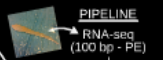
THOUGHTS AND IDEAS TOWARD OPTIMAL

RNA SEQ

RNA AS AN HINDU GODDESS ???

=

RNA's will works in mysterious ways



PIPELINE

RNA-seq (100 bp - PE)

1. CLEANUP (?)

2. NORMALIZE (?)

3. ASSEMBLE

4. REDUNDANCY

5. ANNOTATE

PLUS

Evaluation of *de novo* transcriptome assemblies from RNA-Seq data

http://deweylab.biostat.wisc.edu/rsem/rsem-eval.html

"reference-free de novo transcriptome assembly evaluator"

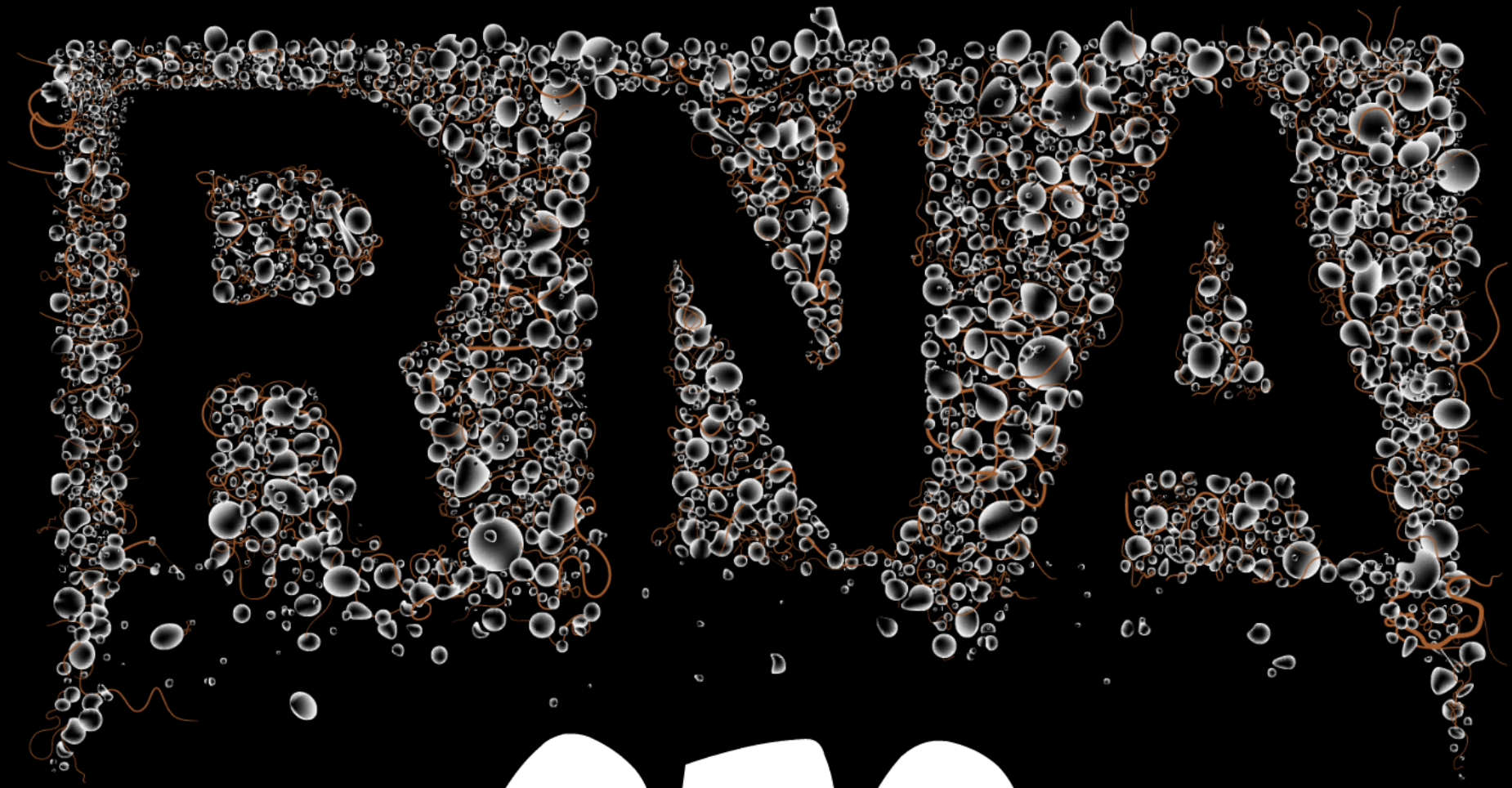
RESULTS SO FAR...

