Gene expression & transcriptomics: Adaptive environmental variation

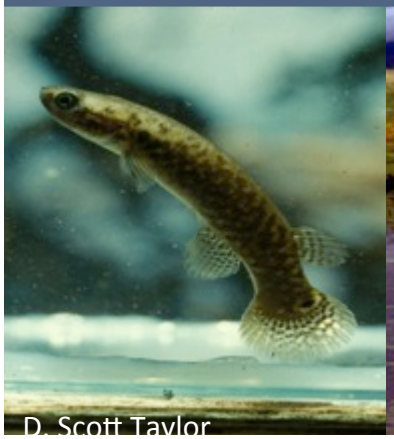
Joanna Kelley

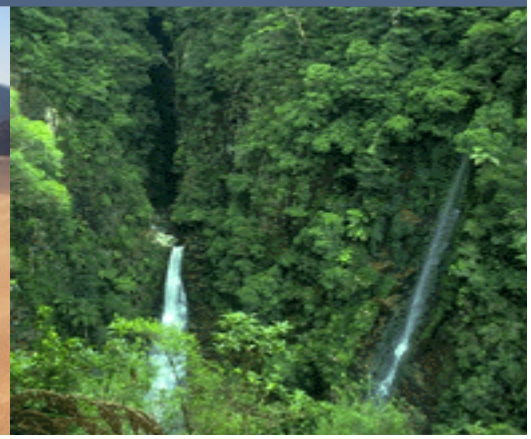
ConGen2015

4 Sept 2015



How do organisms diverge and adapt to the wide-range of environments they encounter?



D. Scott Taylor

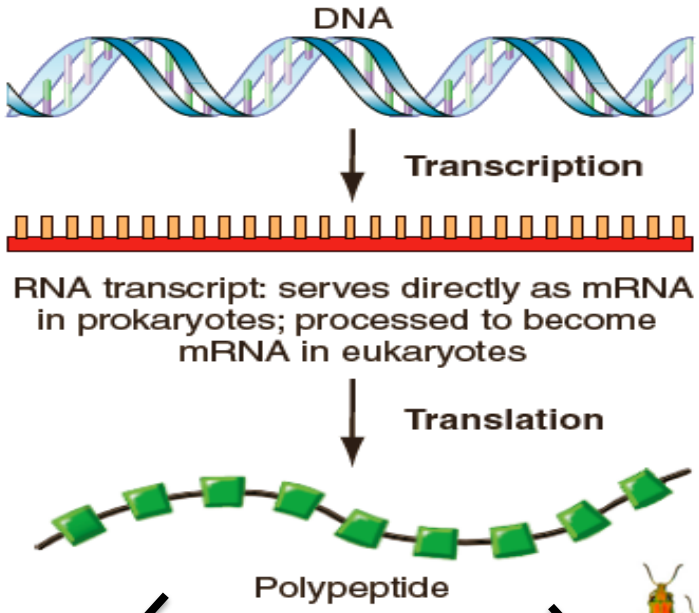
OMICS! approaches

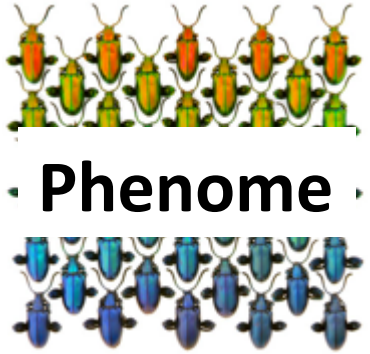
Genome

Transcriptome

Proteome

Metabolome

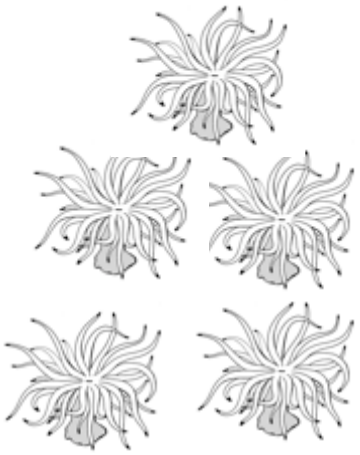




Phenome

Why study gene expression differences among individuals and populations?

- Insights into the molecular basis of phenotypic diversity
- Interpretation of patterns of expression variation in response to environmental conditions, disease, etc.
- Possible management decisions on how and where to manage or transplant populations

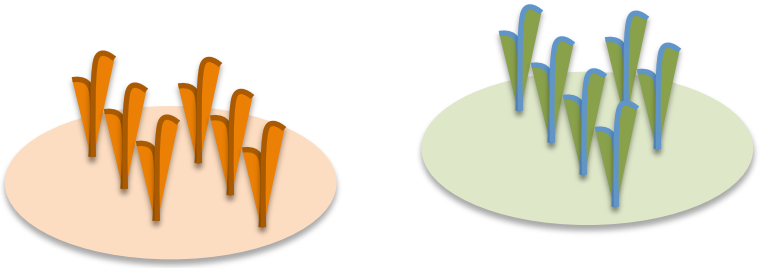


RNAseq to learn about climate change

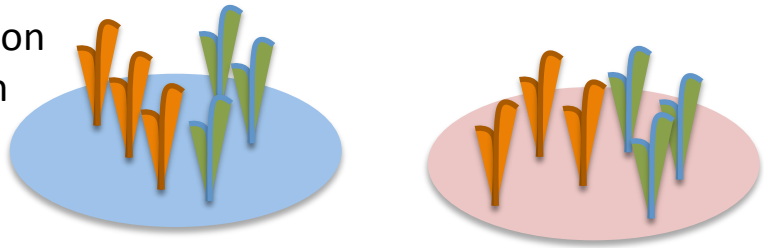
Genomic basis for coral resilience to climate change

Daniel J. Barshis[1,2], Jason T. Ladner, Thomas A. Oliver, François O. Seneca, Nikki Traylor-Knowles, and Stephen R. Palumbi

Wild



Common garden



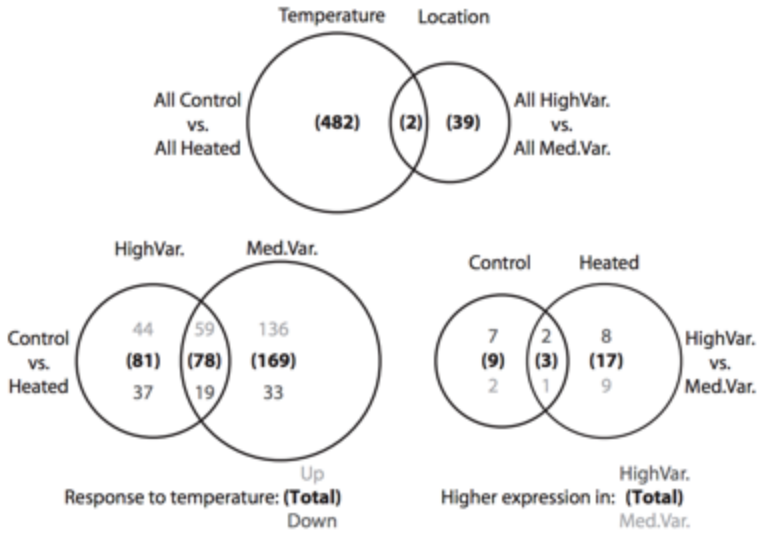


Fig. 2. Venn diagram showing the number of differentially expressed genes detected during analysis based on temperature, location, within-location temperature response, and within-treatment location differences. Bold numbers in parentheses represent totals and respective shades of gray denote up- vs. down-regulated or higher in HV vs. MV, respectively.

RNAseq for management decisions

Research article

Open Access

Transcriptomic response to heat stress among ecologically divergent populations of redband trout

Shawn R Narum* and **Nathan R Campbell**

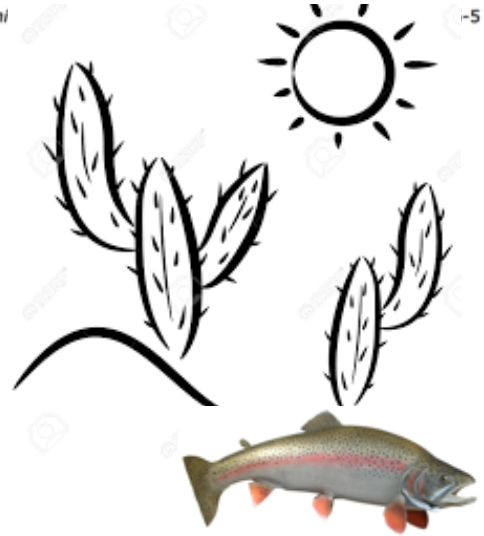
* Corresponding author: Shawn R Narum nars@critfc.org

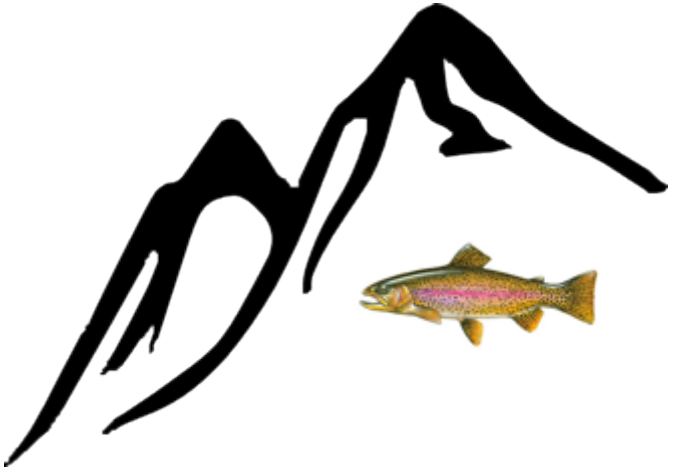
▾ Author Affiliations

Columbia River Inter-Tribal Fish Commission, 3059-F National Fish Hatchery Road, Hagerman 83332, ID, USA

For all author emails, please log on.

BMC Genomi -5





and F1s

RNAseq for management decisions

Research article

Open Access

Transcriptomic response to heat stress among ecologically divergent populations of redband trout

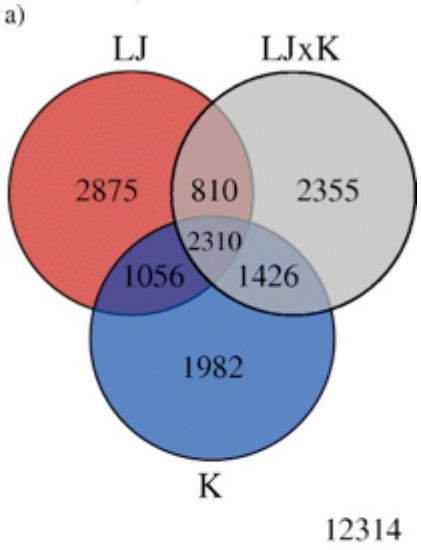
Shawn R Narum* and **Nathan R Campbell**

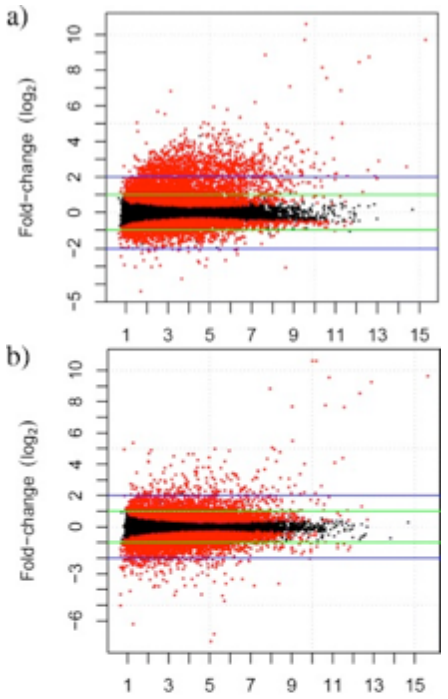
* Corresponding author: Shawn R Narum nars@critfc.org ▾ Author Affiliations

Columbia River Inter-Tribal Fish Commission, 3059-F National Fish Hatchery Road, Hagerman 83332, ID, USA

For all author emails, please log on.

BMC Genomics 2015, **16**:103 doi:10.1186/s12864-015-1246-5



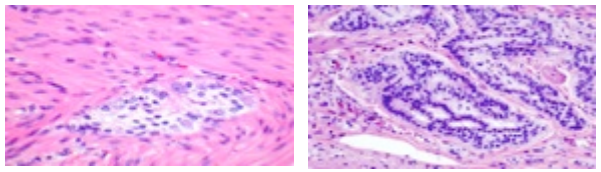


Approaches to measuring gene expression

- Single/few gene studies
  - Northern blots
  - qPCR
- Transcriptome (everything that is transcribed at a single time point in a specific tissue/cell)
  - Microarrays
  - RNA-sequencing (RNA-seq)
    - poly-A+
    - Ribo-minus

RNA sequencing experiment

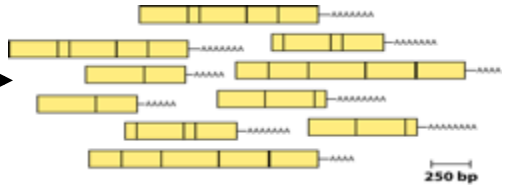
Samples of interest



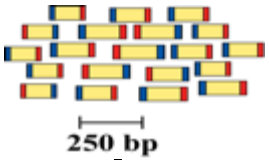
Condition 1 (normal colon)

Condition 2 (colon tumor)

