kallisto live demo

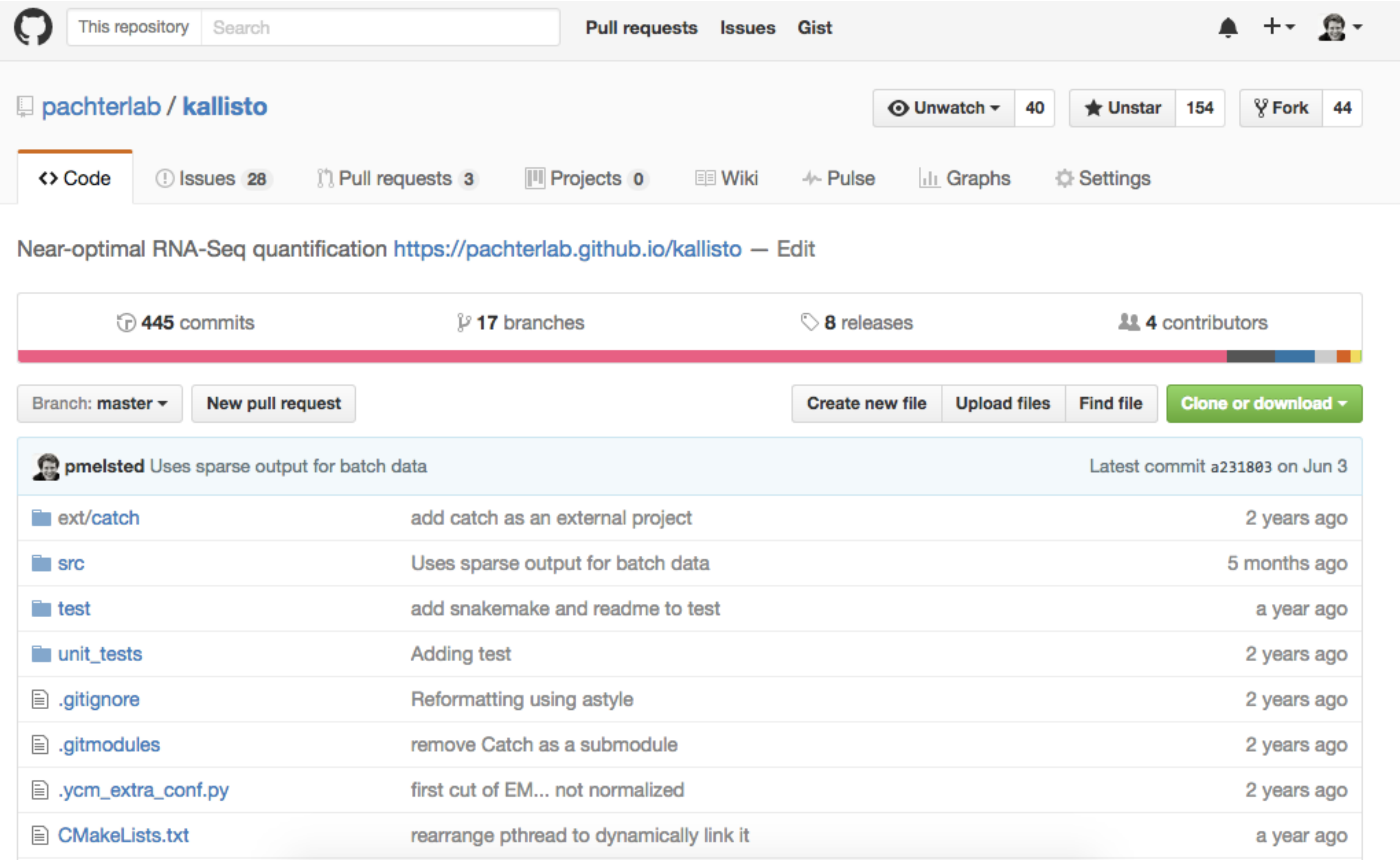
Páll Melsted

Overview

- Downloading and installing kallisto
- Ingredients for a Coelacanth analysis
    - Nikaido *et al.* Genome Research, 2013
- Indexing
- Quantification

Downloading and installing kallisto

- Download the source code and compile?



Near-optimal RNA-Seq quantification https://pachterlab.github.io/kallisto — Edit

445 commits | 17 branches | 8 releases | 4 contributors

pmelsted Uses sparse output for batch data — Latest commit a231803 on Jun 3

ext/catch	add catch as an external project	2 years ago
src	Uses sparse output for batch data	5 months ago
test	add snakemake and readme to test	a year ago
unit_tests	Adding test	2 years ago
.gitignore	Reformatting using astyle	2 years ago
.gitmodules	remove Catch as a submodule	2 years ago
.ycm_extra_conf.py	first cut of EM... not normalized	2 years ago
CMakeLists.txt	rearrange pthread to dynamically link it	a year ago

Downloading and installing kallisto

Mac: use Homebrew science

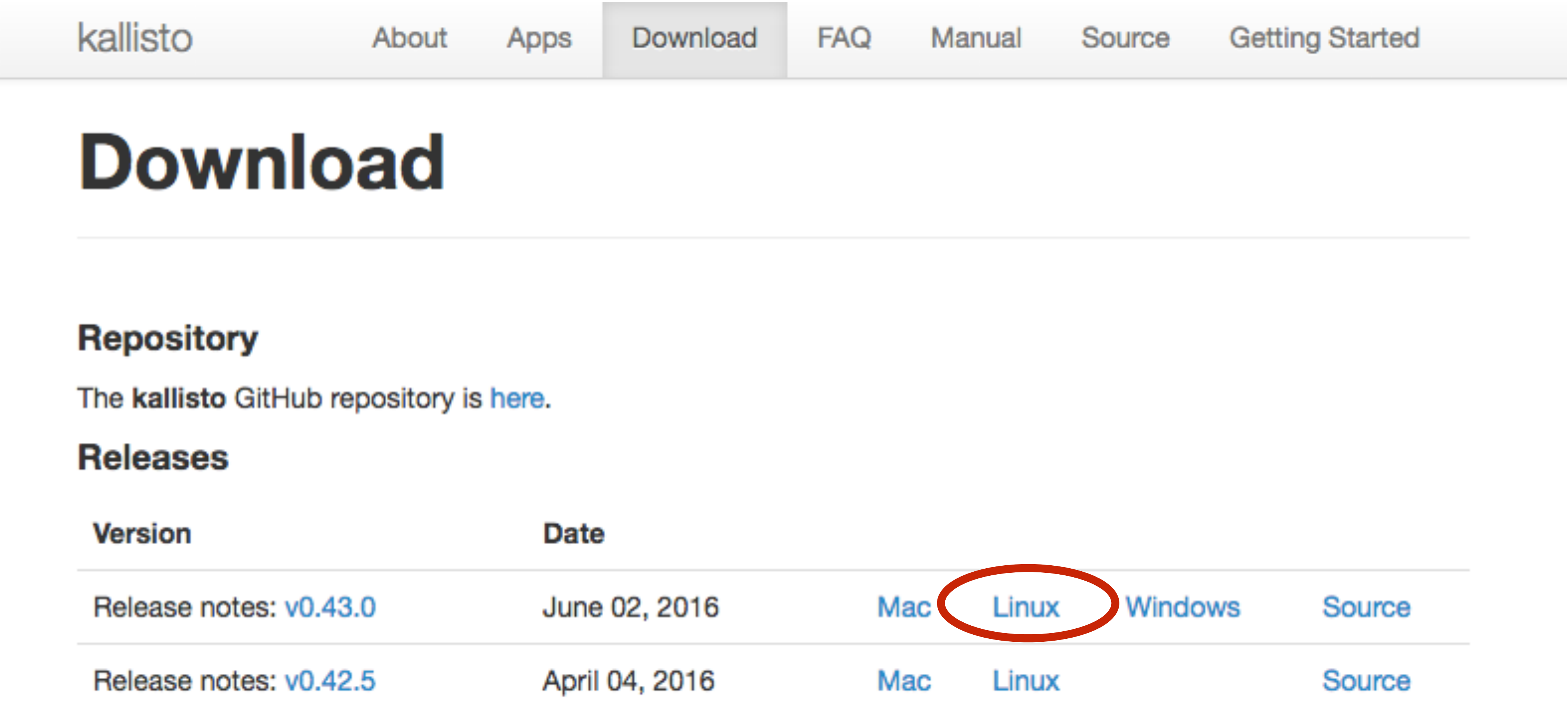
- Install Homebrew from http://brew.sh

```
> brew tap homebrew/science
> brew install kallisto
```