Assembly number	Trim quality (PHRED)	Normalized	Number of transcripts	Number of genes	unigene N50
1	20	Y	168 089	129 917	1655
2	10	Y	168 649	129 965	1640
3	5	Y	168 702	129 894	1661
4	2	Y	168 797	129 729	1650
5	20	N	174 531	144 326	1435
6	10	N	173 373	145 450	1425
7	5	N	175 334	145 410	1426
8	2	N	175 479	145 407	1427

reads = 19 370 403

assembled bases = 4 682 256 300

THOUGHTS AND IDEAS TOWARD OPTIMAL

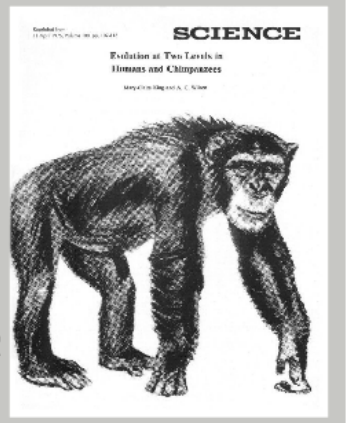


SEQ

RNA

RNA = powerful biological tool

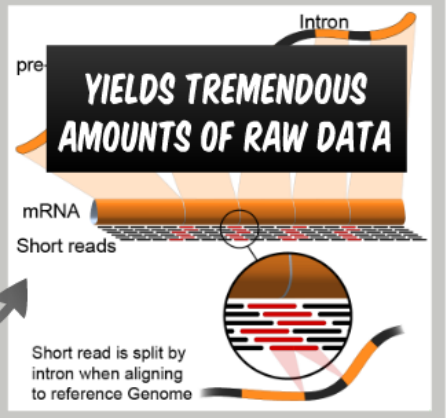
- Response to the environment
- Ecological interactions
- Inter/Intra specific divergence



(King & Wilson 1975)

GETTING RNA = EASY MONEY !

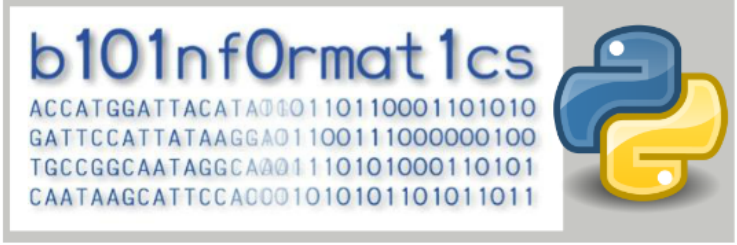
Any biological tissue OR whole organism



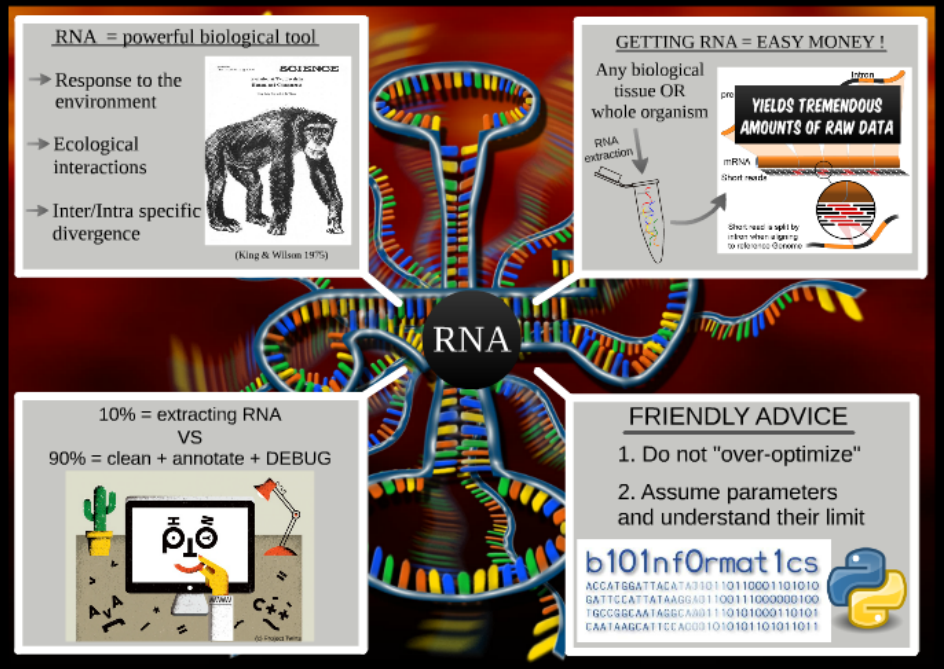
10% = extracting RNA VS 90% = clean + annotate + DEBUG

FRIENDLY ADVICE

1. Do not "over-optimize"
2. Assume parameters and understand their limit



RNA AS AN HINDU GODDESS ???