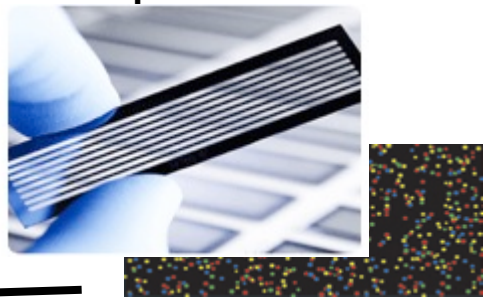
Isolate RNAs



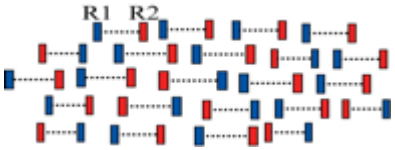
Library preparation



Sequence ends

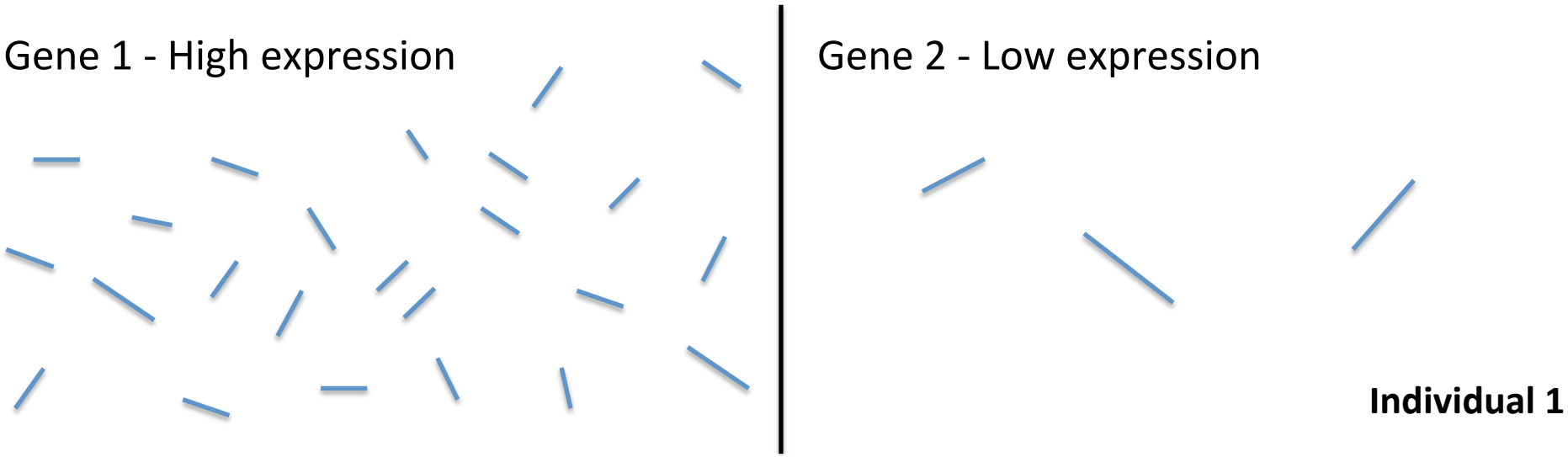


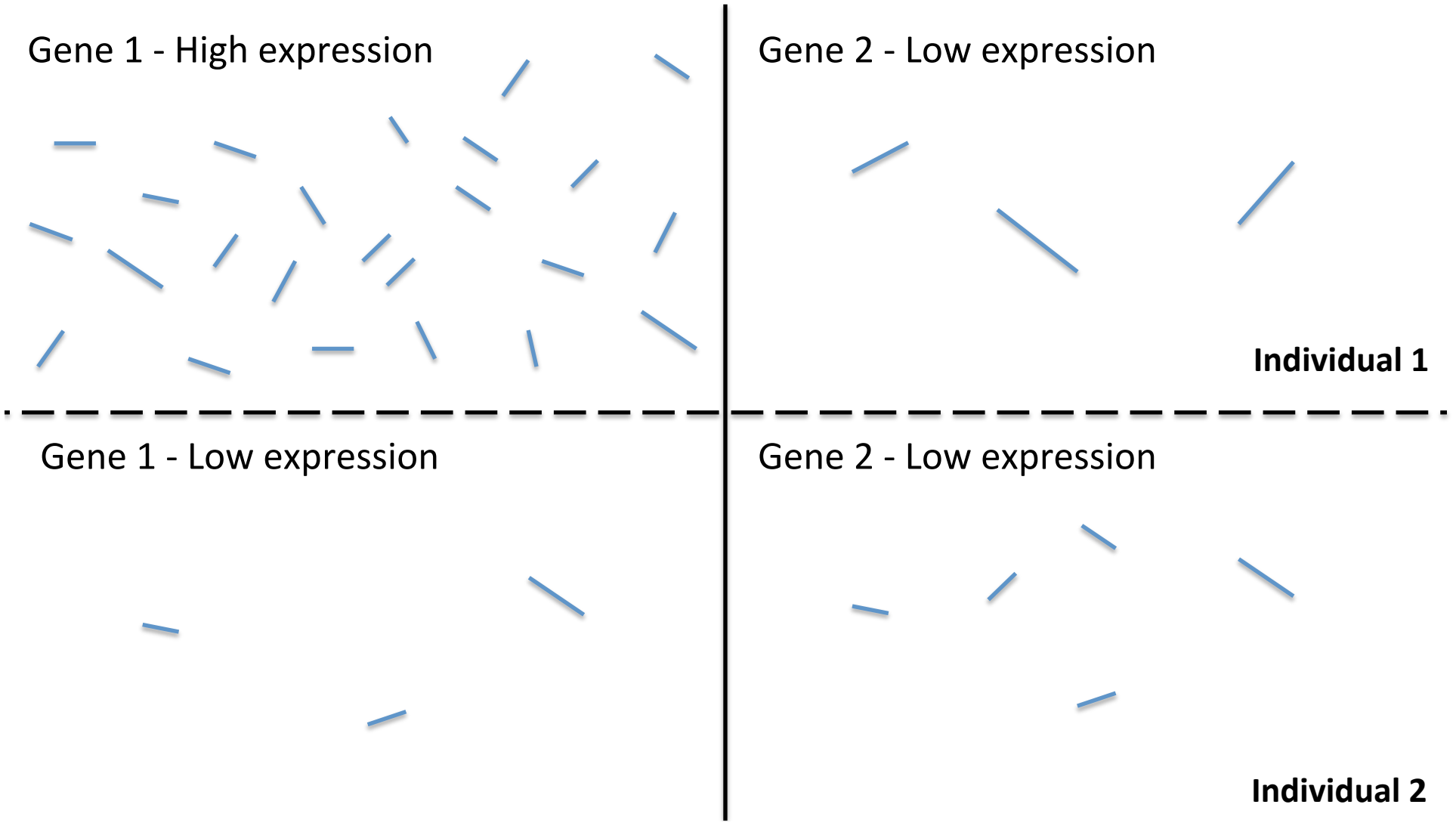
100s of millions of paired reads
10s of billions bases of sequence



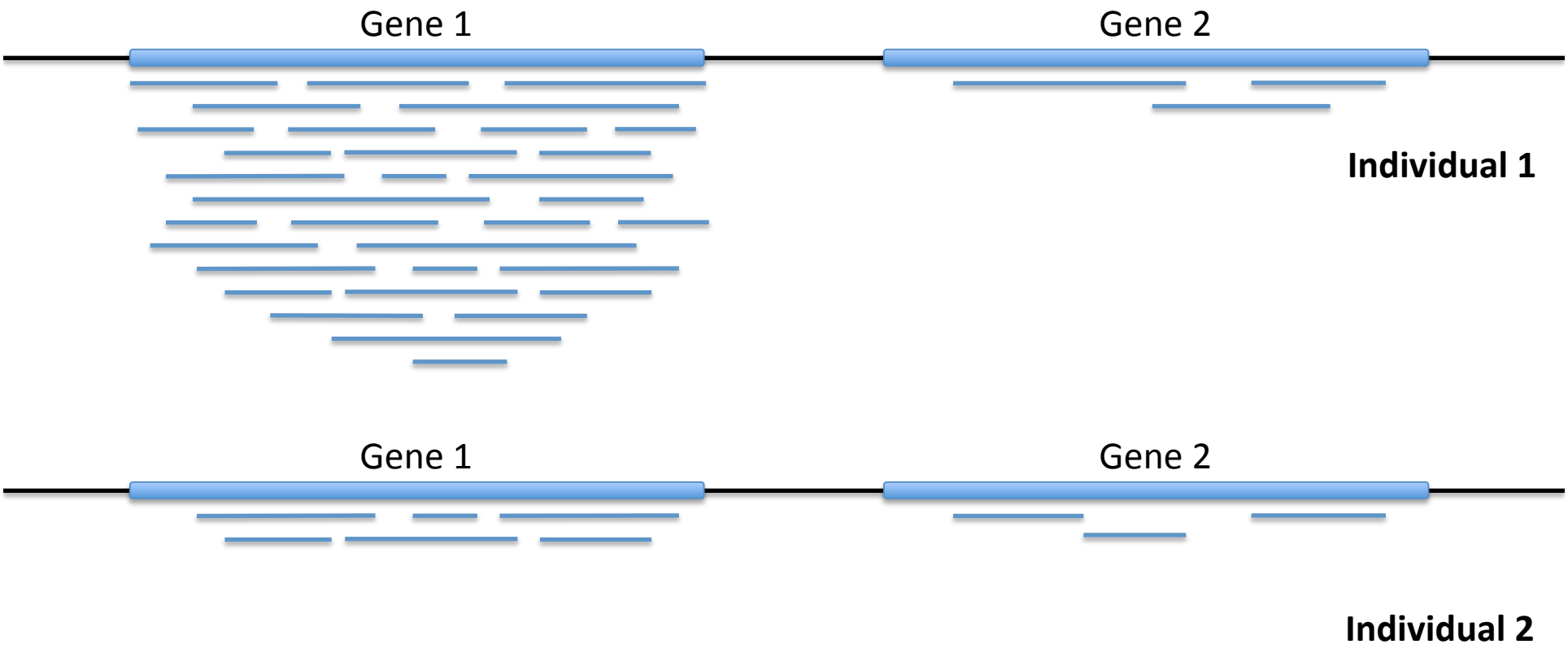
Map to genome OR to de novo assembled transcriptome

Downstream analysis for differential expression and/or SNP calling

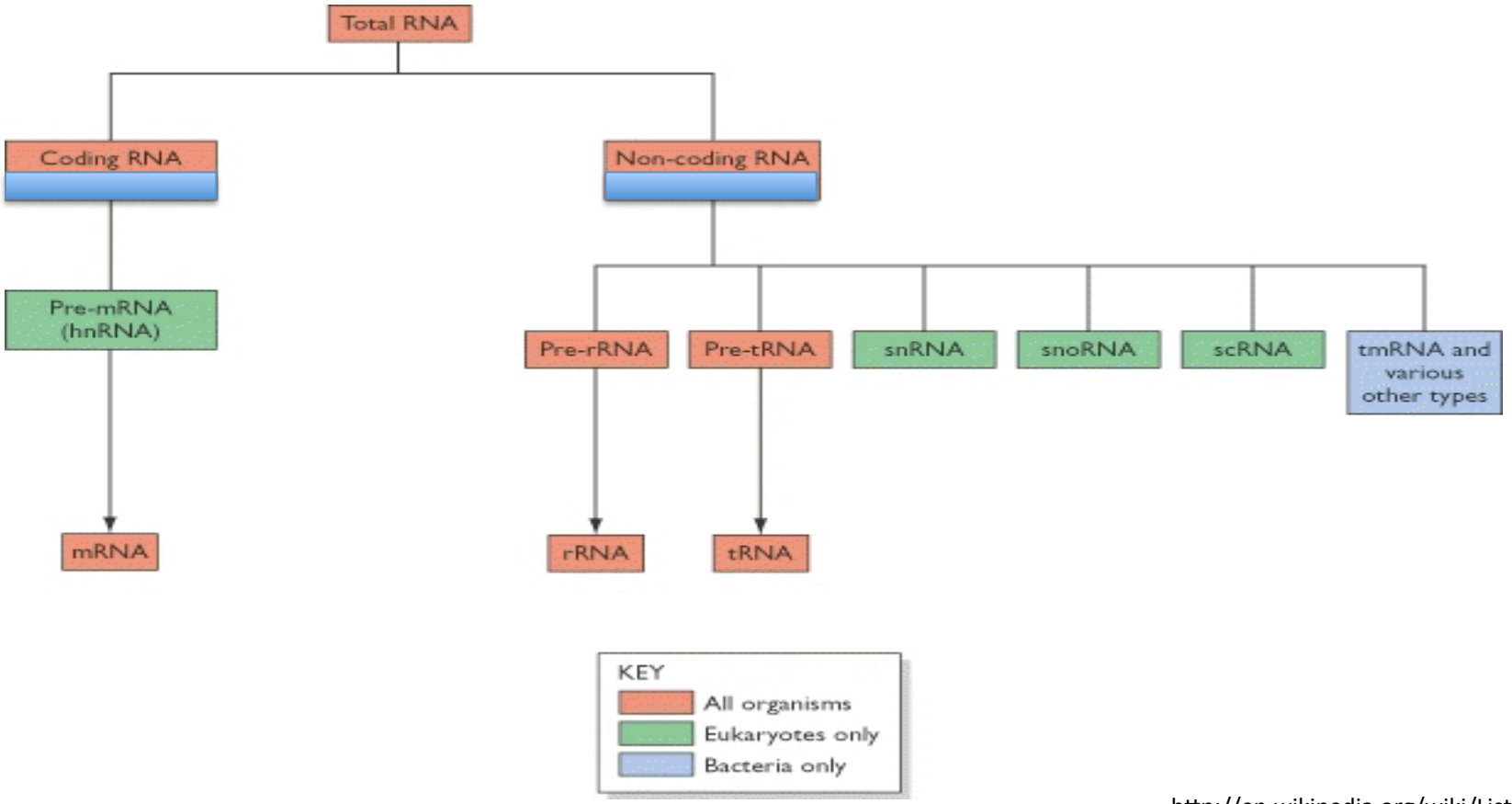

Gene 1 - High expression
Gene 2 - Low expression
Individual 1


Gene 1 - High expression
Gene 2 - Low expression
Individual 1
Gene 1 - Low expression
Gene 2 - Low expression
Individual 2

RNA-Seq reads pile up higher on genes that are highly expressed



Components of total RNA



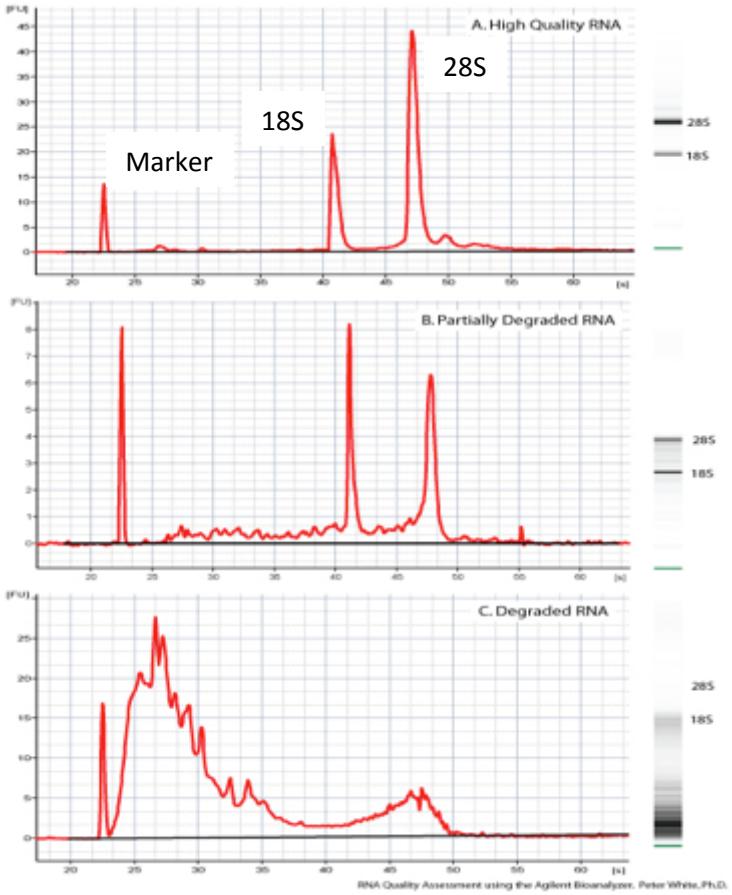
http://en.wikipedia.org/wiki/List_of_RNAs

Extracting total RNA

What would total RNA look like if we ran it on a gel?