- Can also download binary, but Homebrew science adds a lot of useful tools

Downloading and installing kallisto

Linux: download the binary (works on all distributions) from https://pachterlab.github.io/kallisto



Download

Repository

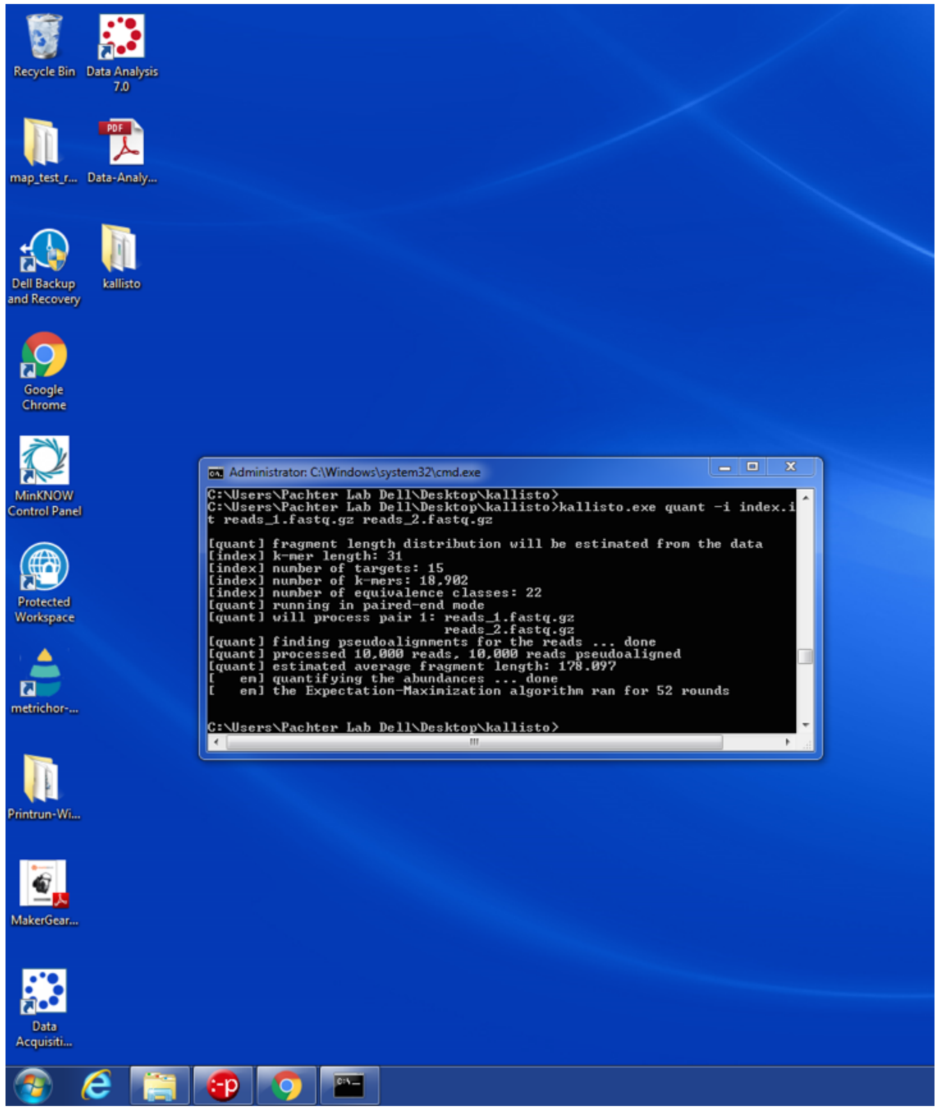
The **kallisto** GitHub repository is here.

Releases

Version	Date				
Release notes: v0.43.0	June 02, 2016	Mac	[✓ Linux]	Windows	Source
Release notes: v0.42.5	April 04, 2016	Mac	Linux		Source

Downloading and installing kallisto

Windows: download binary and run using command line (cmd)



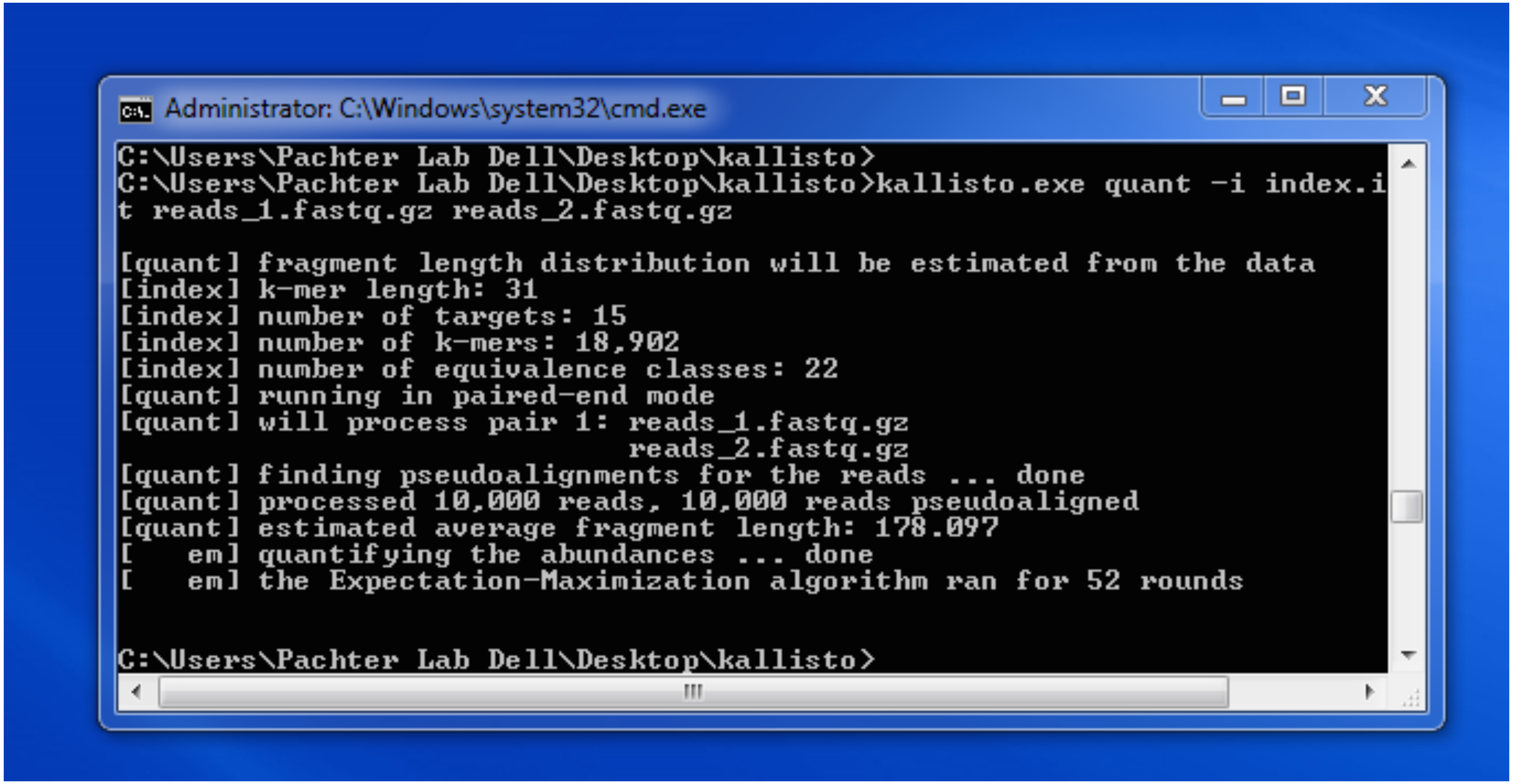
```
C:\Users\Pachter Lab Dell\Desktop\kallisto>
C:\Users\Pachter Lab Dell\Desktop\kallisto>kallisto.exe quant -i index.i
t reads_1.fastq.gz reads_2.fastq.gz

[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 15
[index] number of k-mers: 18,902
[index] number of equivalence classes: 22
[quant] running in paired-end mode
[quant] will process pair 1: reads_1.fastq.gz
                             reads_2.fastq.gz
[quant] finding pseudoalignments for the reads ... done
[quant] processed 10,000 reads, 10,000 reads pseudoaligned
[quant] estimated average fragment length: 178.097
[   em] quantifying the abundances ... done
[   em] the Expectation-Maximization algorithm ran for 52 rounds


C:\Users\Pachter Lab Dell\Desktop\kallisto>
```