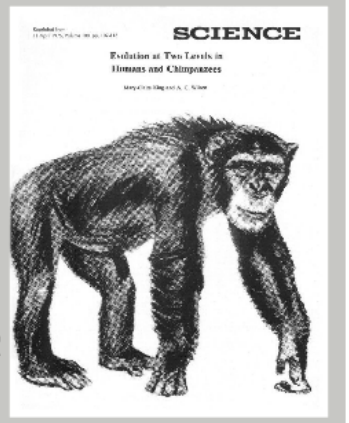
=

RNA's will works in mysterious ways

RNA

RNA = powerful biological tool

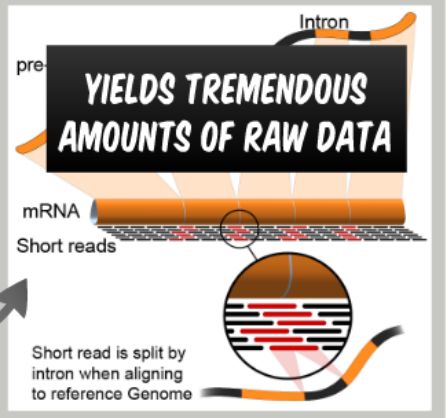
- Response to the environment
- Ecological interactions
- Inter/Intra specific divergence



(King & Wilson 1975)

GETTING RNA = EASY MONEY !