- High quality RNA?
- Slightly degraded RNA?
- Degraded RNA?

Degraded total RNA?

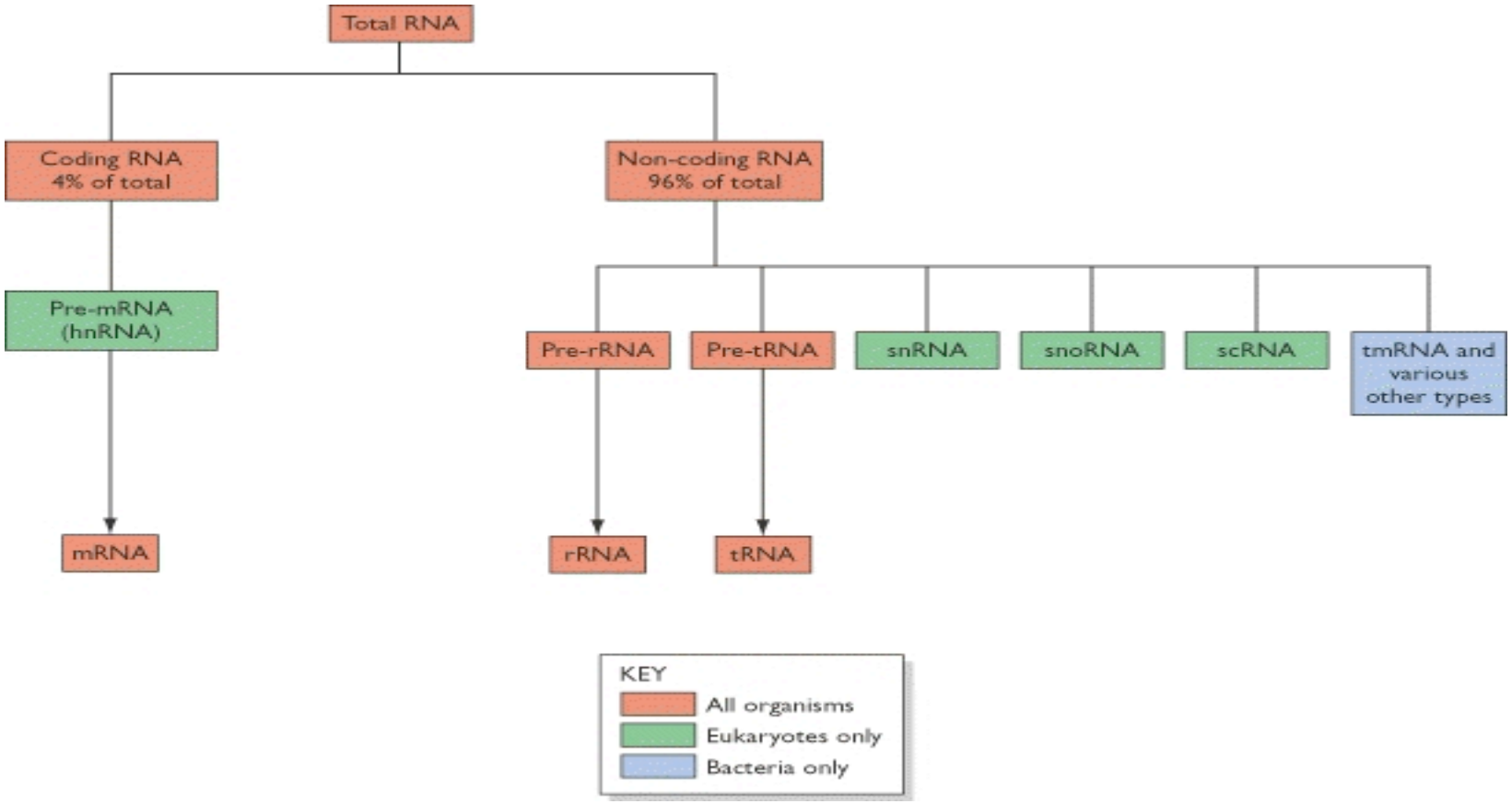


High quality RNA

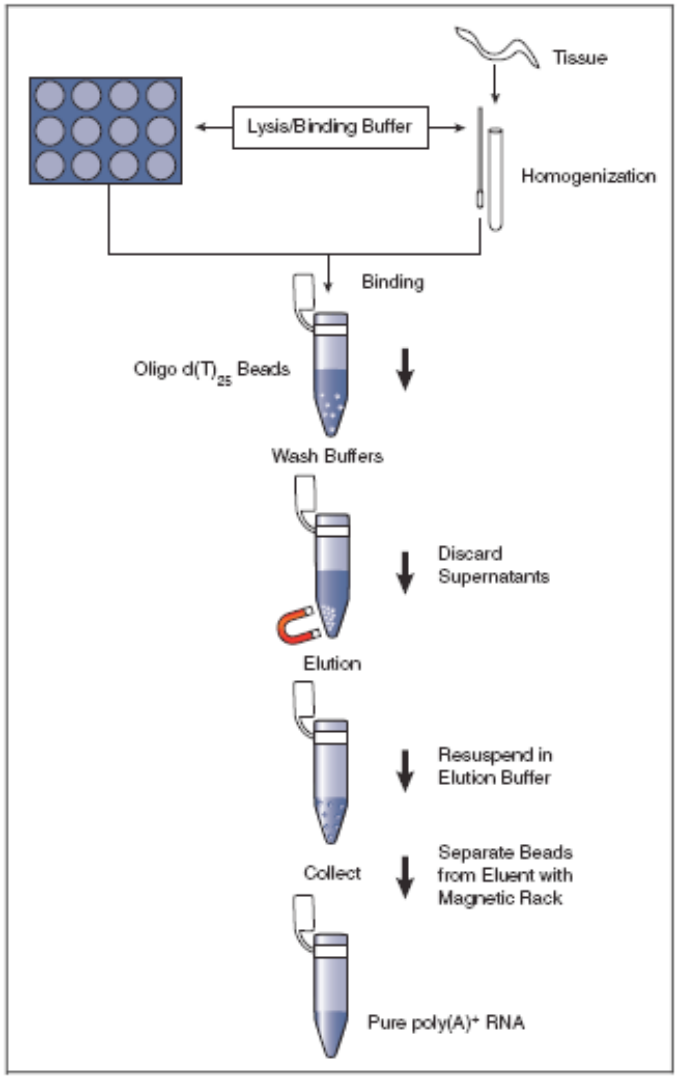
Slightly degraded RNA

Degraded RNA

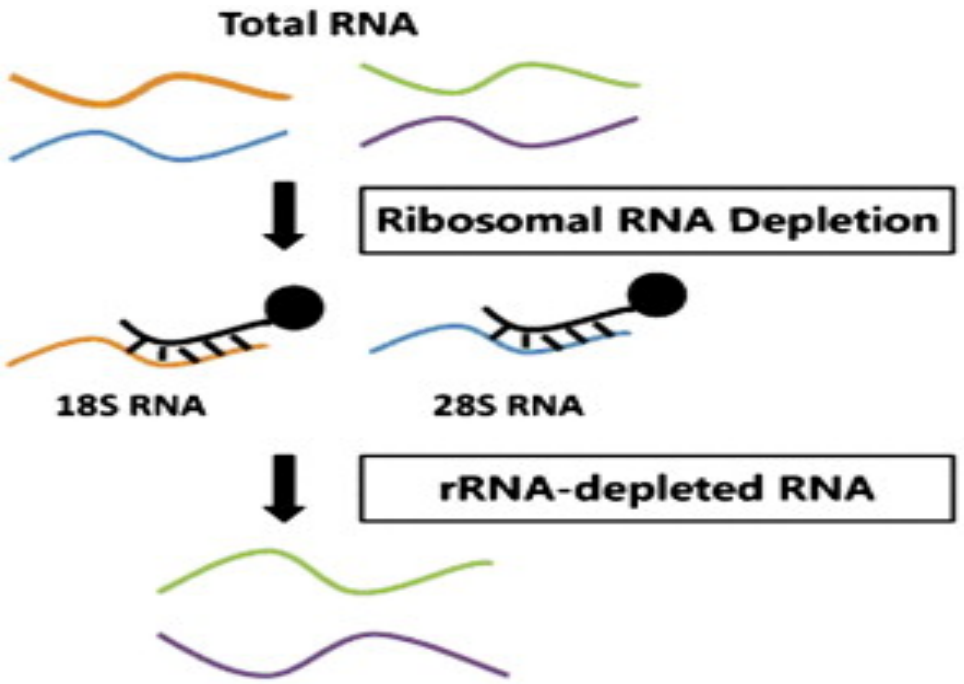
How to enrich for the RNA that we are interested in?



Poly-A+ selection

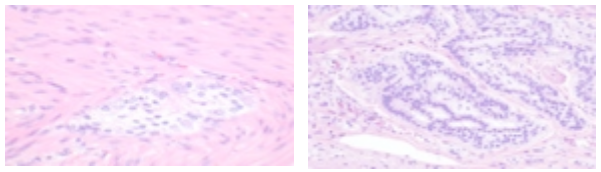


Ribo-minus


Total RNA
Ribosomal RNA Depletion
18S RNA
28S RNA
rRNA-depleted RNA

RNA sequencing experiment

Samples of interest

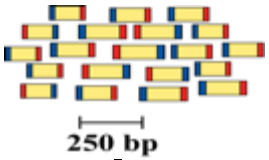


Condition 1 (normal colon)

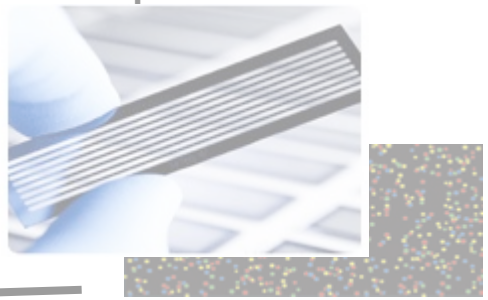
Condition 2 (colon tumor)

Isolate RNAs

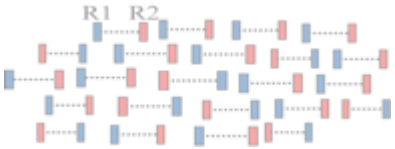
Library preparation



Sequence ends



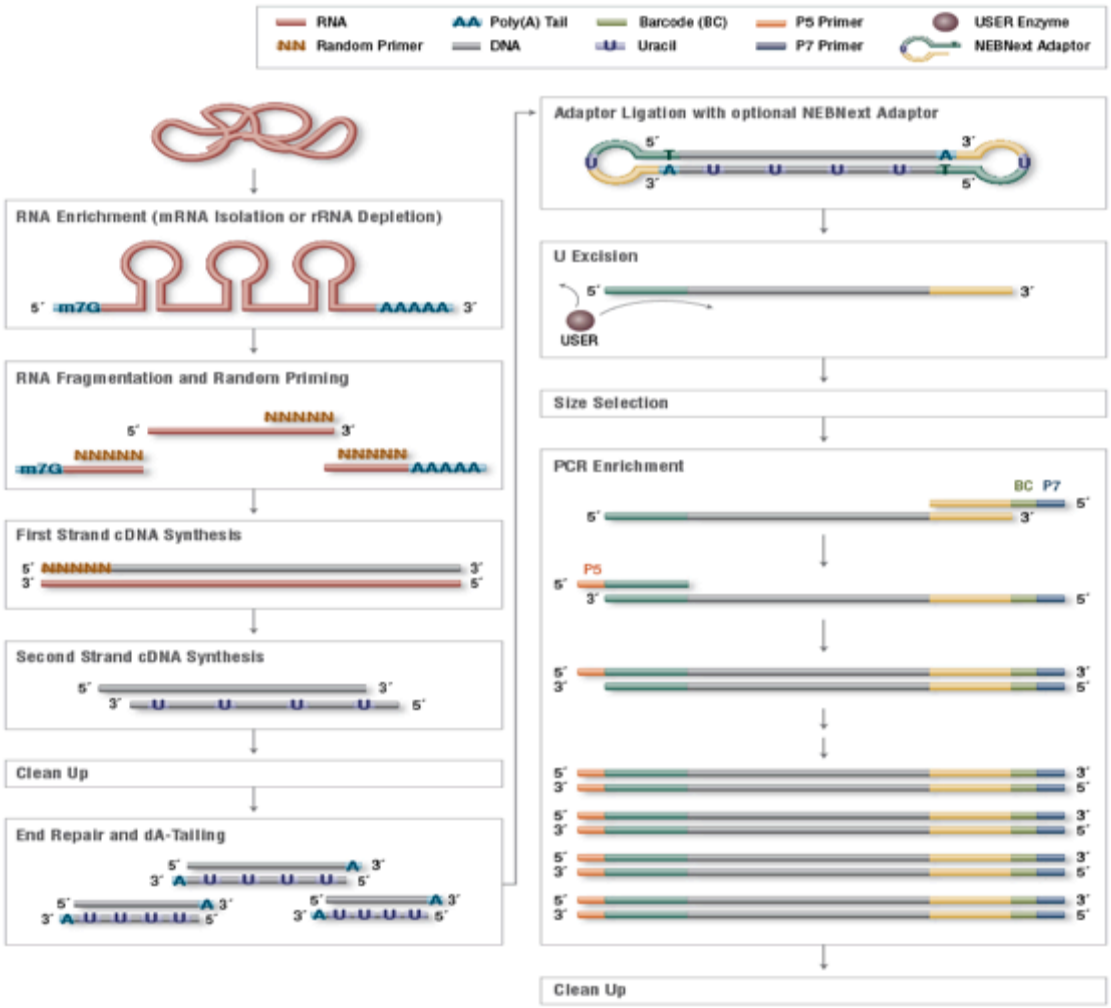
100s of millions of paired reads
10s of billions bases of sequence