Downloading and installing kallisto

Windows: download binary and run using command line (cmd)



```
C:\Users\Pachter Lab Dell\Desktop\kallisto>
C:\Users\Pachter Lab Dell\Desktop\kallisto>kallisto.exe quant -i index.i
t reads_1.fastq.gz reads_2.fastq.gz

[quant] fragment length distribution will be estimated from the data
[index] k-mer length: 31
[index] number of targets: 15
[index] number of k-mers: 18,902
[index] number of equivalence classes: 22
[quant] running in paired-end mode
[quant] will process pair 1: reads_1.fastq.gz
                             reads_2.fastq.gz
[quant] finding pseudoalignments for the reads ... done
[quant] processed 10,000 reads, 10,000 reads pseudoaligned
[quant] estimated average fragment length: 178.097
[   em] quantifying the abundances ... done
[   em] the Expectation-Maximization algorithm ran for 52 rounds


C:\Users\Pachter Lab Dell\Desktop\kallisto>
```

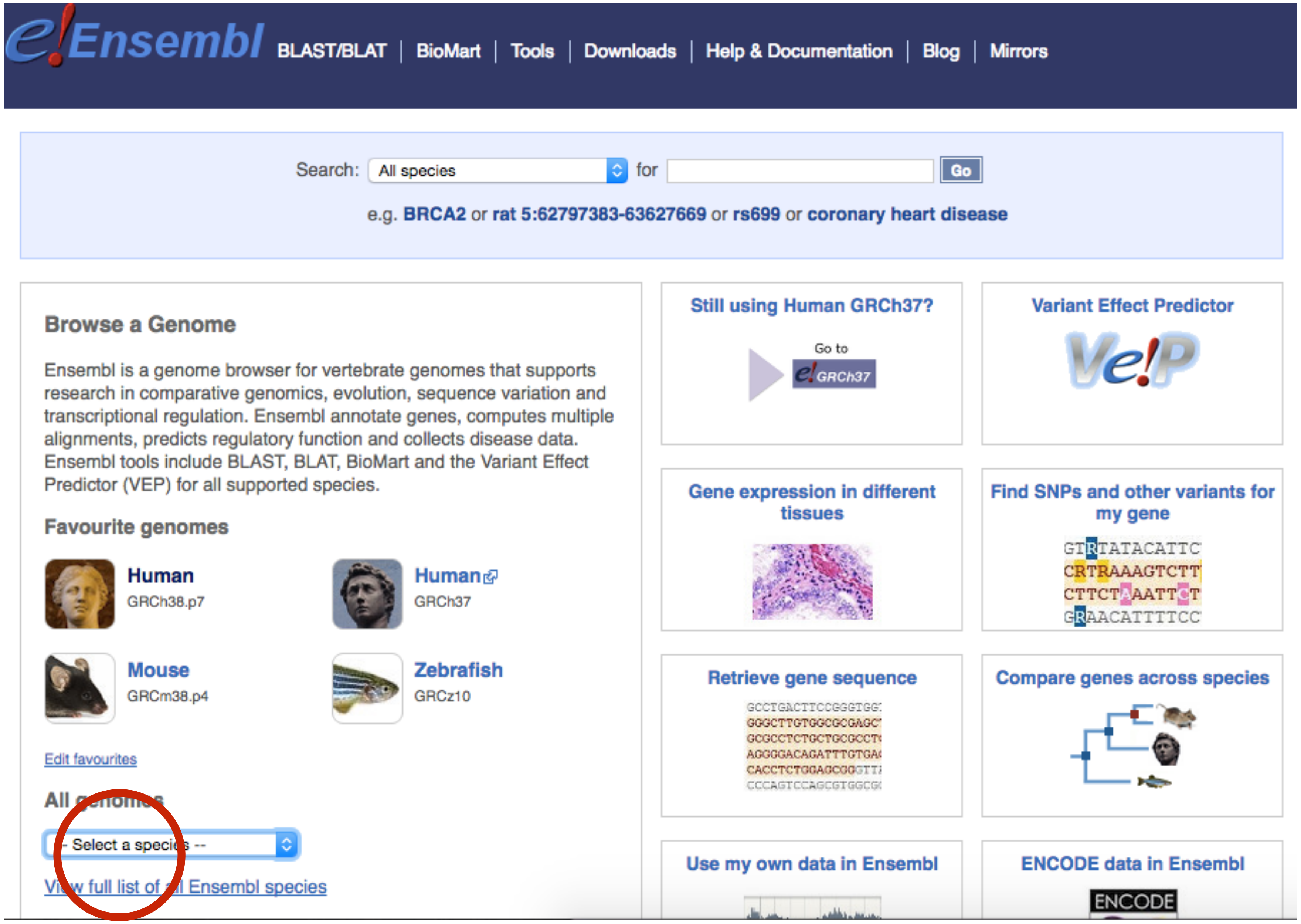
Requirements for running kallisto

- Reference transcriptome
- Paired end reads, FASTQ files

Coelacanth analysis

- Reference transcriptome
- Paired end reads, FASTQ files
  - Downloaded from the short read archive (SRA)
    - 19 samples
      - 6 tissues - gill, kidney, pharynx, pectoral fin, pelvic fin, muscle
      - ~3 technical replicates per tissue

Reference transcriptome



BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: All species for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Favourite genomes