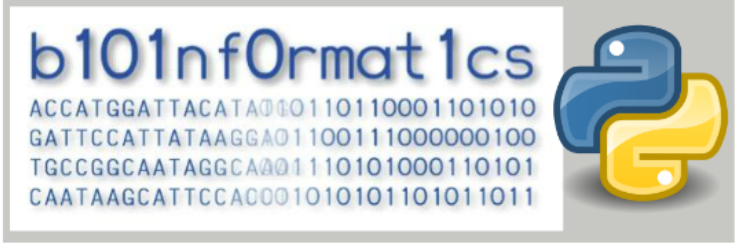
Any biological tissue OR whole organism



10% = extracting RNA VS 90% = clean + annotate + DEBUG

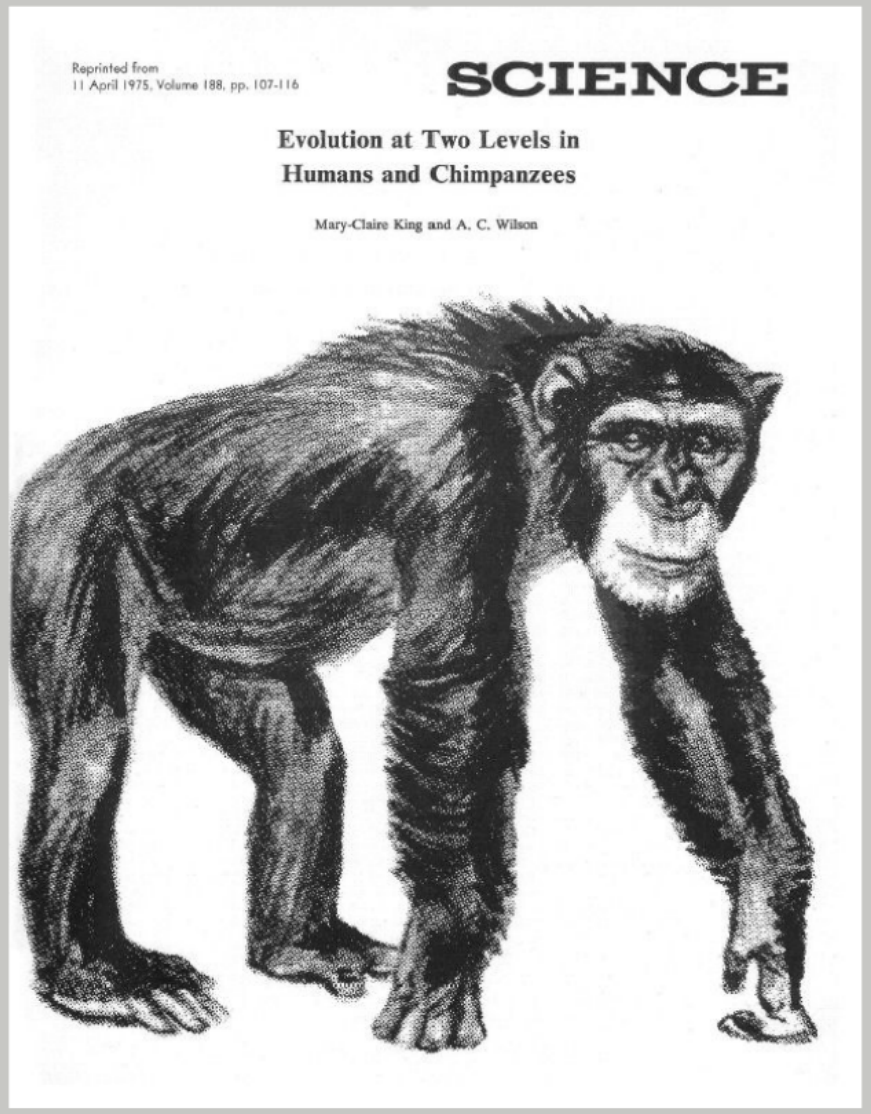
FRIENDLY ADVICE

1. Do not "over-optimize"
2. Assume parameters and understand their limit



RNA = powerful biological tool

- Response to the environment
- Ecological interactions
- Inter/Intra specific divergence

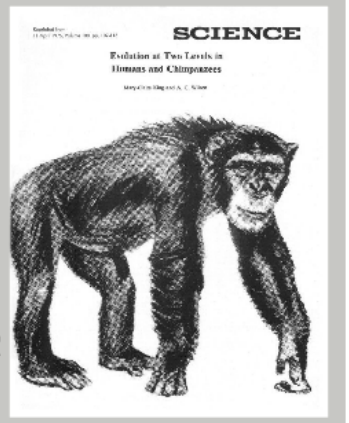


(King & Wilson 1975)

RNA

RNA = powerful biological tool

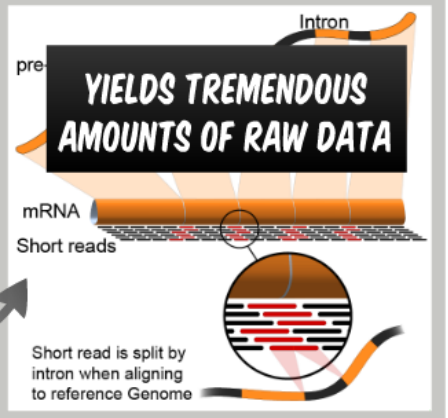
- Response to the environment
- Ecological interactions
- Inter/Intra specific divergence



(King & Wilson 1975)

GETTING RNA = EASY MONEY !