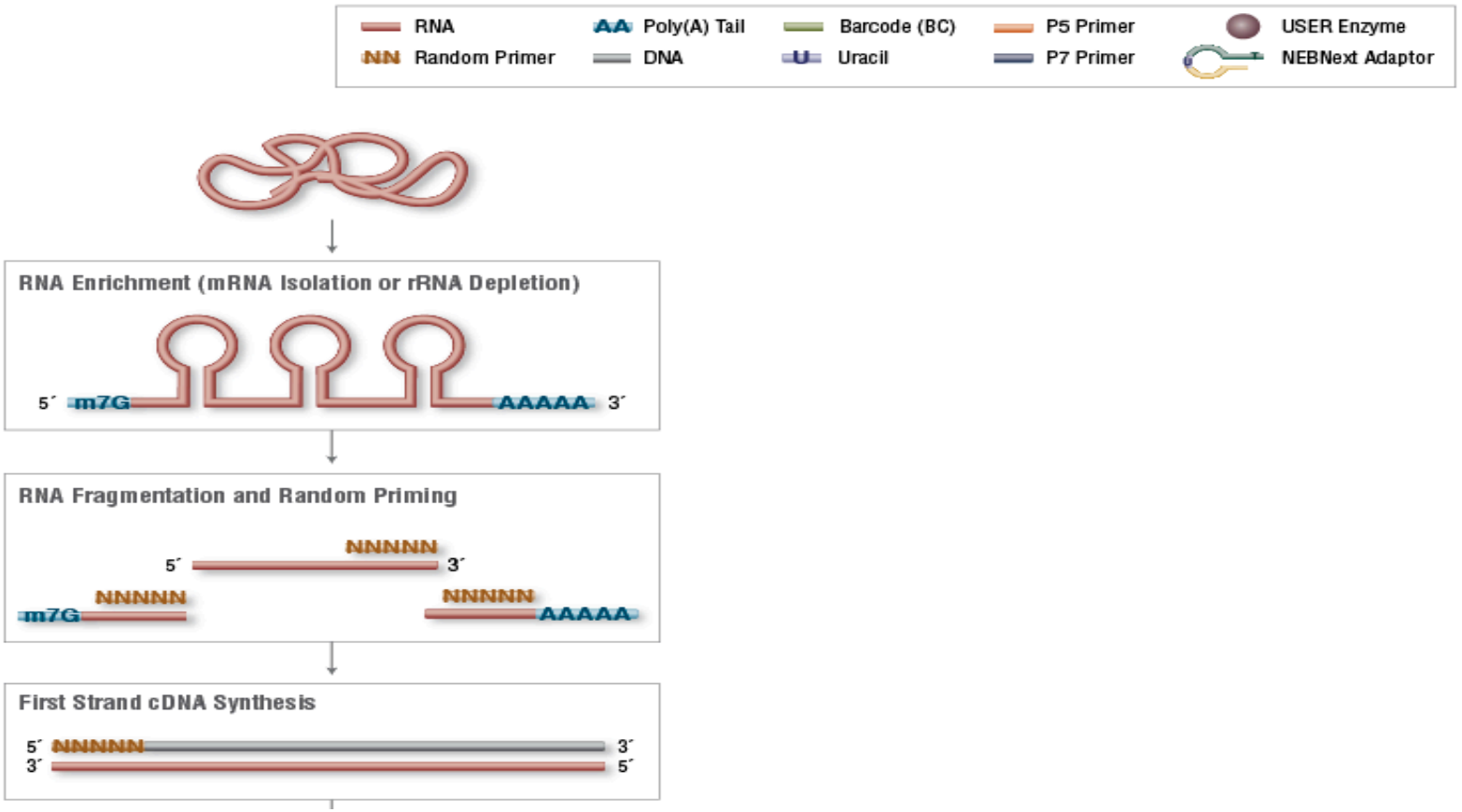
Map to genome OR to de novo assembled transcriptome

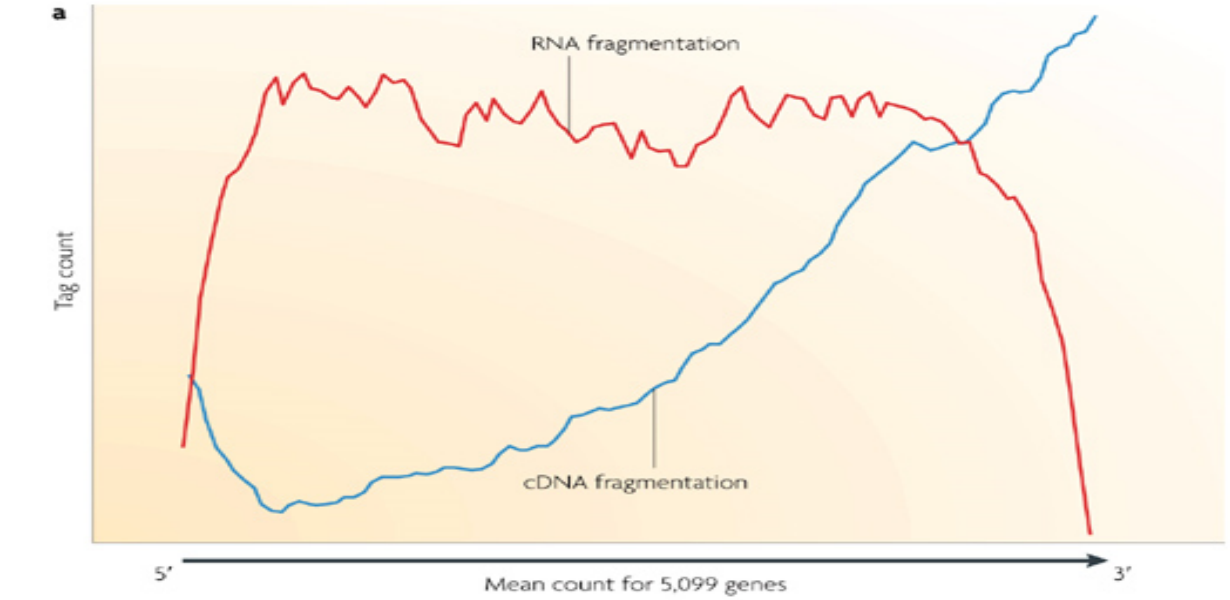
Downstream analysis for differential expression and/or SNP calling

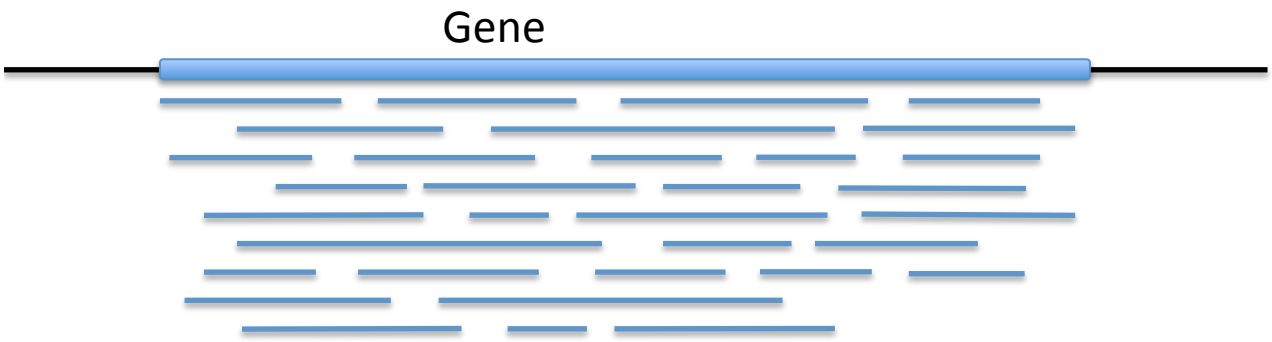
RNA-seq Library preparation process



Steps in the library prep process



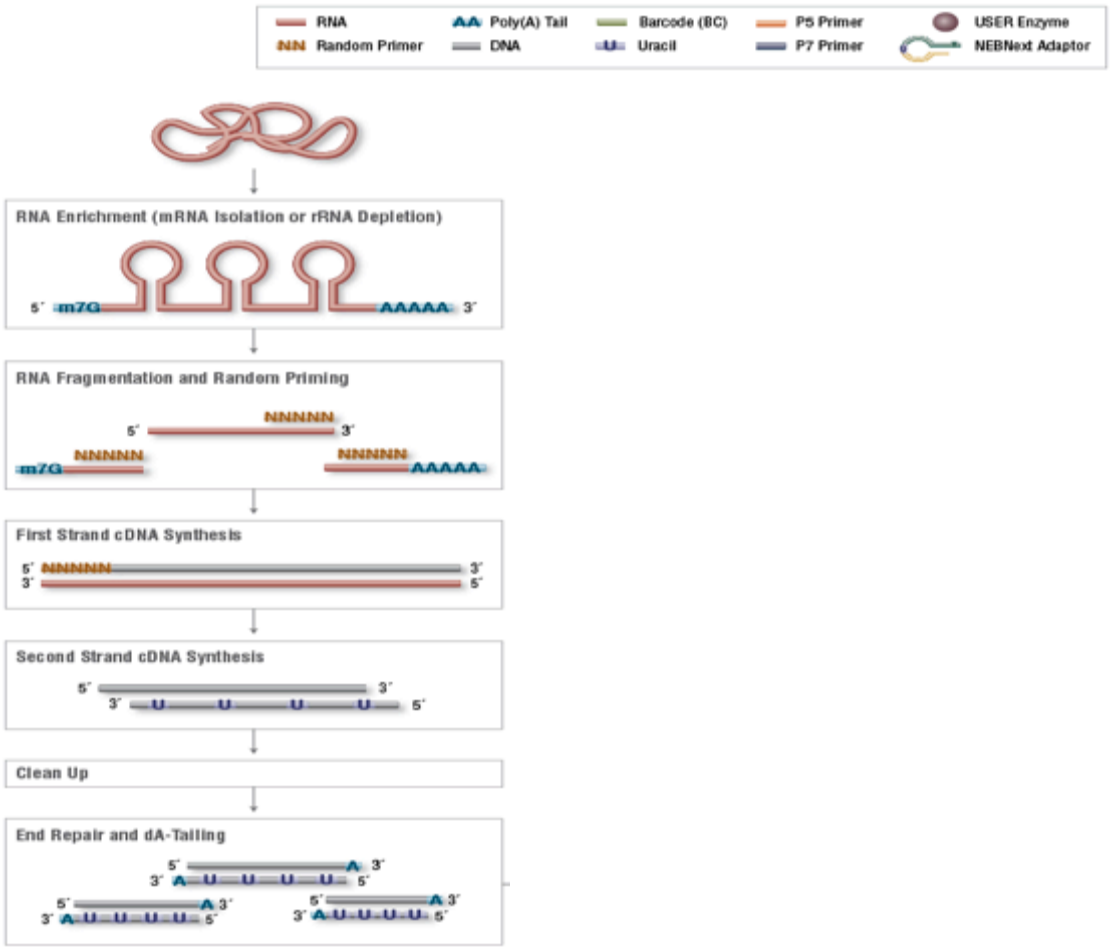




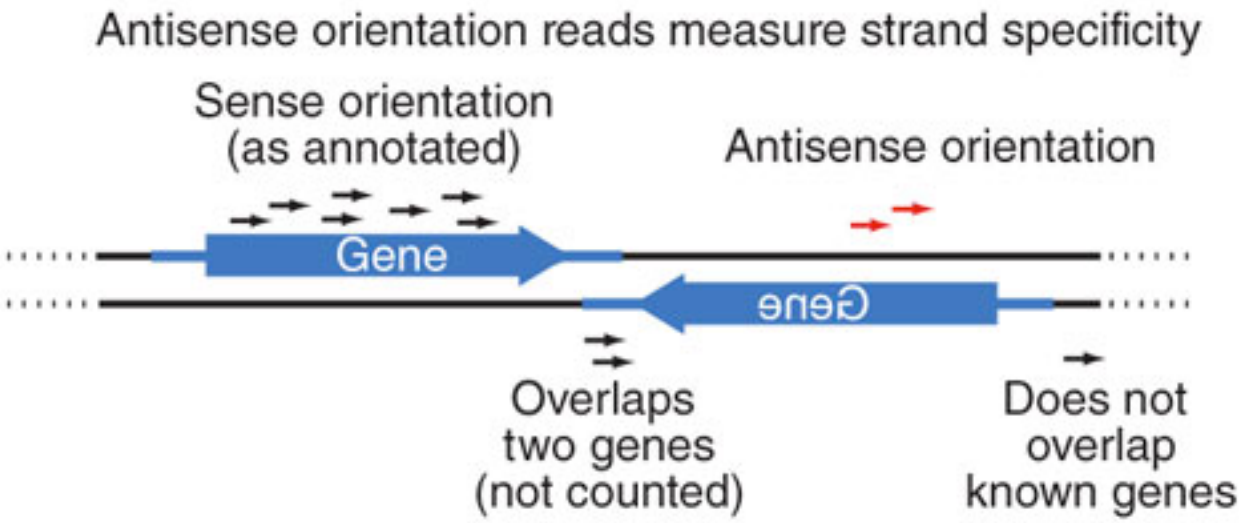
Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends.

RNA-seq Library preparation process

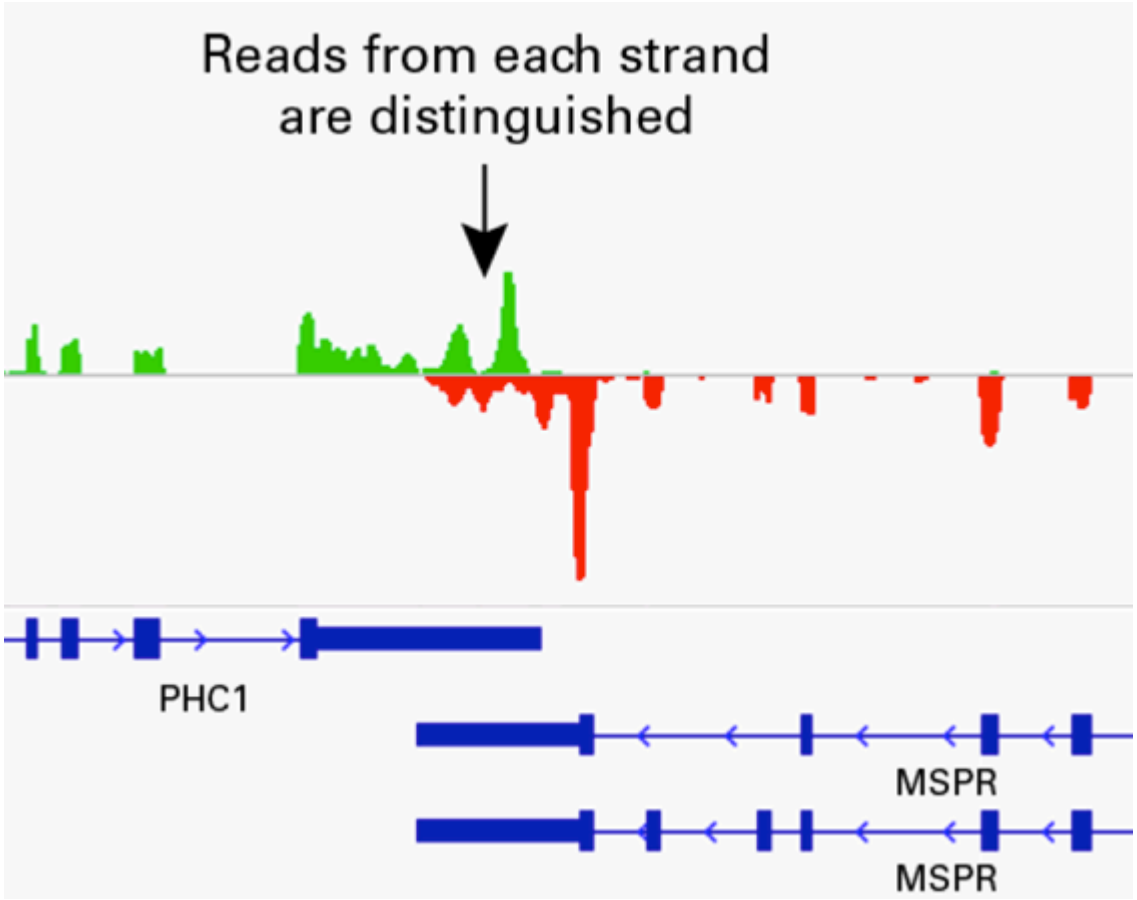
Directional (Stranded) vs non-directional