Human GRCh38.p7

Human GRCh37

Mouse GRCm38.p4

Zebrafish GRCz10

Edit favourites

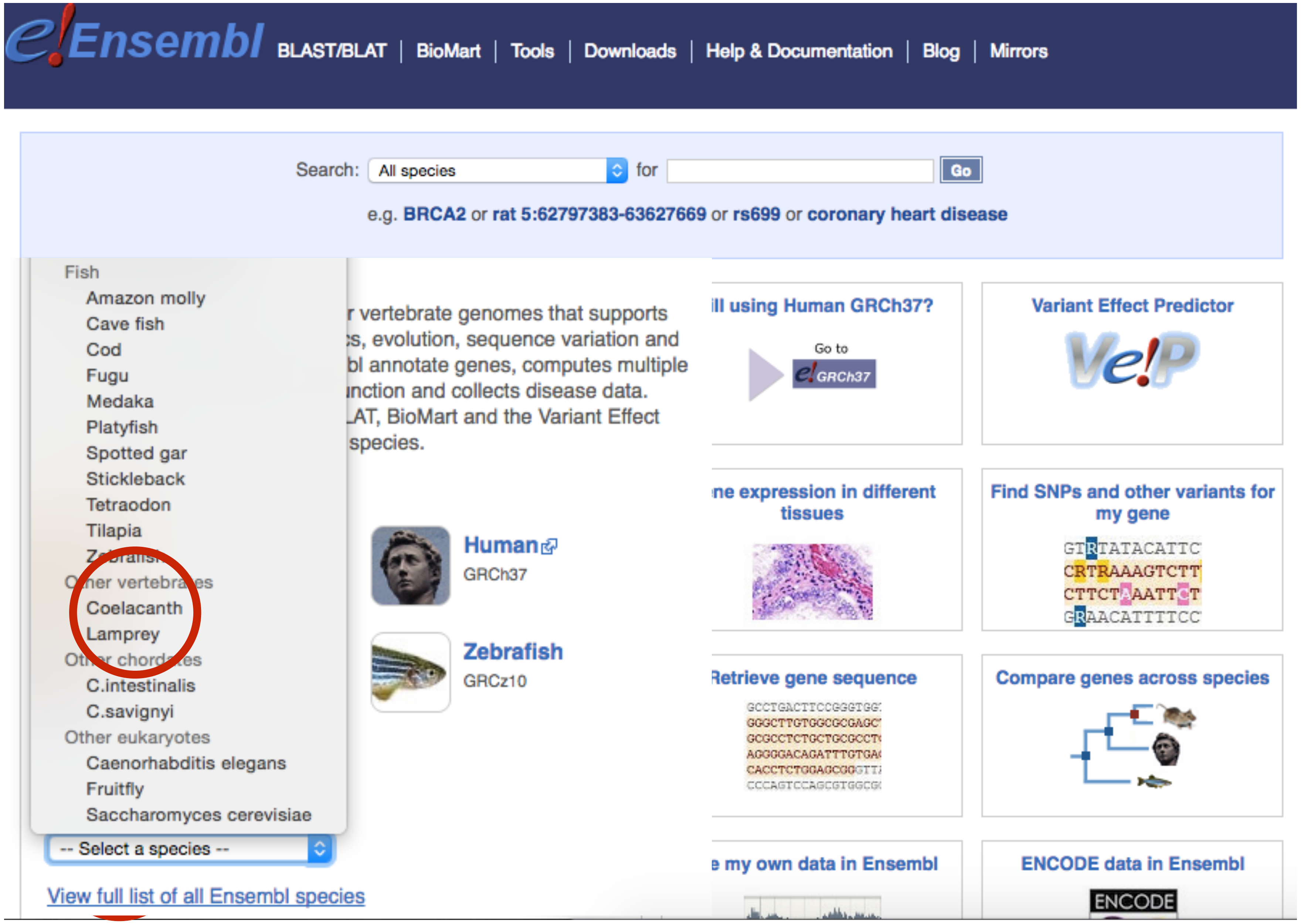
All genomes

-- Select a species --

View full list of all Ensembl species

Still using Human GRCh37?
Go to
GRCh37
Variant Effect Predictor
VEP
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
Compare genes across species
Use my own data in Ensembl
ENCODE data in Ensembl
ENCODE

Reference transcriptome



e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: All species for Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

Fish
- Amazon molly
- Cave fish
- Cod
- Fugu
- Medaka
- Platyfish
- Spotted gar
- Stickleback
- Tetraodon
- Tilapia
- Zebrafish

Other vertebrates
- Coelacanth
- Lamprey

Other chordates
- C.intestinalis
- C.savignyi

Other eukaryotes
- Caenorhabditis elegans
- Fruitfly
- Saccharomyces cerevisiae

-- Select a species --

View full list of all Ensembl species

r vertebrate genomes that supports
s, evolution, sequence variation and
bl annotate genes, computes multiple
nction and collects disease data.
AT, BioMart and the Variant Effect
species.

Human GRCh37

Zebrafish GRCz10

ll using Human GRCh37?

Go to e! GRCh37

Variant Effect Predictor

ne expression in different tissues

Find SNPs and other variants for my gene

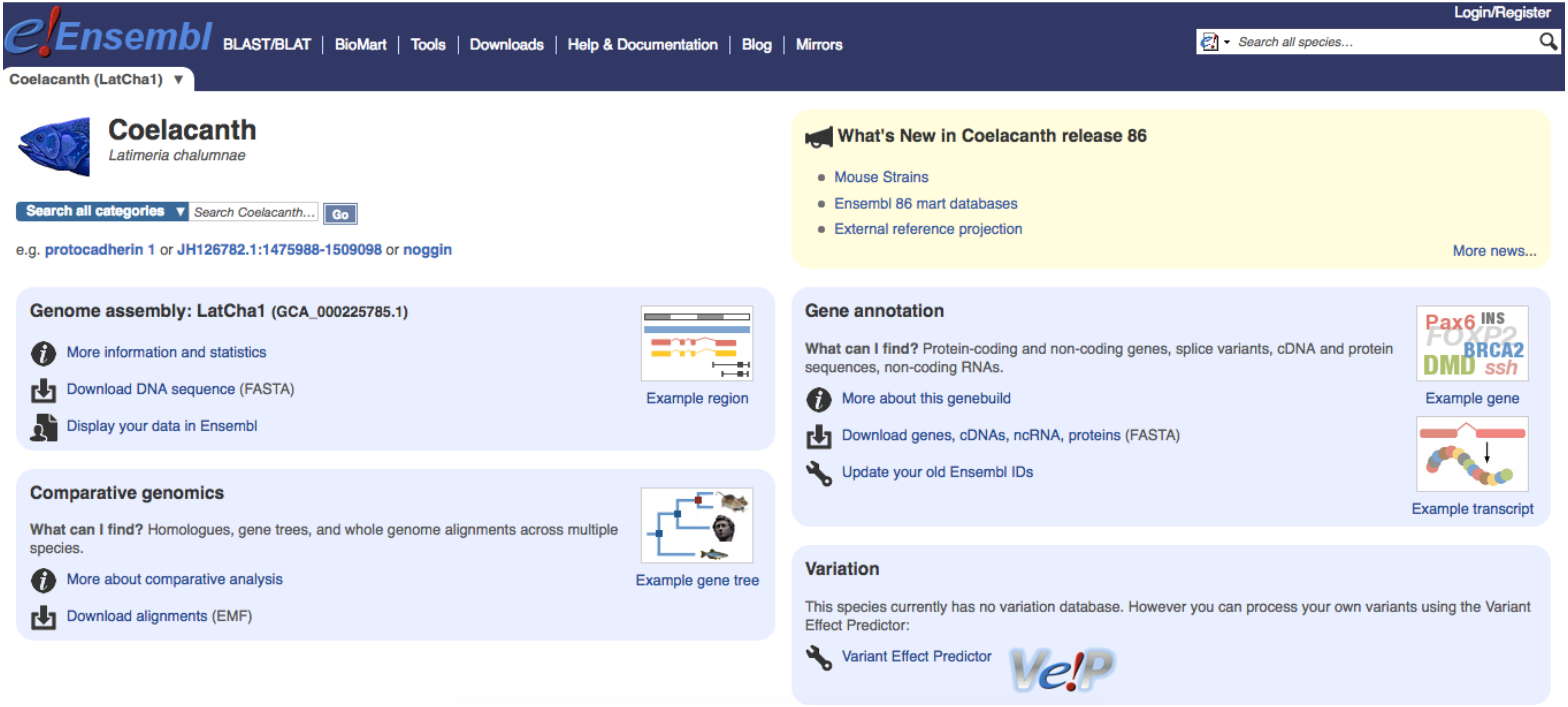
Retrieve gene sequence

Compare genes across species

e my own data in Ensembl

ENCODE data in Ensembl

Reference transcriptome



Coelacanth

Latimeria chalumnae

e.g. protocadherin 1 or JH126782.1:1475988-1509098 or noggin

What's New in Coelacanth release 86

- Mouse Strains
- Ensembl 86 mart databases
- External reference projection

More news...

Genome assembly: LatCha1 (GCA_000225785.1)

More information and statistics

Download DNA sequence (FASTA)