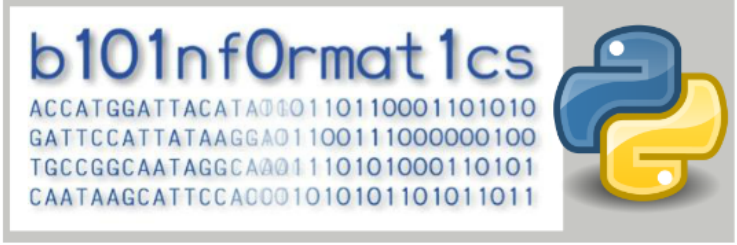
Any biological tissue OR whole organism



10% = extracting RNA VS 90% = clean + annotate + DEBUG

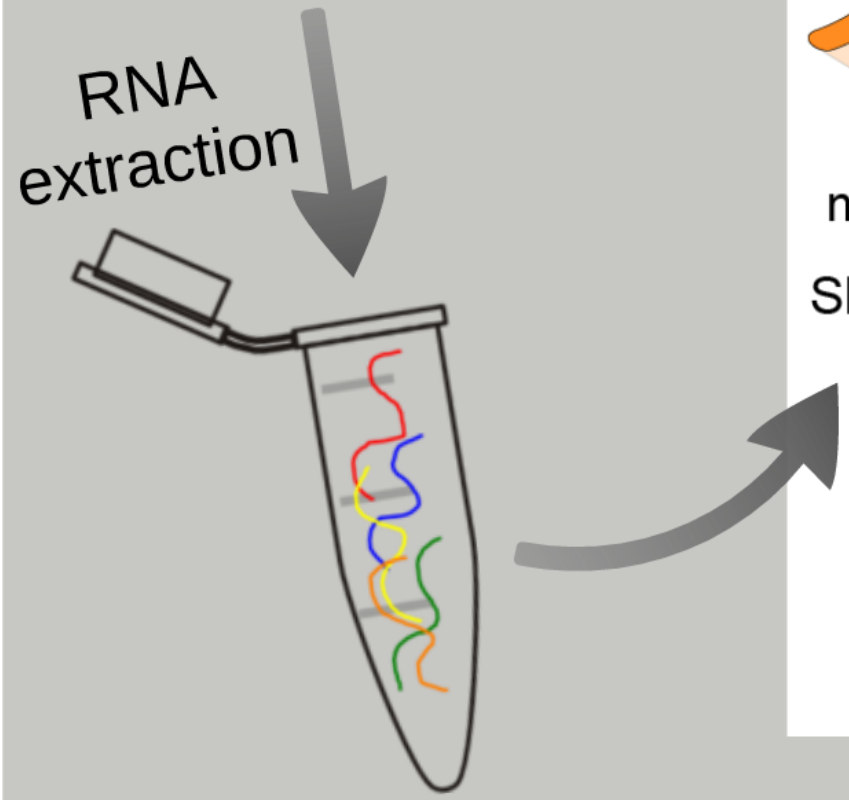
FRIENDLY ADVICE

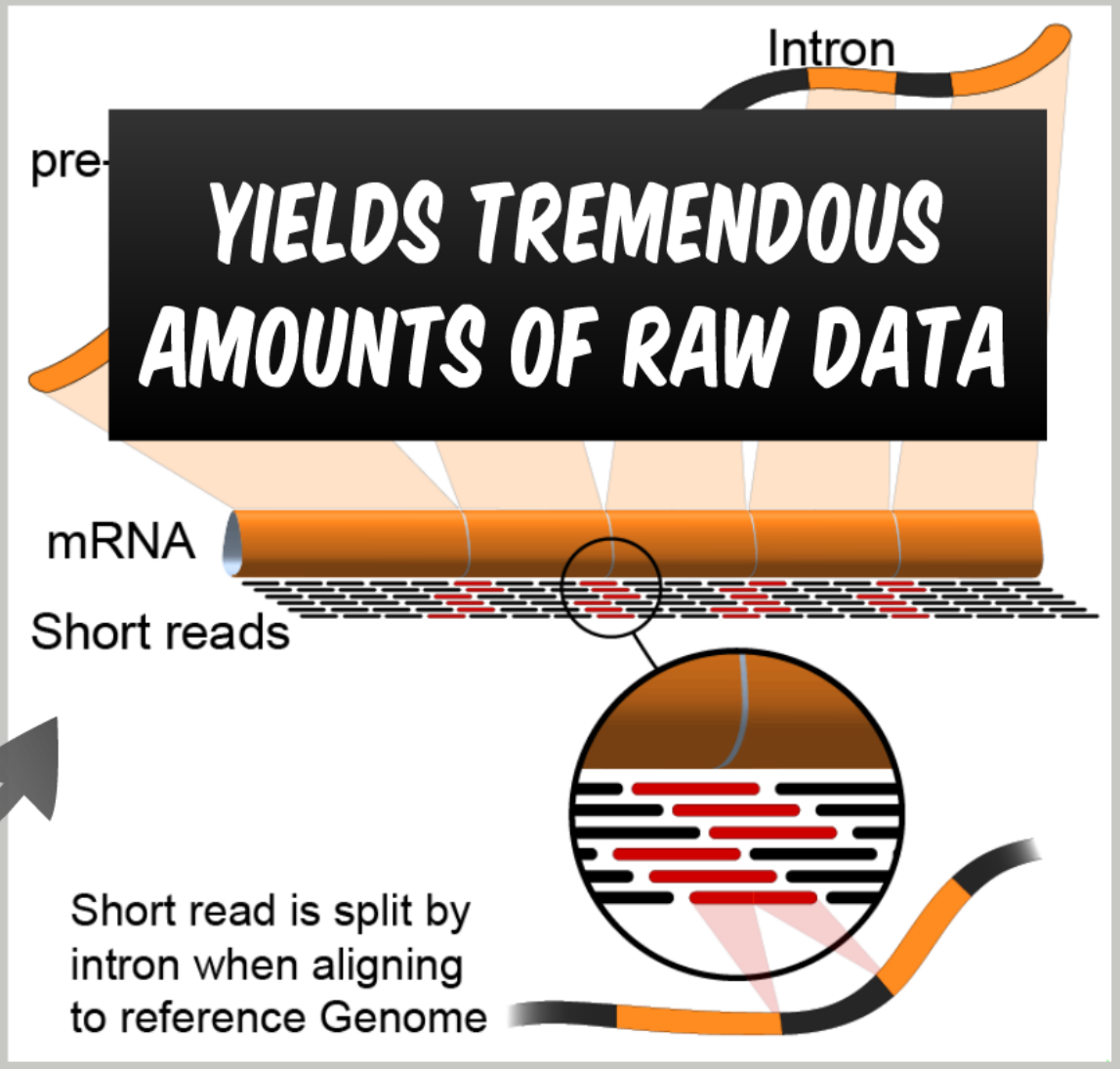
1. Do not "over-optimize"
2. Assume parameters and understand their limit



GETTING RNA = EASY MONEY !