Possible insights from a directional RNAseq experiment

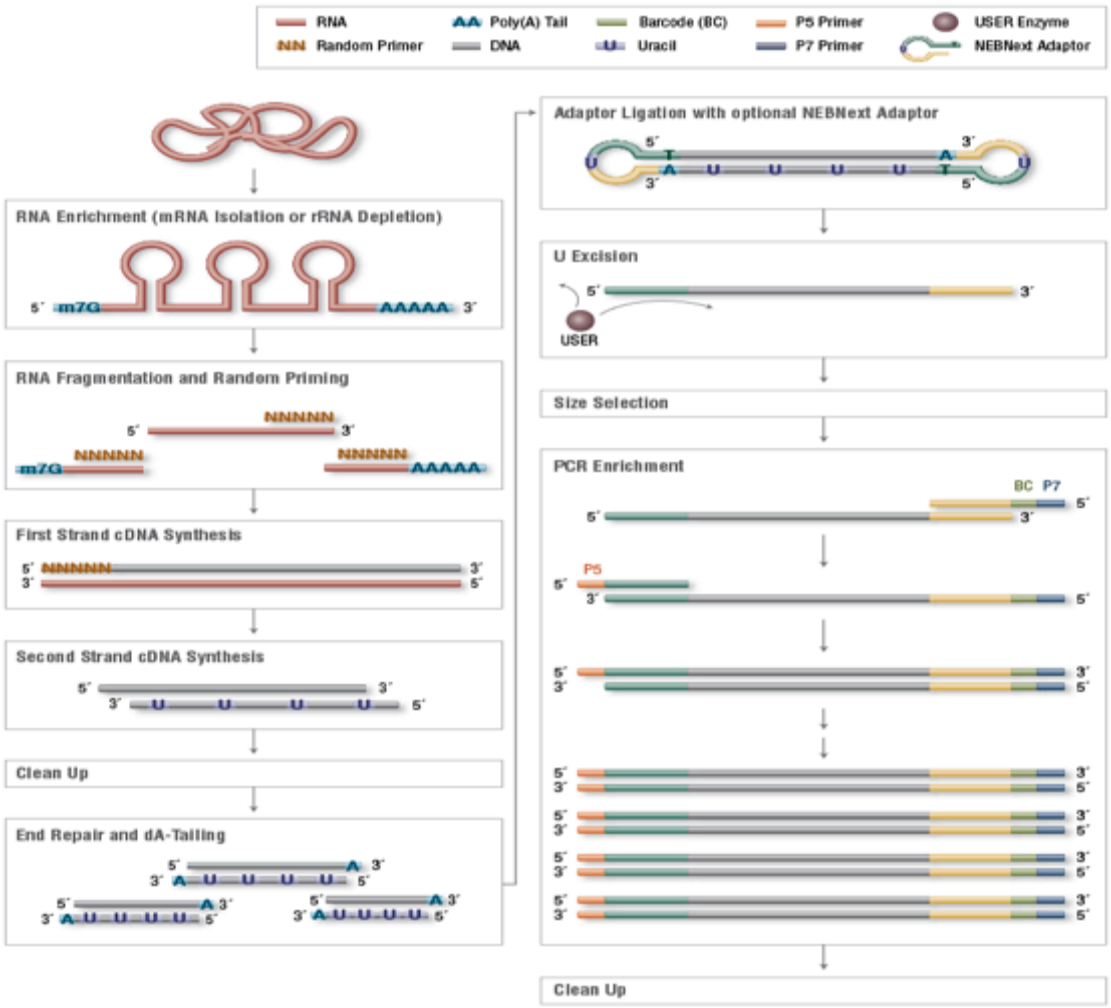


Possible insights from a directional RNAseq experiment



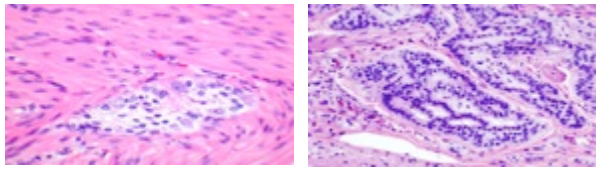
RNA-seq Library preparation process

- Adapter ligation
- Barcoding
- Amplification



RNA sequencing experiment

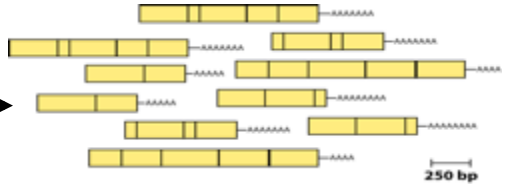
Samples of interest



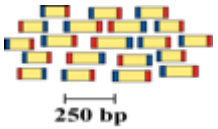
Condition 1 (normal colon)

Condition 2 (colon tumor)

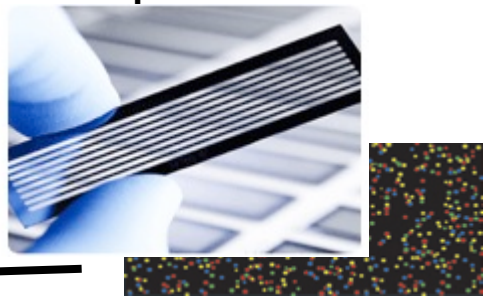
Isolate RNAs



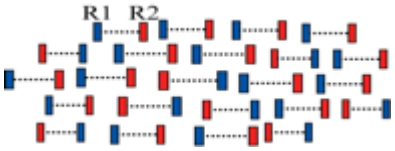
Library preparation



Sequence ends



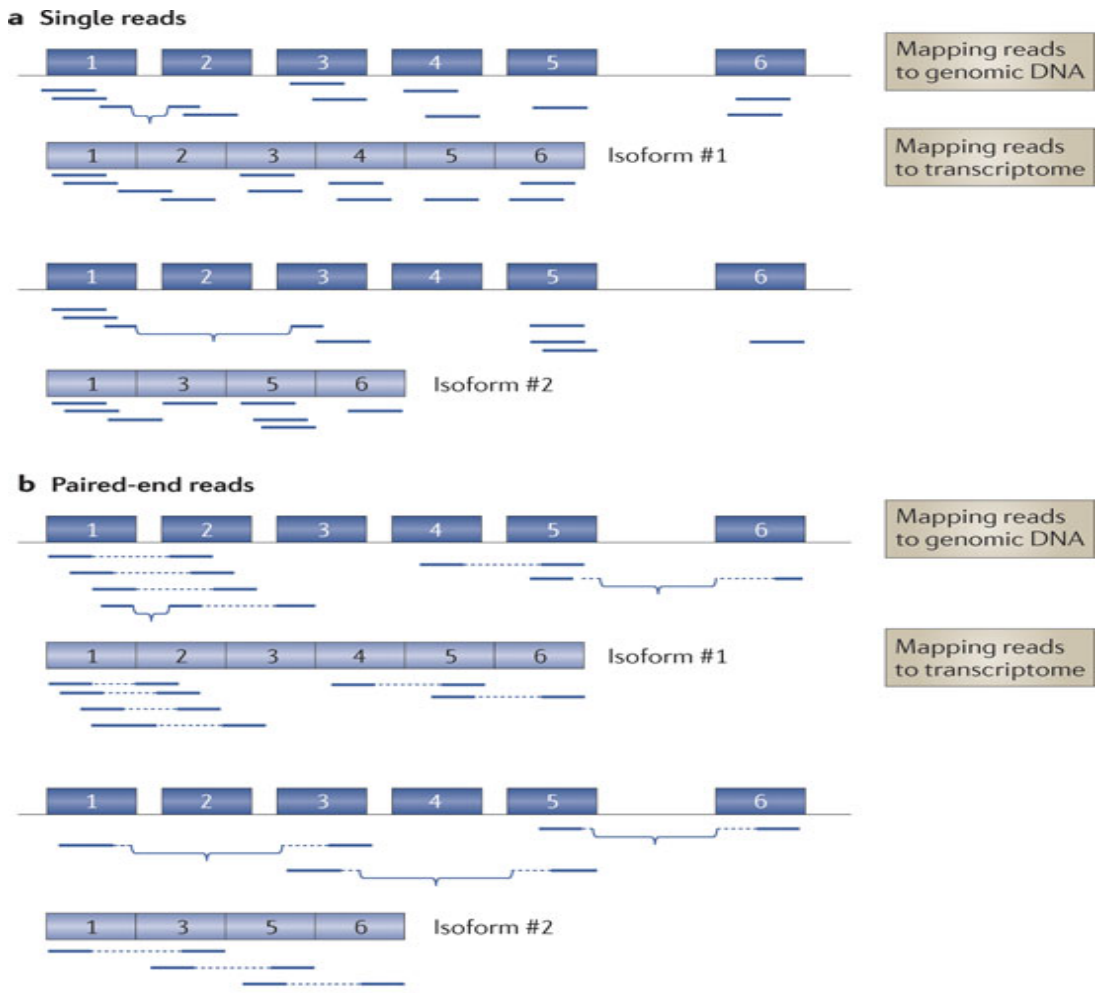
100s of millions of paired reads
10s of billions bases of sequence



Map to genome, transcriptome, and predicted exon junctions

Downstream analysis for differential expression, SNP calling, etc

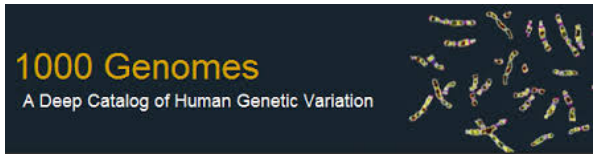
Reference genome



Nature Reviews | Genetics

The closest genome? Is it close "enough"?

http://arthropodgenomes.org/wiki/i5K



http://www.1000genomes.org/

http://genome10k.soe.ucsc.edu/

All depends on your question:

- Could be close enough – but might lose important information
- Advantages – annotations already done

No reference? No problem!

- De novo reference transcriptome assembly!
- Trinity de novo
- Other programs

Divergent selection by a natural toxicant

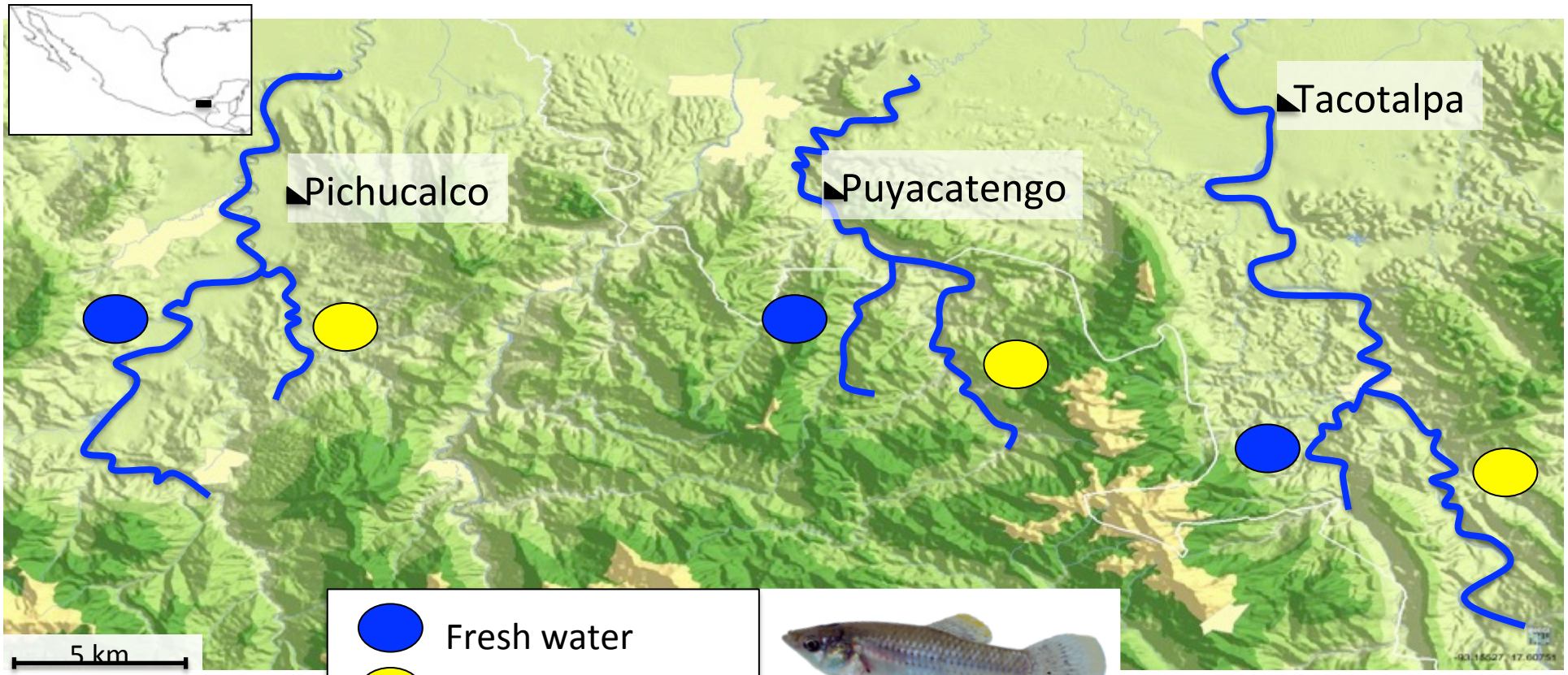
Fish living in streams with high levels of hydrogen sulfide (H_2S)

H_2S creates physiologically explicit environmental gradients:

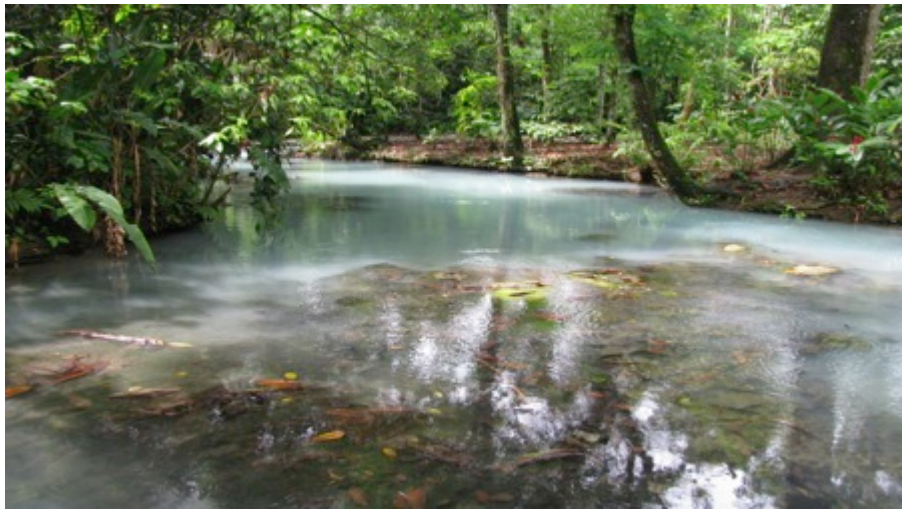
- Naturally occurring in volcanic regions
- Inhibition of oxygen transport and cellular respiration
- Acute toxicity in micromolar concentrations
- Concentrations up to 1100 µM present (~19 ppm)
- Causes and aggravates hypoxia

Tobler et al. 2006, *Extremophiles*

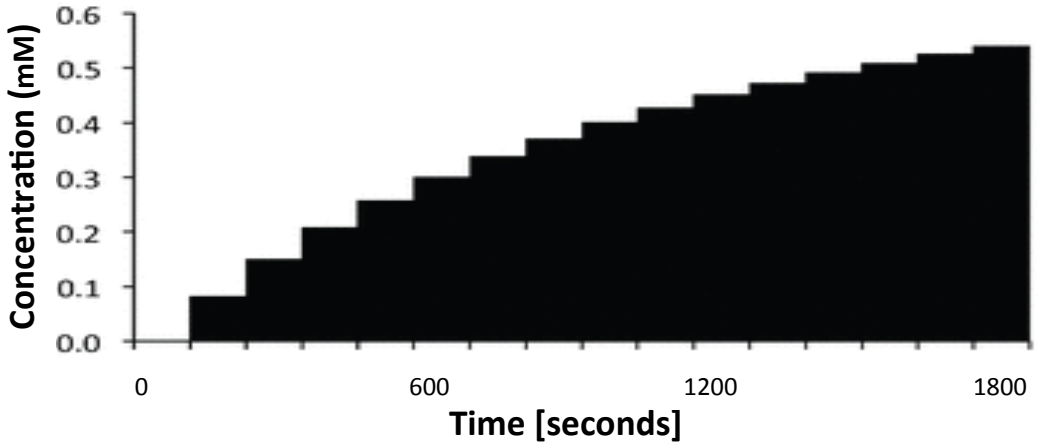
Three distinct sulfidic watersheds

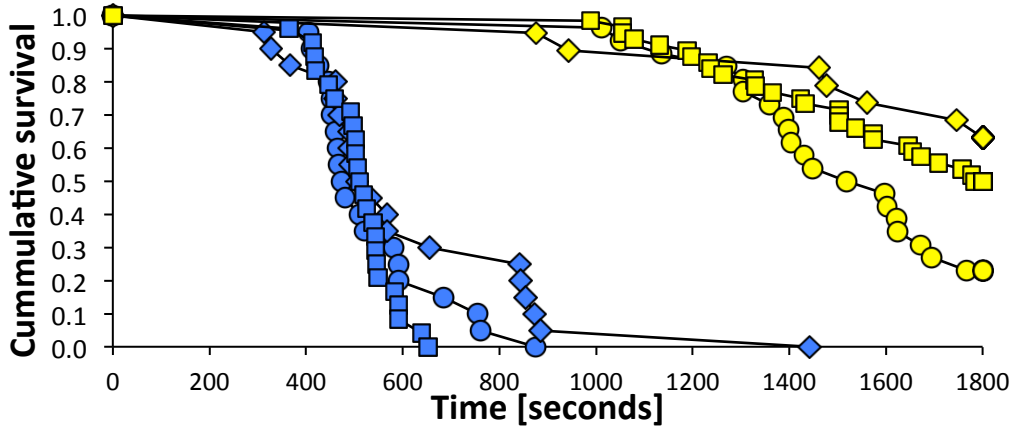


Poecilia mexicana



Physiological sulfide tolerance





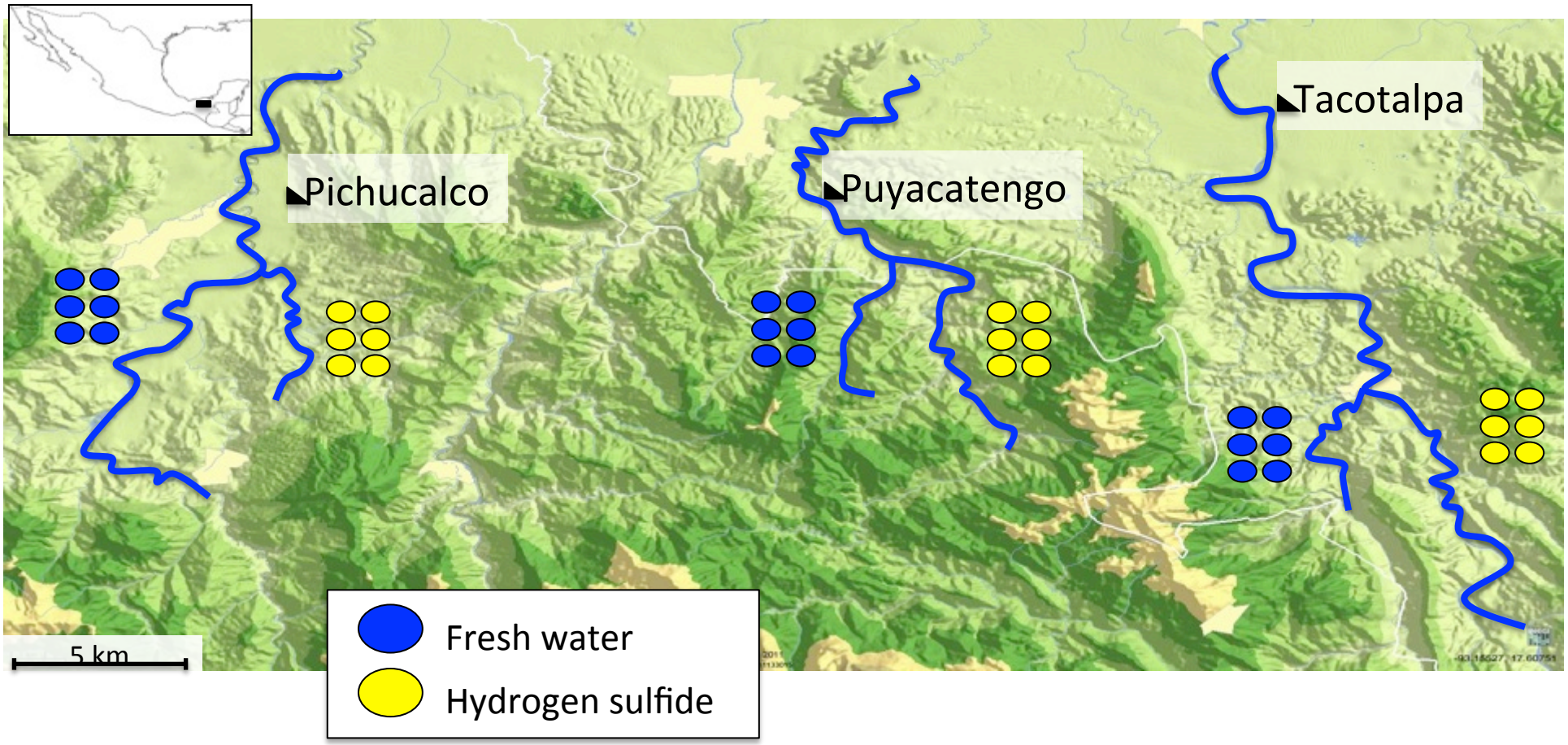
- Tacotalpa, non-sulfidic
- Tacotalpa, sulfidic
- Puyacatengo, non-sulfidic
- Puyacatengo, sulfidic
- Pichucalco, non-sulfidic

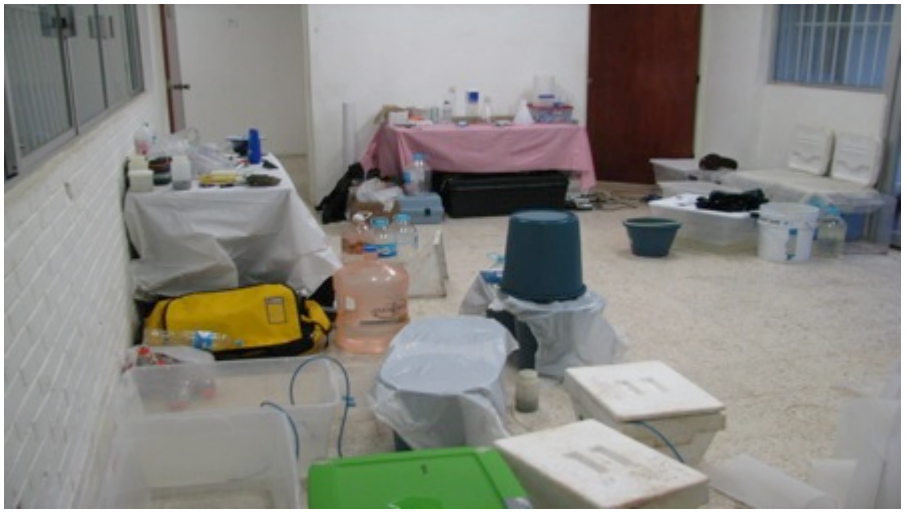
Tobler et al. 2011, *Evolution*

Importance of respiratory adaptations

- H_2S leads to and aggravates hypoxia
- Oxygen is needed for sulfide detoxification
- Aquatic surface respiration (ASR)

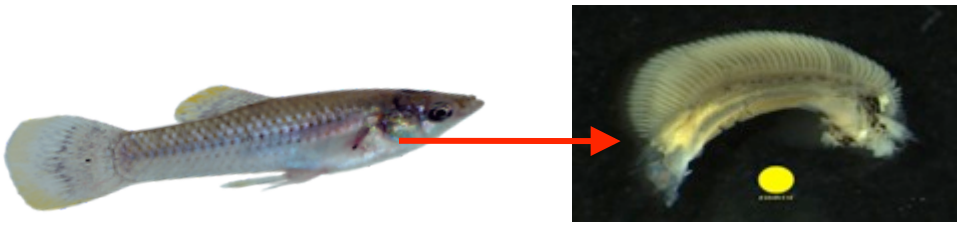
Sampling individuals from parallel environments



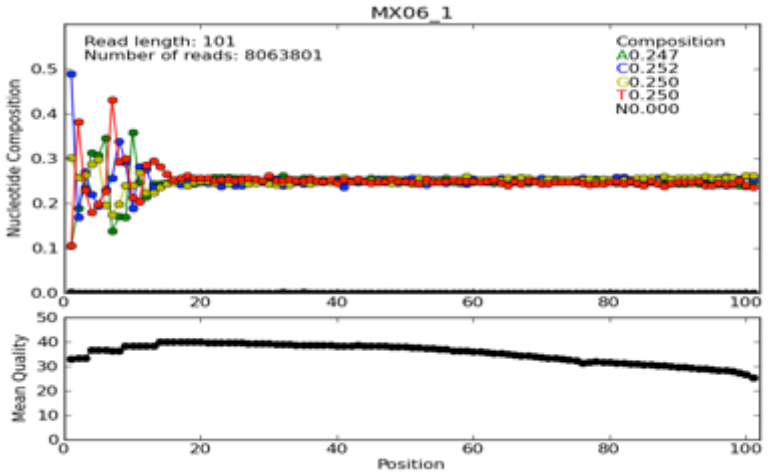


Transcriptome analyses

RNAsequencing approach:



- RNA extraction (polyA+)
- cDNA library construction
- Individual barcoding
- Paired-end cDNA sequencing on Illu
- Process reads
  - Sequencing adaptor (~0.02% of reads)
  - Composition skew and low quality
  - Poly A/Ts

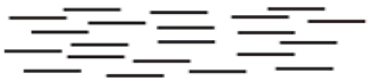


Transcriptome analyses

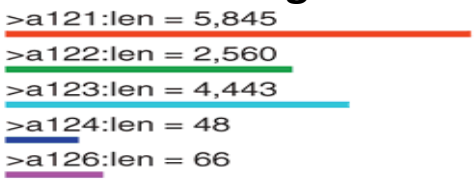
De novo transcript assembly:

- Use Trinity (Grabherr et al. 2011, *Nat Biotechnology*)
- Combine data from 36 individuals
- Normalize the data *in silico:* include up to 20x of a given kmer

Normalized RNA-seq reads



Linear contigs



de Bruijn graphs

Transcripts and isoforms

```
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TGATCGGAT...
```