Display your data in Ensembl

Example region

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

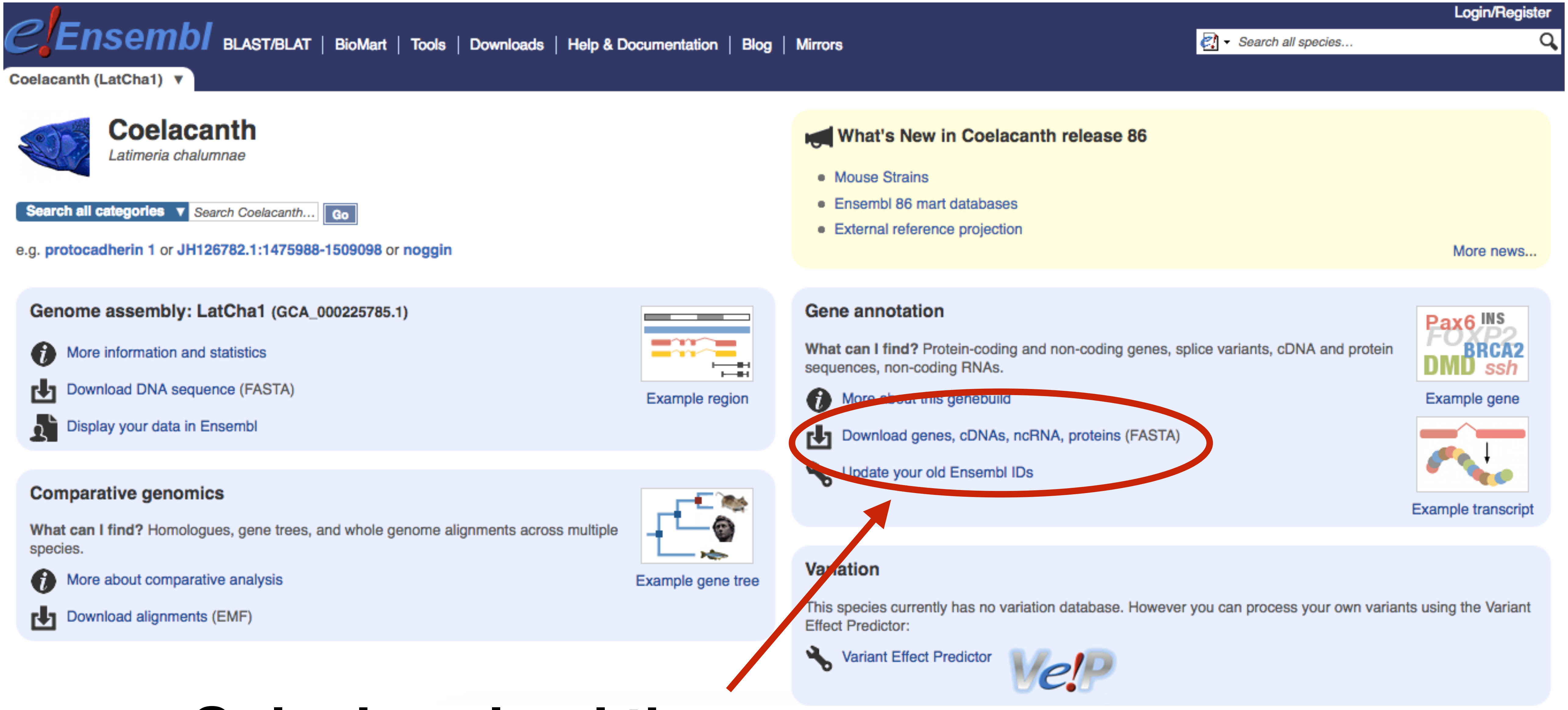
Example gene tree

Variation

This species currently has no variation database. However you can process your own variants using the Variant Effect Predictor:

Variant Effect Predictor

Reference transcriptome



Coelacanth

Latimeria chalumnae

e.g. protocadherin 1 or JH126782.1:1475988-1509098 or noggin

What's New in Coelacanth release 86

- Mouse Strains
- Ensembl 86 mart databases
- External reference projection

More news...

Genome assembly: LatCha1 (GCA_000225785.1)

More information and statistics

Download DNA sequence (FASTA)

Display your data in Ensembl

Example region

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Example gene tree

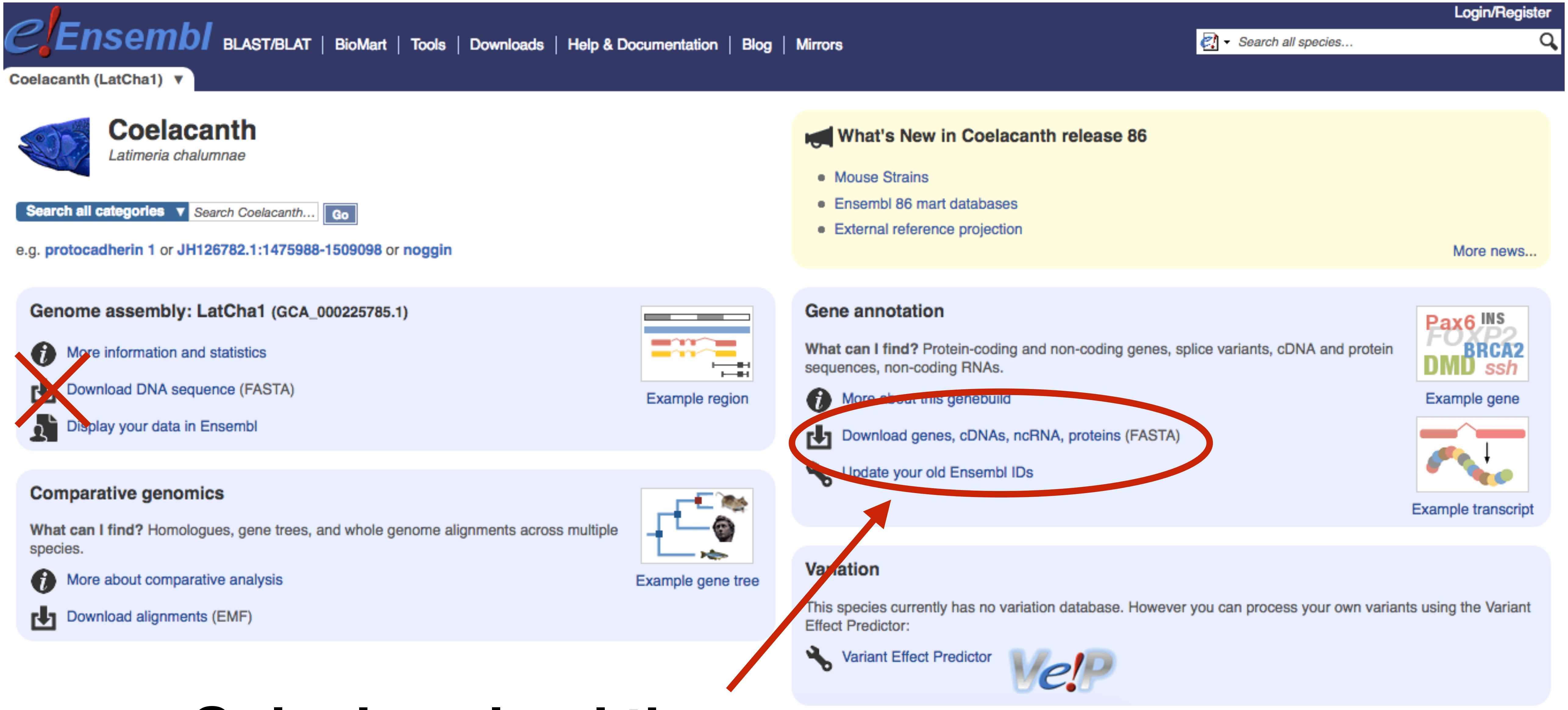
Variation

This species currently has no variation database. However you can process your own variants using the Variant Effect Predictor:

Variant Effect Predictor

Only download the transcriptome

Reference transcriptome



Login/Register

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search all species...

Coelacanth (LatCha1) ▼

Coelacanth

Latimeria chalumnae

Search all categories ▼ Search Coelacanth... Go

e.g. **protocadherin 1** or **JH126782.1:1475988-1509098** or **noggin**

What's New in Coelacanth release 86

- Mouse Strains
- Ensembl 86 mart databases
- External reference projection

More news...

Genome assembly: LatCha1 (GCA_000225785.1)

More information and statistics

Download DNA sequence (FASTA)

Display your data in Ensembl

Example region

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Example gene tree

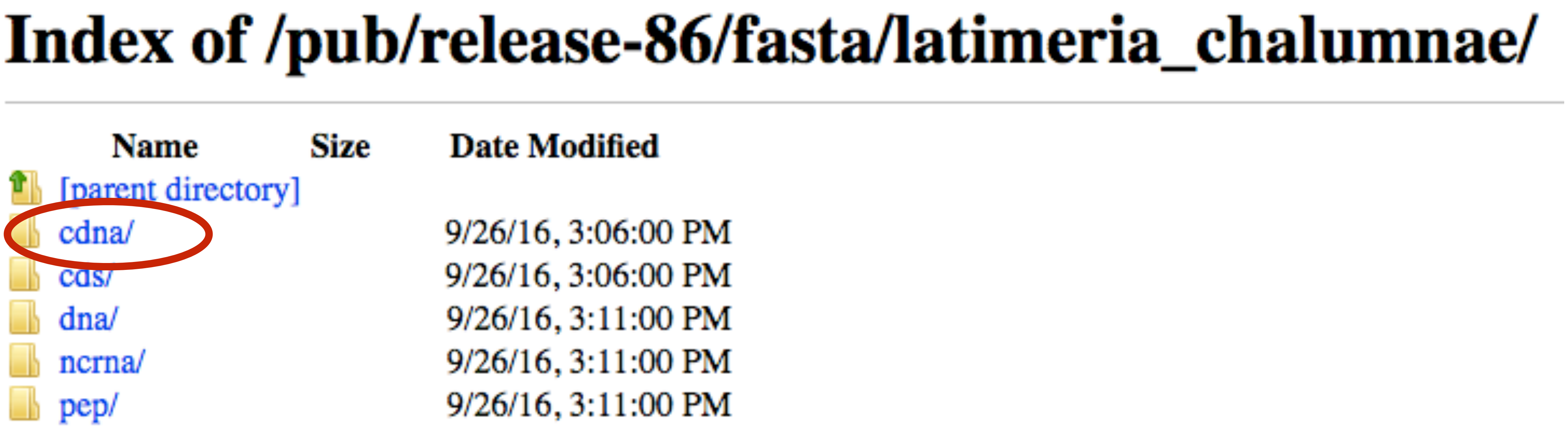
Variation

This species currently has no variation database. However you can process your own variants using the Variant Effect Predictor:

Variant Effect Predictor

Only download the transcriptome

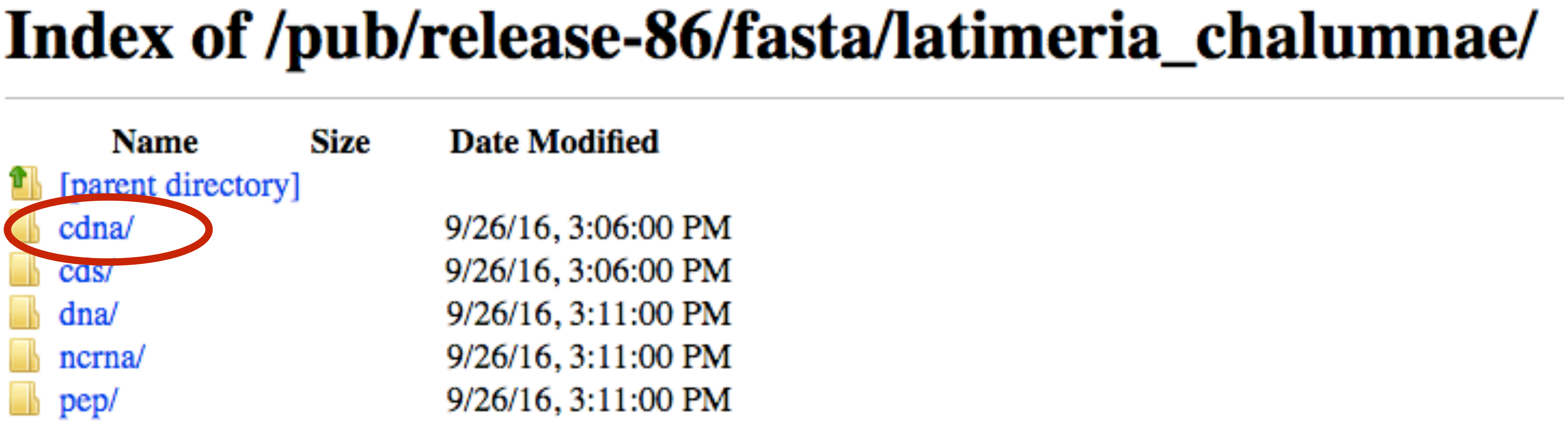
Reference transcriptome



Index of /pub/release-86/fasta/latimeria_chalumnae/

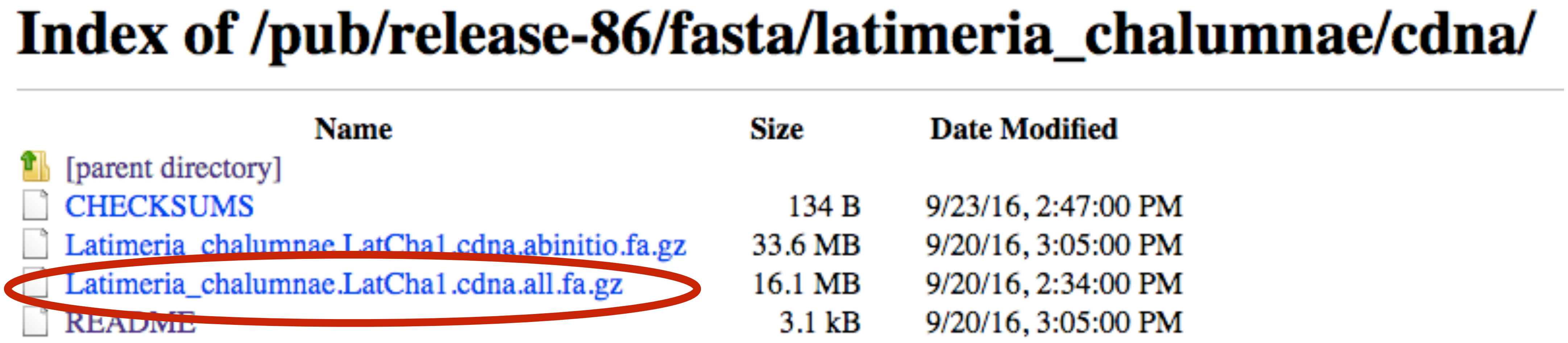
Name	Size	Date Modified
[parent directory]		
cdna/		9/26/16, 3:06:00 PM
cds/		9/26/16, 3:06:00 PM
dna/		9/26/16, 3:11:00 PM
ncrna/		9/26/16, 3:11:00 PM
pep/		9/26/16, 3:11:00 PM

Reference transcriptome



Index of /pub/release-86/fasta/latimeria_chalumnae/

Name	Size	Date Modified
[parent directory]		
cdna/		9/26/16, 3:06:00 PM
cds/		9/26/16, 3:06:00 PM
dna/		9/26/16, 3:11:00 PM
ncrna/		9/26/16, 3:11:00 PM
pep/		9/26/16, 3:11:00 PM



Index of /pub/release-86/fasta/latimeria_chalumnae/cdna/

Name	Size	Date Modified
[parent directory]		
CHECKSUMS	134 B	9/23/16, 2:47:00 PM
Latimeria_chalumnae.LatCha1.cdna.abinitio.fa.gz	33.6 MB	9/20/16, 3:05:00 PM
Latimeria_chalumnae.LatCha1.cdna.all.fa.gz	16.1 MB	9/20/16, 2:34:00 PM
README	3.1 kB	9/20/16, 3:05:00 PM

Indexing command

- Creates an index of the transcriptome reference that kallisto will use for quantification
    - Required only once per reference transcriptome

```
> time kallisto index -i index.idx
  Latimeria_chalumnae.LatCha1.cdna.all.fa.gz
```

Indexing command

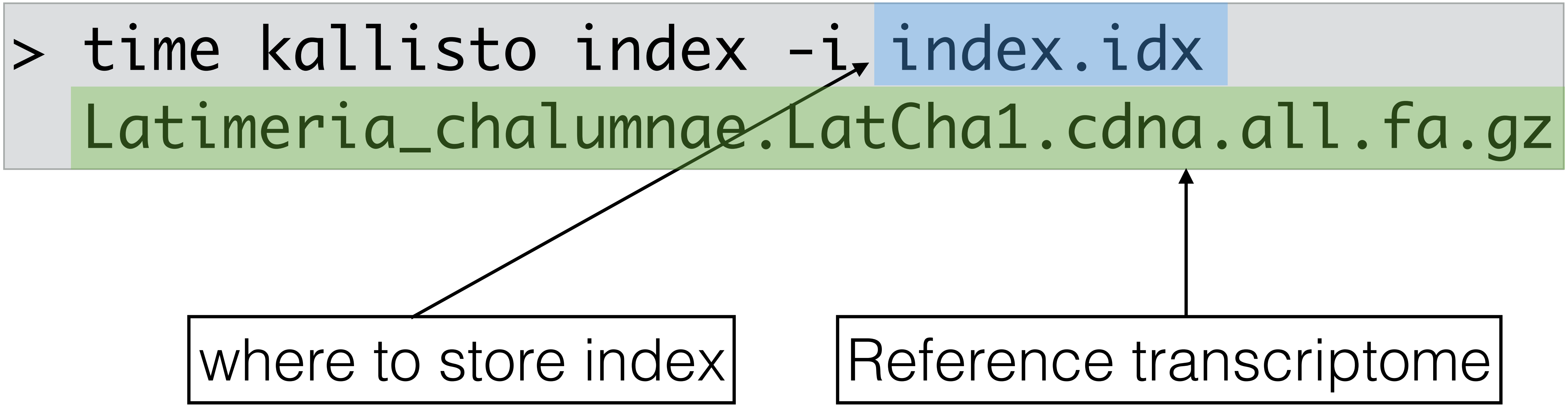
- Creates an index of the transcriptome reference that kallisto will use for quantification
  - Required only once per reference transcriptome

```
> time kallisto index -i index.idx
  Latimeria_chalumnae.LatCha1.cdna.all.fa.gz
```

where to store index

Indexing command

- Creates an index of the transcriptome reference that kallisto will use for quantification
    - Required only once per reference transcriptome



```
> time kallisto index -i index.idx
  Latimeria_chalumnae.LatCha1.cdna.all.fa.gz
```

where to store index

Reference transcriptome

Indexing guide

What value of k should be chosen?