Any biological tissue OR whole organism

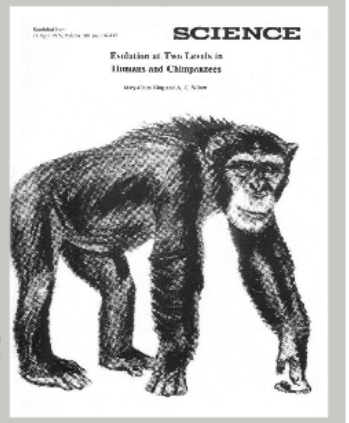




RNA

RNA = powerful biological tool

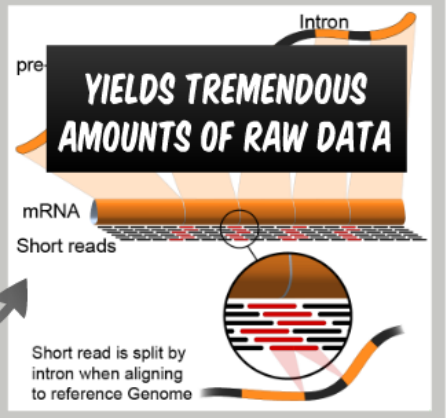
- Response to the environment
- Ecological interactions
- Inter/Intra specific divergence



(King & Wilson 1975)

GETTING RNA = EASY MONEY !

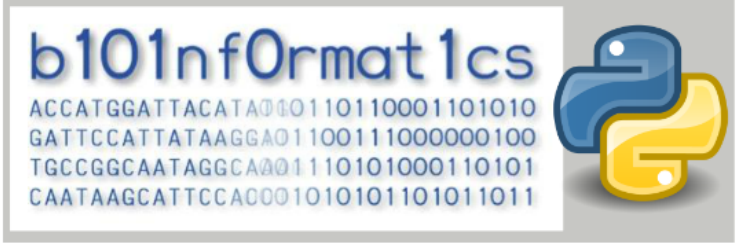
Any biological tissue OR whole organism



10% = extracting RNA VS 90% = clean + annotate + DEBUG

FRIENDLY ADVICE

1. Do not "over-optimize"
2. Assume parameters and understand their limit



10% = extracting RNA

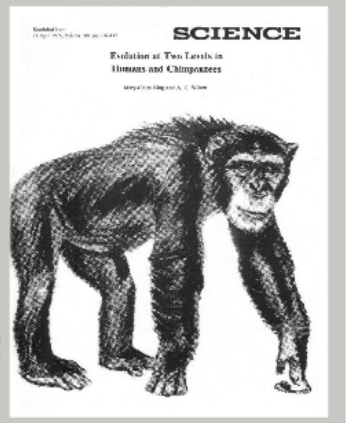
VS

90% = clean + annotate + DEBUG

RNA

RNA = powerful biological tool

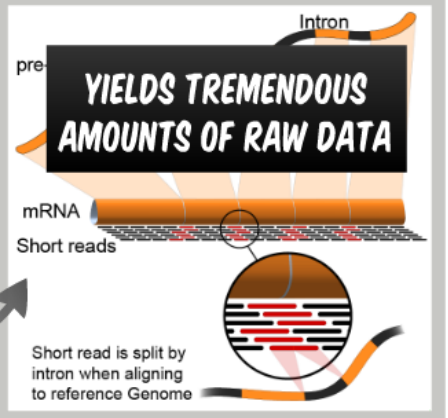
- Response to the environment
- Ecological interactions
- Inter/Intra specific divergence



(King & Wilson 1975)

GETTING RNA = EASY MONEY !