- Validate 6 of the predicted transcripts using RT-PCR and sequencing

Poecilia mexicana transcriptome

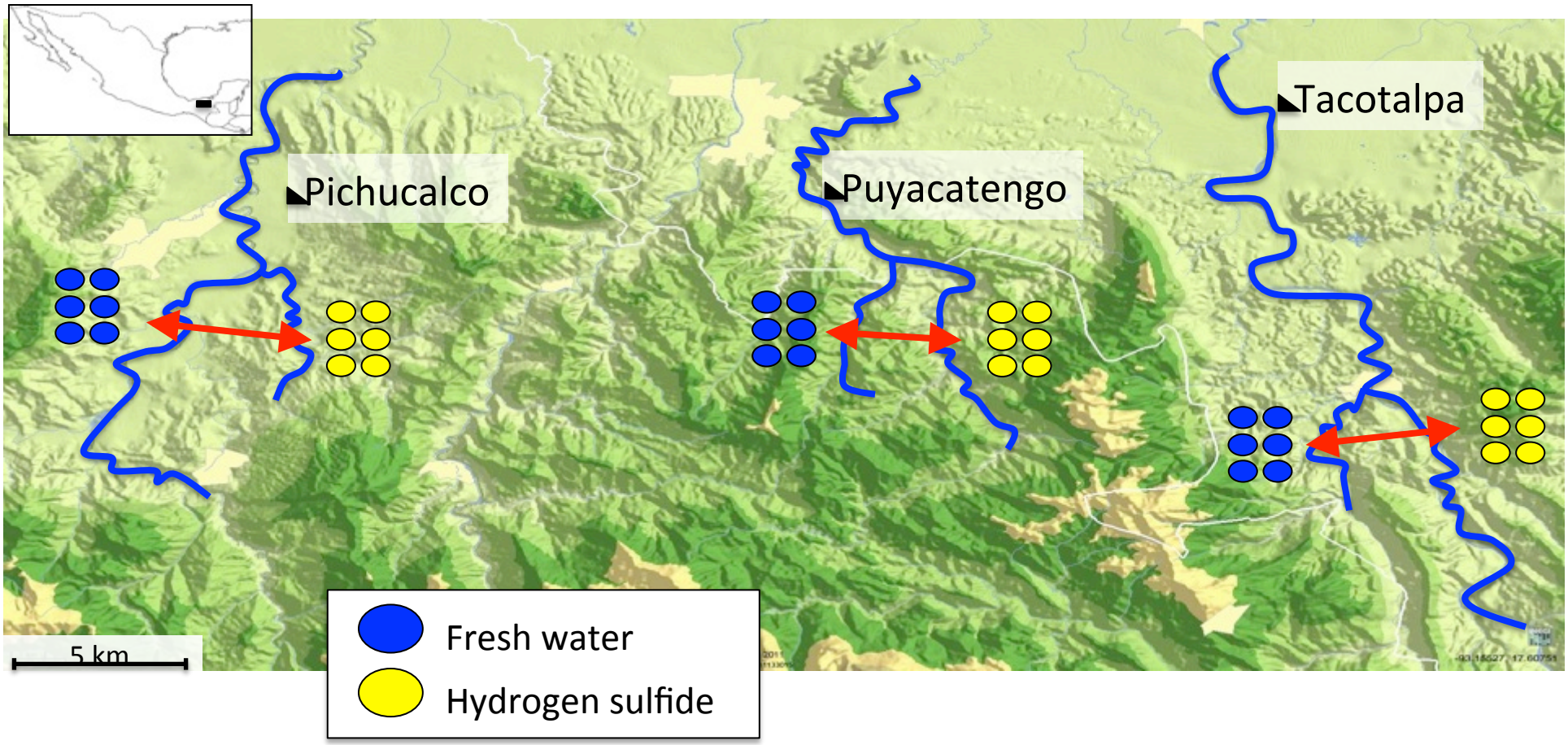
	Sample	# reads	# base pairs
Pichucalco	MX05	4,986,074	1,007,186,948
	MX06	8,063,801	1,628,887,802
	MX07	5,220,840	1,054,609,680
Tacotalpa	MX61	6,618,861	1,337,009,922
	MX62	6,150,007	1,242,301,414
	MX63	4,916,184	993,069,168
	Average	**5,992,628**	**1,210,510,822**

Assembly statistics:

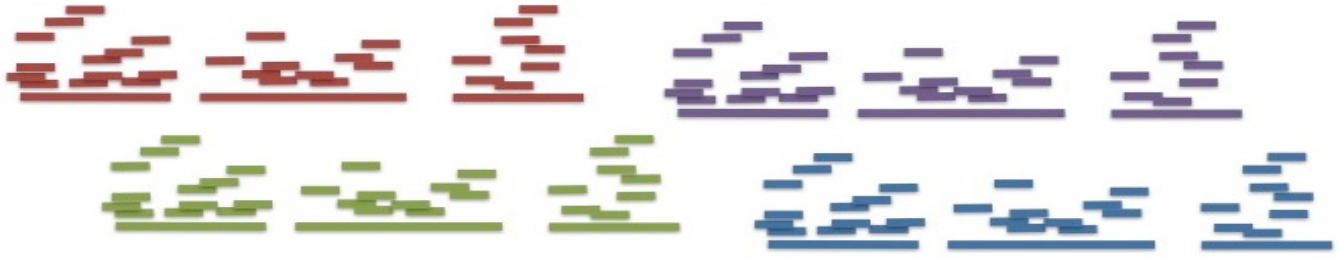
- Mean transcript length: 808 bp
- Median transcript length: 429 bp
- Maximum transcript length: 15,812 bp
- 78,193 transcripts (including different isoforms)
  - 63,004 unique predicted loci (includes isoforms with >97% sequence identity)
  - 62,558 loci with predicted open reading frames

Kelley et al 2012, *BMC Genomics*

Identifying patterns of differential gene expression

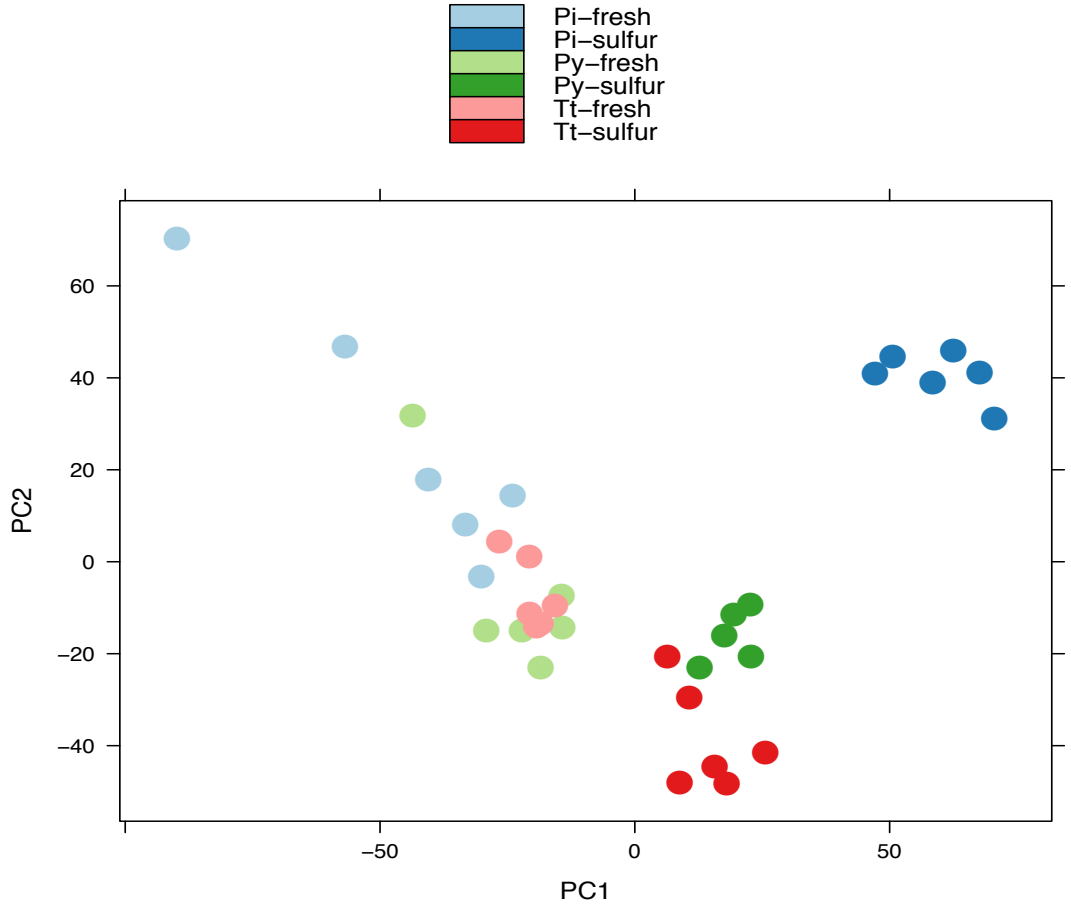


Identifying patterns of differential gene expression

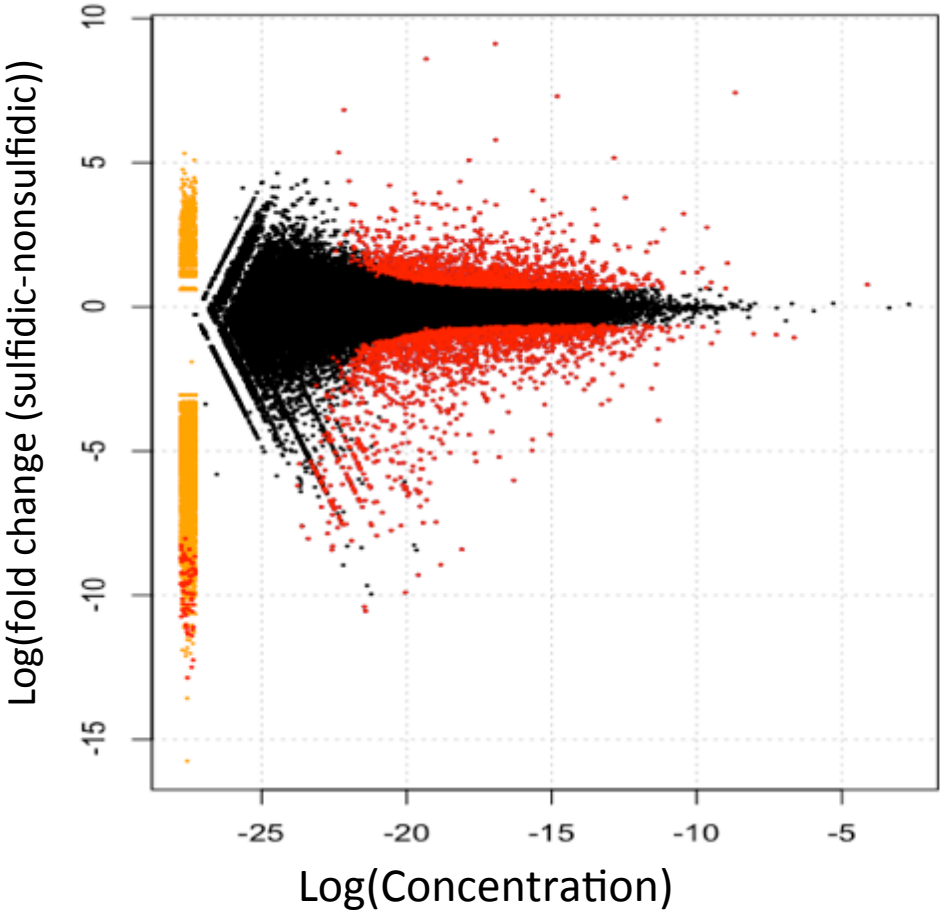


- **Map reads to reference transcriptome to estimate expression levels for each individual**
- **Test for differential expression using an exact test (or GLM) on normalized expression levels**

PCA of top variable genes separates sulfur from non-sulfur ecotypes

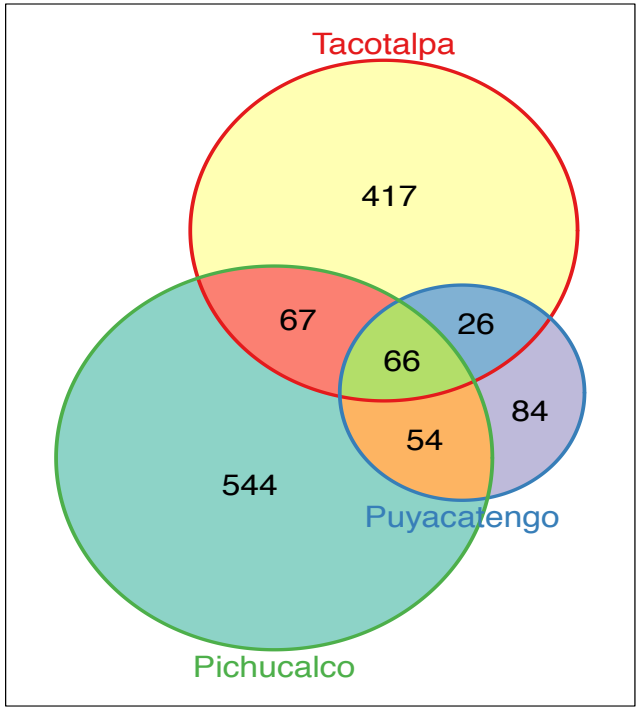


Identifying patterns of differential gene expression

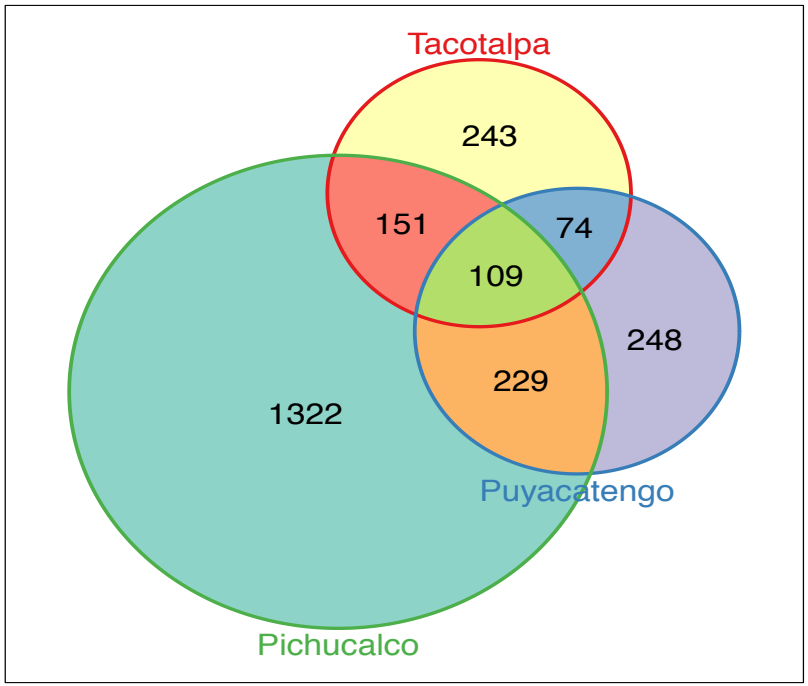


- ~2000 differentially expressed genes between sulfidic and non-sulfidic ecotypes
- ~10% of differently expressed genes are shared among sulfidic and non-sulfidic ecotypes from different drainages

Extent of differential expression is consistent with inferred population divergences