- k-mer tradeoffs
  - high k - more specific
  - low k - robust to sequencing errors

Indexing guide

What value of k should be chosen?

- k-mer tradeoffs
    - high k - more specific
    - low k - robust to sequencing errors
- general guide
    - for 75bp reads, use default, k=31
    - for 50bp reads, use default except if known issues
    - shorter reads, lower k=25 or k=21.

Quantification

Quantification is run separately for each sample

```
> time kallisto quant -i index/index.idx
  -o results/DRR002318 -b 30 -t 4
  data/DRR002318_1.fastq.gz
  data/DRR002318_2.fastq.gz
```

Quantification

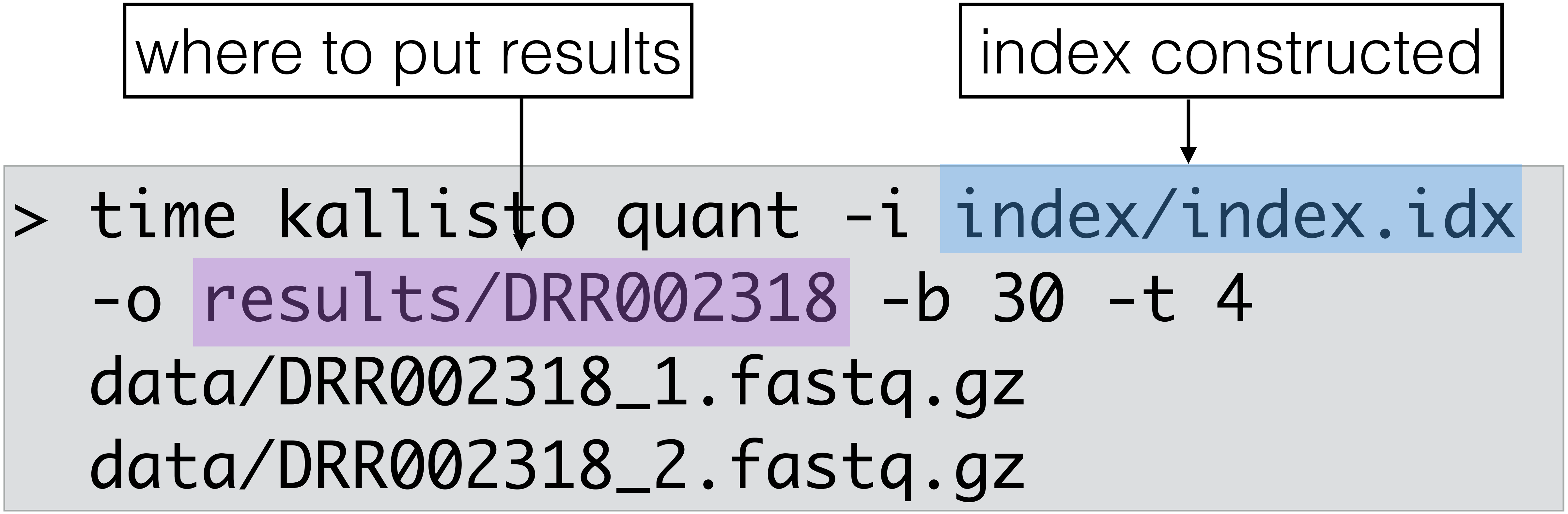
Quantification is run separately for each sample

index constructed

```
> time kallisto quant -i index/index.idx
  -o results/DRR002318 -b 30 -t 4
  data/DRR002318_1.fastq.gz
  data/DRR002318_2.fastq.gz
```

Quantification

Quantification is run separately for each sample



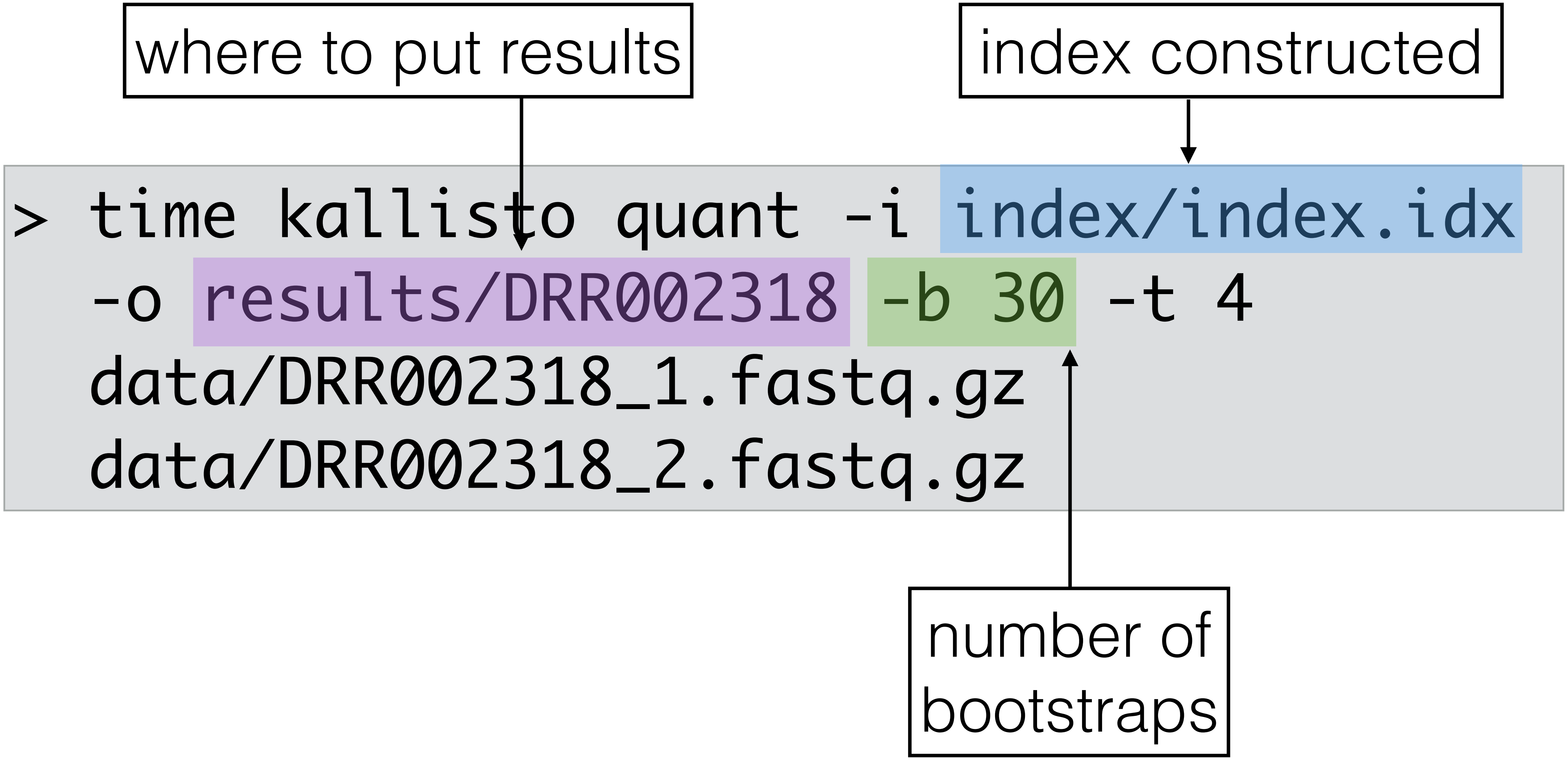
where to put results

index constructed

```
> time kallisto quant -i index/index.idx
  -o results/DRR002318 -b 30 -t 4
  data/DRR002318_1.fastq.gz
  data/DRR002318_2.fastq.gz
```