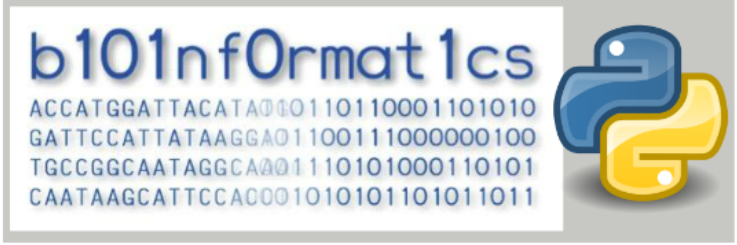
Any biological tissue OR whole organism



10% = extracting RNA VS 90% = clean + annotate + DEBUG

FRIENDLY ADVICE

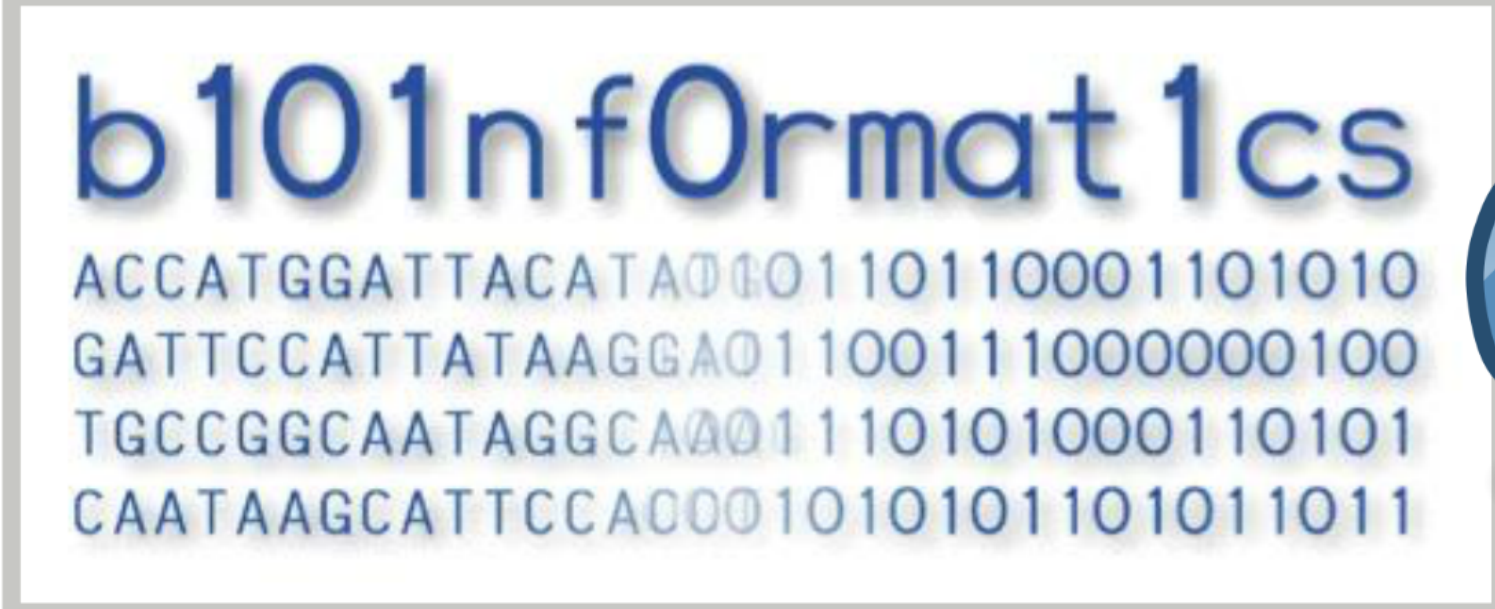
1. Do not "over-optimize"
2. Assume parameters and understand their limit



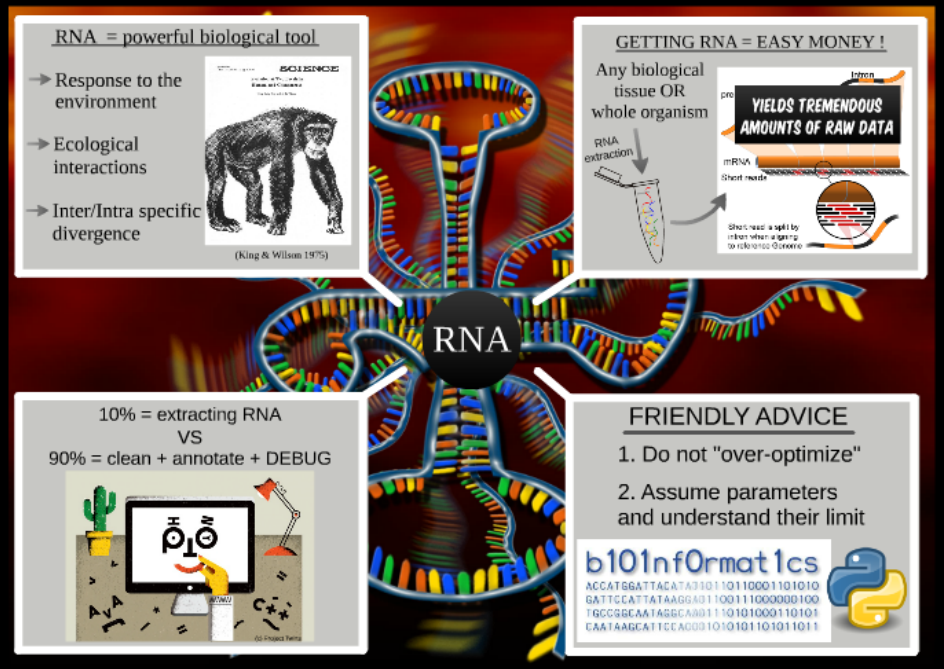
FRIENDLY ADVICE

1. Do not "over-optimize"

2. Assume parameters and understand their limit



RNA AS AN HINDU GODDESS ???