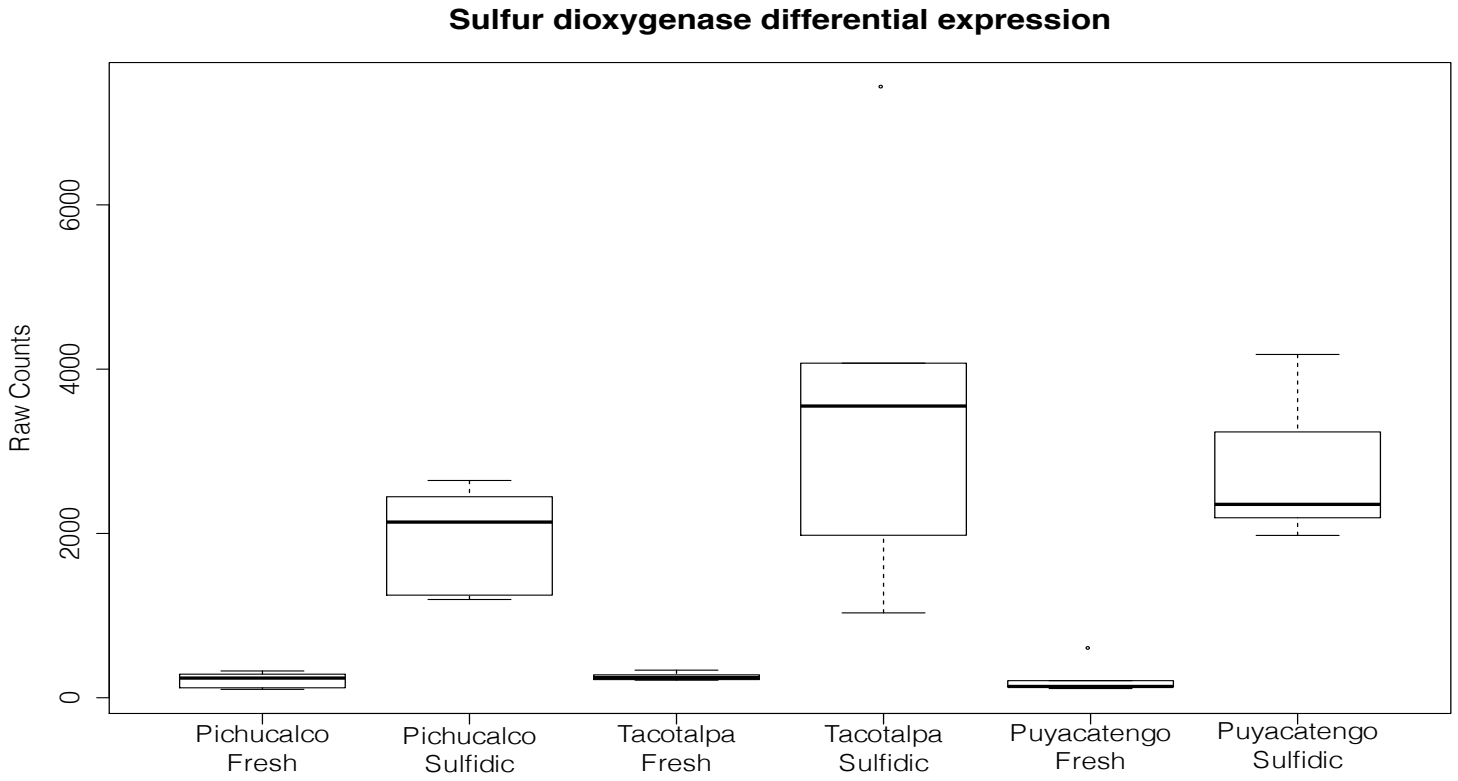


Genes that are up-regulated in sulfidic environment(s)



Genes that are down-regulated in sulfidic environment(s)

Highly expressed in sulfidic environments



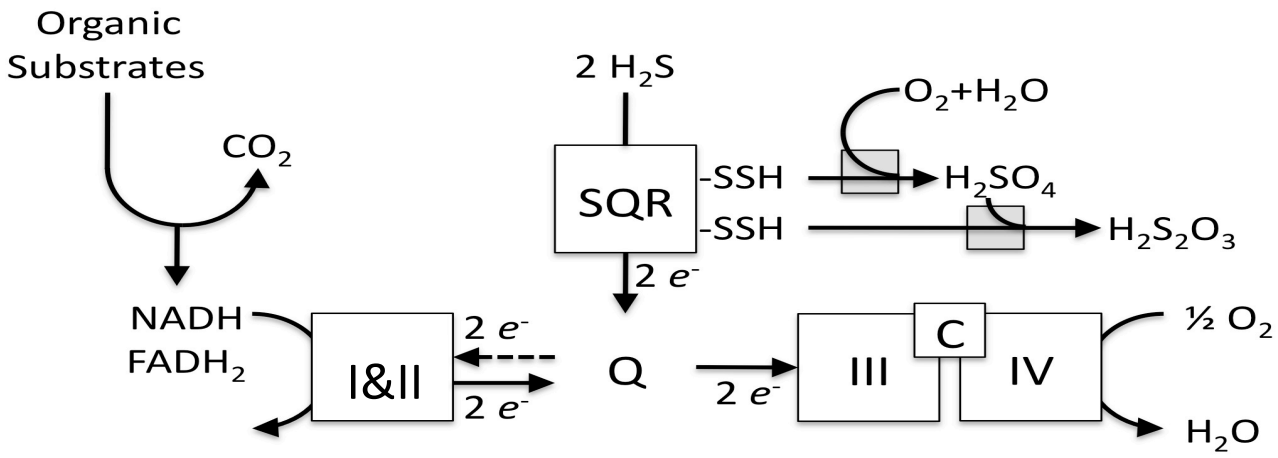
Highly expressed in sulfidic environments

Annotation using BLAST (or BLAST2GO or Trinotate)

Contig Name	Annotation similarity	Annotation
comp10393_c0_seq1	78%	NADH dehydrogenase 1 alpha subcomplex subunit 4-like 2
comp11093_c0_seq1	87%	Mitochondrial uncoupling protein 2
comp11162_c0_seq1	64%	Ferric-chelate reductase 1
comp13810_c0_seq1	70%	Retinol dehydrogenase 2
comp14313_c0_seq1	54%	Peroxisome proliferator-activated receptor gamma coactivator 1-beta
comp14320_c0_seq1	87%	Phosphoenolpyruvate cytosolic
comp14332_c0_seq1	71%	Nocturnin
comp22827_c0_seq1	91%	Cytochrome c
comp390_c1_seq1	91%	Myoglobin
comp1681_c0_seq1	78%	Sulfur dioxygenase
…	…	…

Underlying physiology and genes: Sulfide toxicity and detoxification

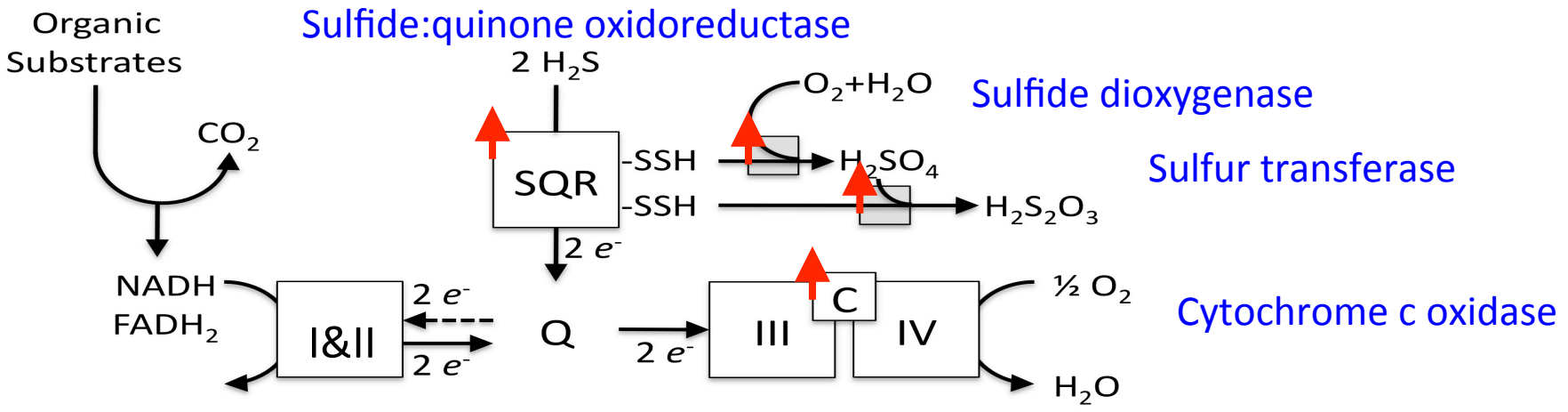
Sulfide toxicity and detoxification in mitochondria



Lagoutte et al. 2010, *Biochim Biophys Acta*

Sulfide toxicity and detoxification

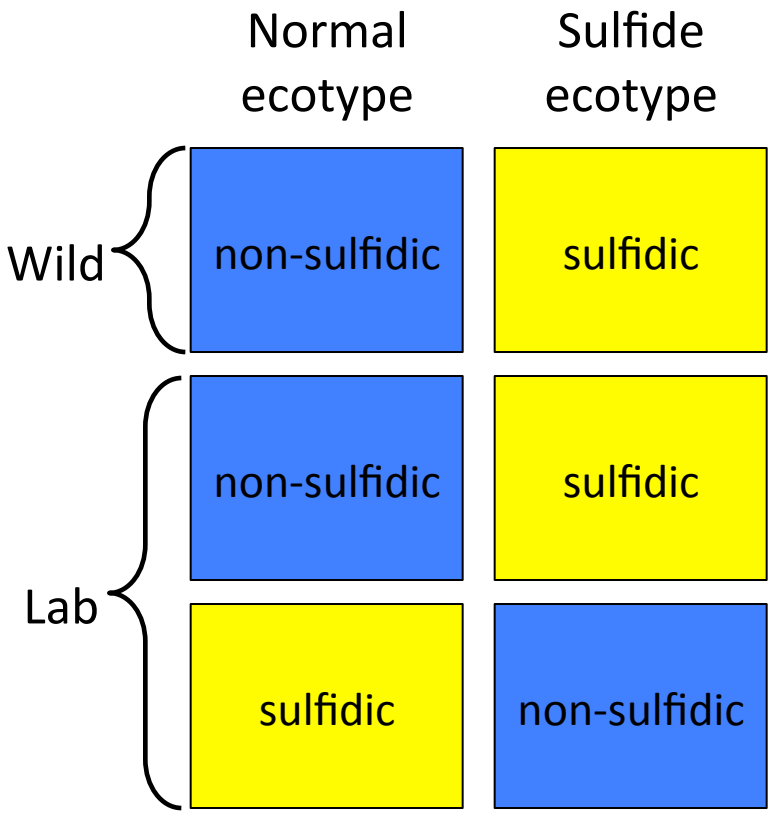
Sulfide toxicity and detoxification in mitochondria



↑ Differentially expressed genes

Lagoutte et al. 2010, *Biochim Biophys Acta*

Heritability of gene expression



- Common garden rearing experiment
- Exposure experiment of lab-raised fish to sulfidic and non-sulfidic conditions
- What differences in gene expression between ecotypes are fixed/ facultative?
- Are there fixed differences in coding sequences between ecotypes?

SNPs from RNAseq data

MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 2310–2323 doi: 10.1111/mec.13165

INVITED REVIEWS AND SYNTHESES

SNP genotyping and population genomics from expressed sequences – current advances and future possibilities

PIERRE DE WIT,* MELISSA H. PESPENI† and STEPHEN R. PALUMBI‡

Am J Hum Genet. 2013 Oct 3; 93(4): 641–651. PMCID: PMC3791257
doi: 10.1016/j.ajhg.2013.08.008

Reliable Identification of Genomic Variants from RNA-Seq Data

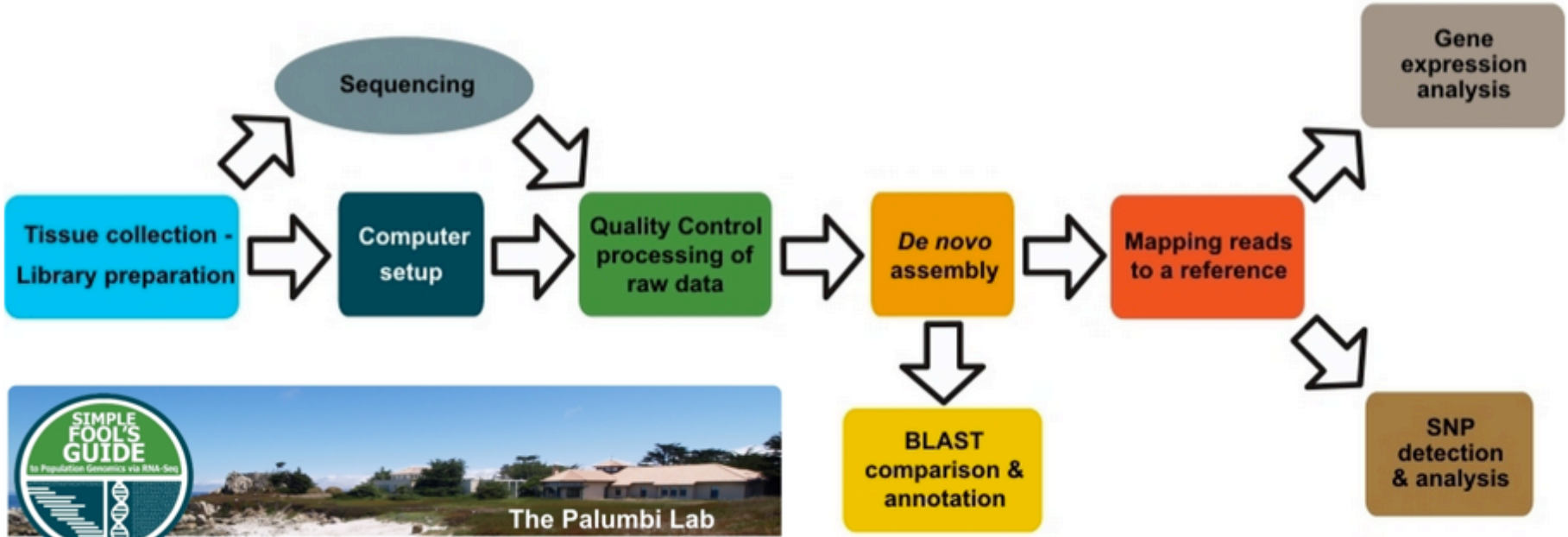
Robert Piskol,[1] Gokul Ramaswami,[1] and Jin Billy Li[1,*]

RESEARCH ARTICLE

Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data

Emma M. Quinn, Paul Cormican, Elaine M. Kenny, Matthew Hill, Richard Anney, Michael Gill, Aiden P. Corvin, Derek W. Morris

Published: March 26, 2013 • DOI: 10.1371/journal.pone.0058815



The Simple Fool's Guide to Population Genomics via RNA-Seq: An Introduction to High-Throughput Sequencing Data Analysis

This website and accompaning documents are intended as a tool to help researchers dealing with non-model organisms acquire and process transcriptomic high-throughput sequencing data without having to learn extensive bioinformatics skills. It covers all steps from tissue collection, sample preparation and computer setup, through addressing biological questions with gene expression and SNP data.

You may cite this work as follows:

De Wit P, Pespeni MH, Ladner JT, Barshis DJ, Seneca F, Jaris H, Overgaard Therkildsen N, Morikawa M and Palumbi SR (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources* **12**, 1058-1067.

Data analysis (and there are no pipelines)

- Demultiplex
- Look at the data (for example, fastQC)
- Trim adapters
- Trim low quality reads and/or bases
- Continue with **worksheet**!

Thanks!

Joanna Kelley
joanna.l.kelley@wsu.edu
https://kelleylab.wordpress.com/

WASHINGTON STATE UNIVERSITY
World Class. Face to Face.