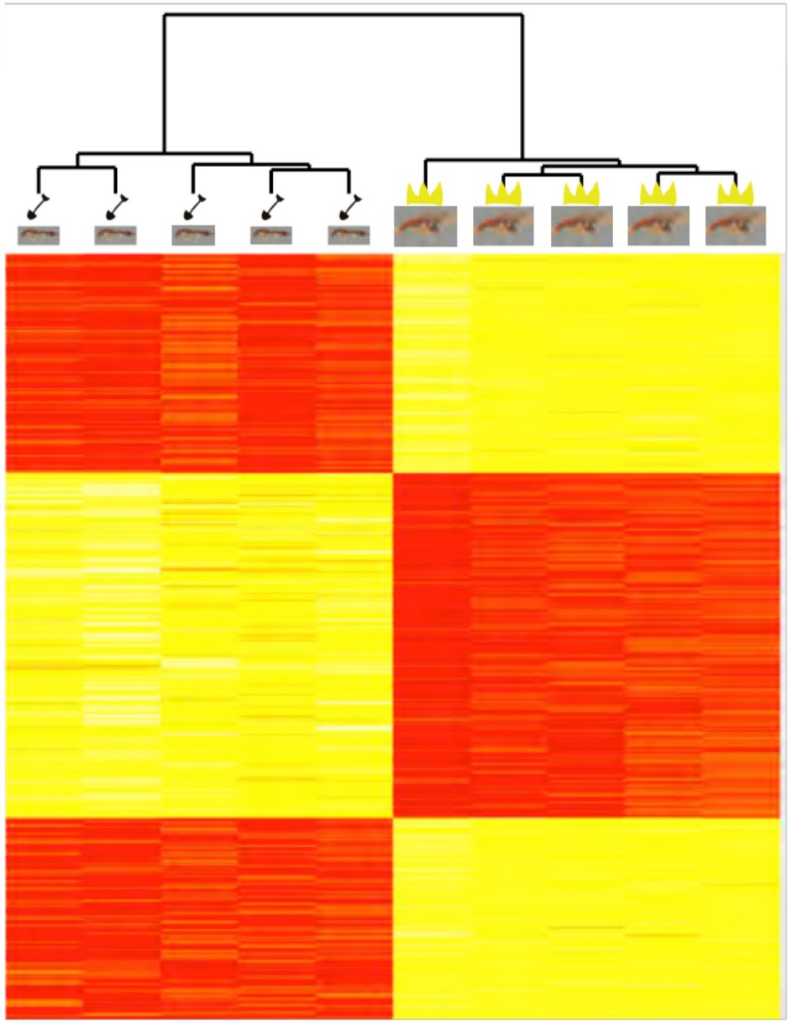
** number of scaffolds for de novo genomes and number of scaffolds covered at least by 90% for the mapped genomes

Two approaches to search for worker genes

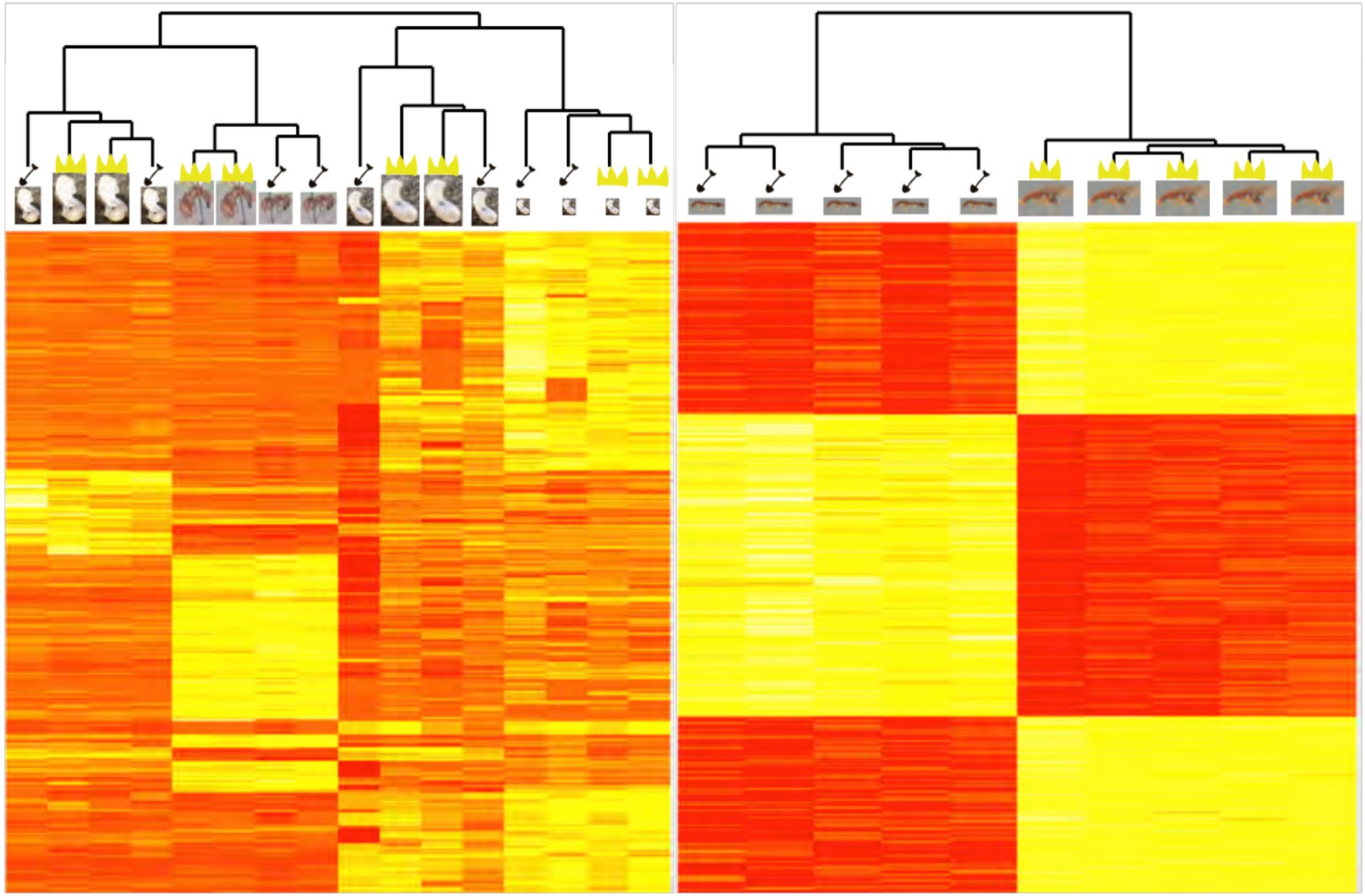
- We would expect that in species that have lost their worker caste (phenotype) genes that are worker specific would be lost as well

- Genes that show a significant worker bias in their expression pattern should evolve faster in social parasites since they are essentially now under neutral selection and if not necessary should also vanish.

Expression differences between worker and queens



Expression differences between worker and queens



Take Home: Expression is time and stage specific and it is important to take this into account!

Important for the RNAseq week: You must have installed Microsoft Excel. Alternatively, you must have access to Office 365 (available via the university license).
https://www.uni-muenster.de/IT/en/services/arbeitsplatz/software/office365.html?lang=en

Wieviel Prozent der Vorlesung haben sie verstanden?

61

34 Antworten

Mit Google Tabellen verknüpfen

Antworten möglich

Zusammenfassung | Frage | Einzelansicht

Wieviel Prozent der letzten Vorlesung haben sie verstanden?

Kopieren

34 Antworten