Quantification

Quantification is run separately for each sample



where to put results

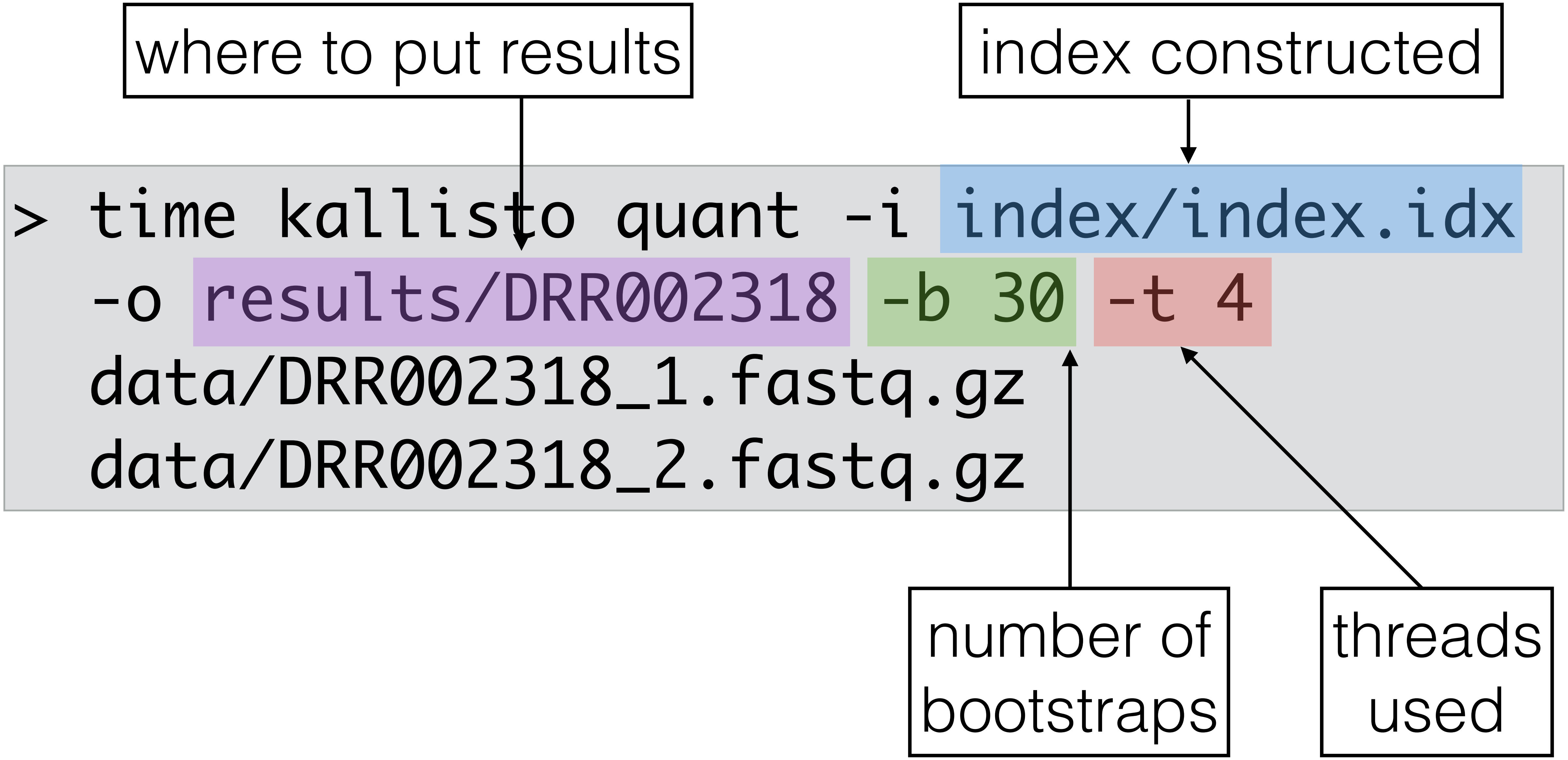
index constructed

```
> time kallisto quant -i index/index.idx
  -o results/DRR002318 -b 30 -t 4
  data/DRR002318_1.fastq.gz
  data/DRR002318_2.fastq.gz
```

number of bootstraps

Quantification

Quantification is run separately for each sample



```
> time kallisto quant -i index/index.idx
  -o results/DRR002318 -b 30 -t 4
  data/DRR002318_1.fastq.gz
  data/DRR002318_2.fastq.gz
```

where to put results

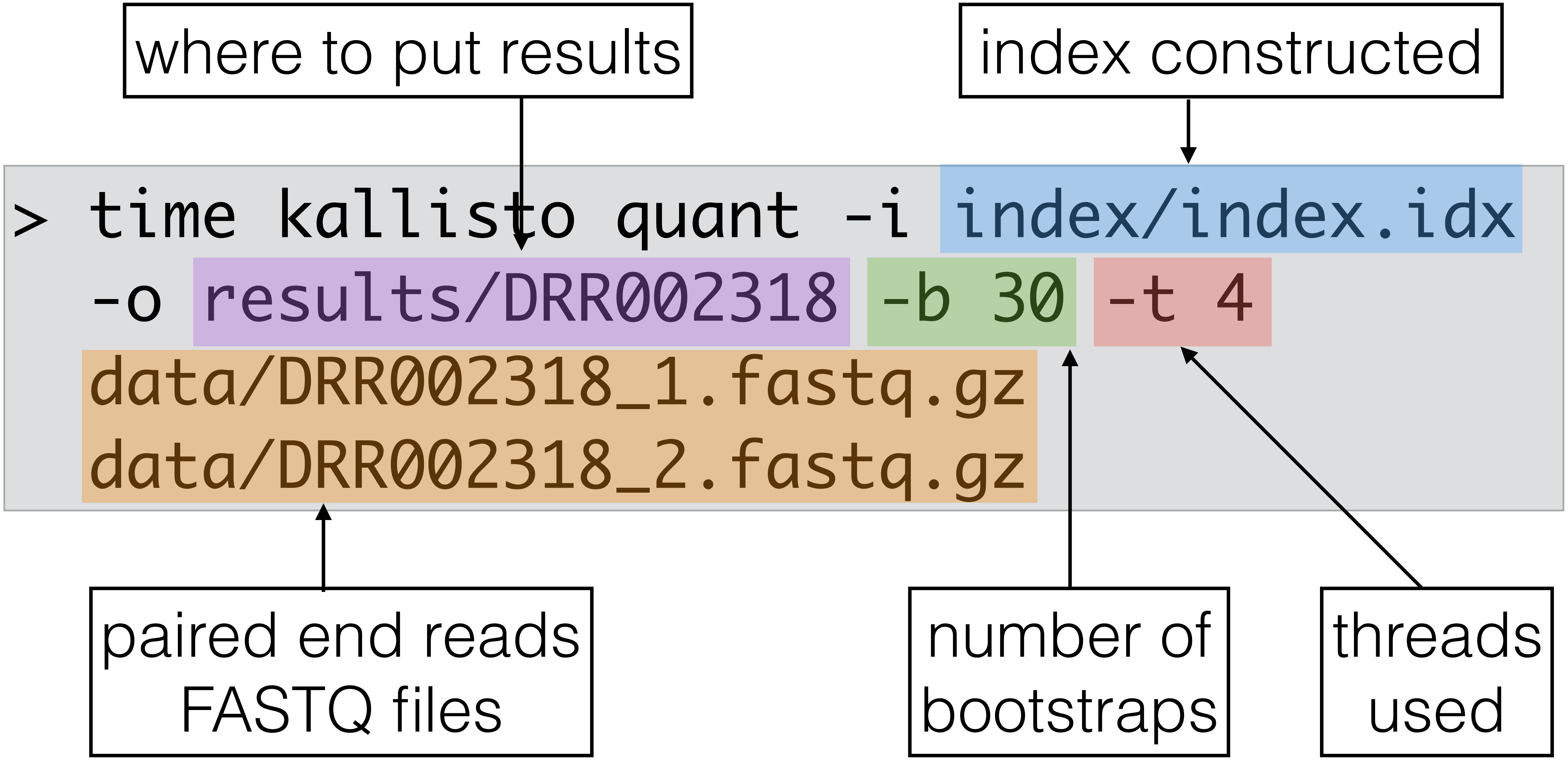
index constructed

number of bootstraps

threads used

Quantification

Quantification is run separately for each sample



where to put results

index constructed

```
> time kallisto quant -i index/index.idx
  -o results/DRR002318 -b 30 -t 4
  data/DRR002318_1.fastq.gz
  data/DRR002318_2.fastq.gz
```

paired end reads FASTQ files

number of bootstraps

threads used

Results

- The output is stored in the specified directory
- `abundance.h5` - HDF5 compressed file, not human readable
  - contains the quantifications, all bootstraps and other information
- `abundance.tsv` - Tab Separated file
  - abundances and counts for each transcript
- `run_info.json` - JSON formatted file
  - information about the run