=

RNA's will works in mysterious ways

ess ???

PIPELINE

RNA-seq
(100 bp - PE)

1. CLEANUP (?)

2. NORMALIZE (?)

3. ASSEMBLE

4. REDUNDANCY

5. ANNOTATE

in

PIPELINE

RNA-seq
(100 bp - PE)

1. CLEANUP (

RNA-seq
(100 bp - PE)

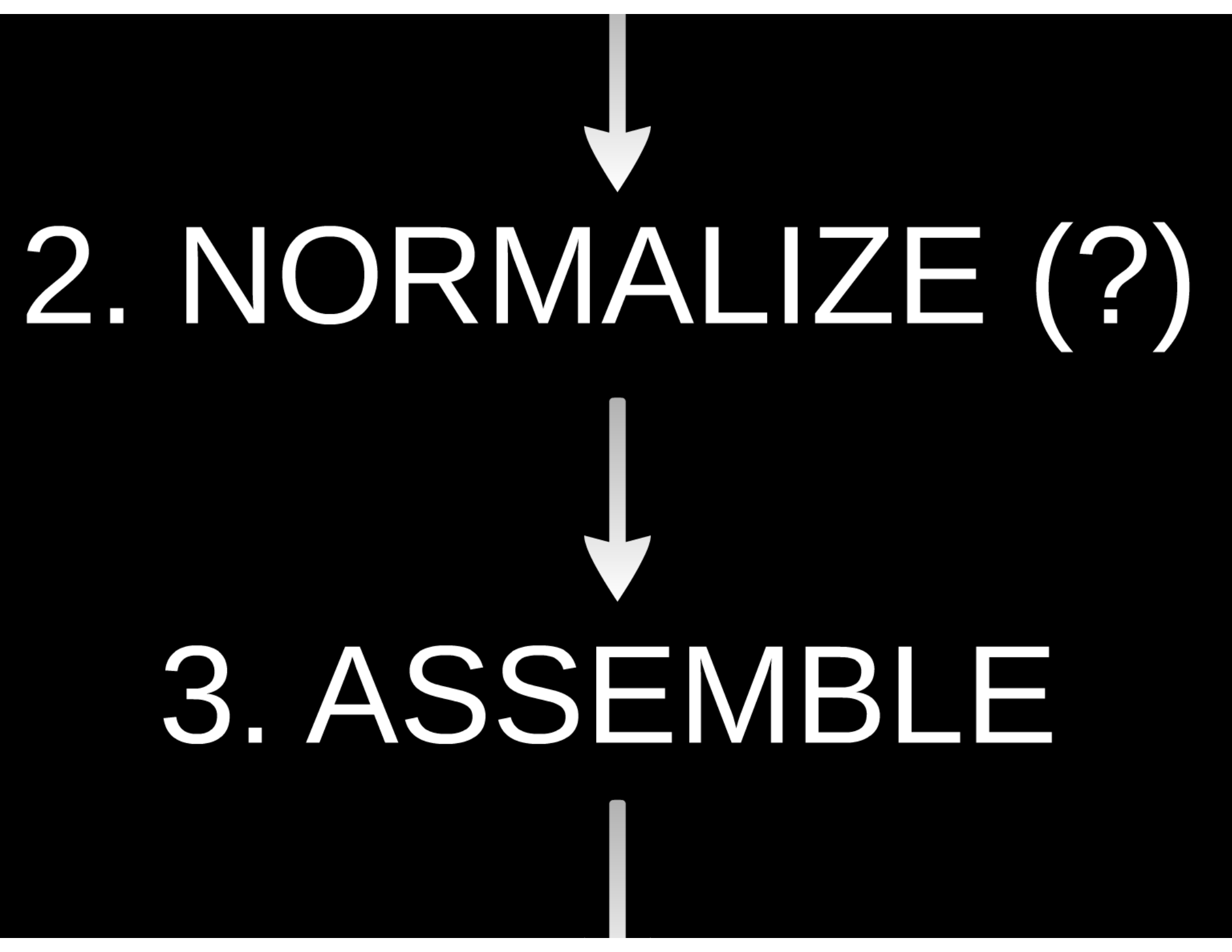
1. CLEANUP (?)

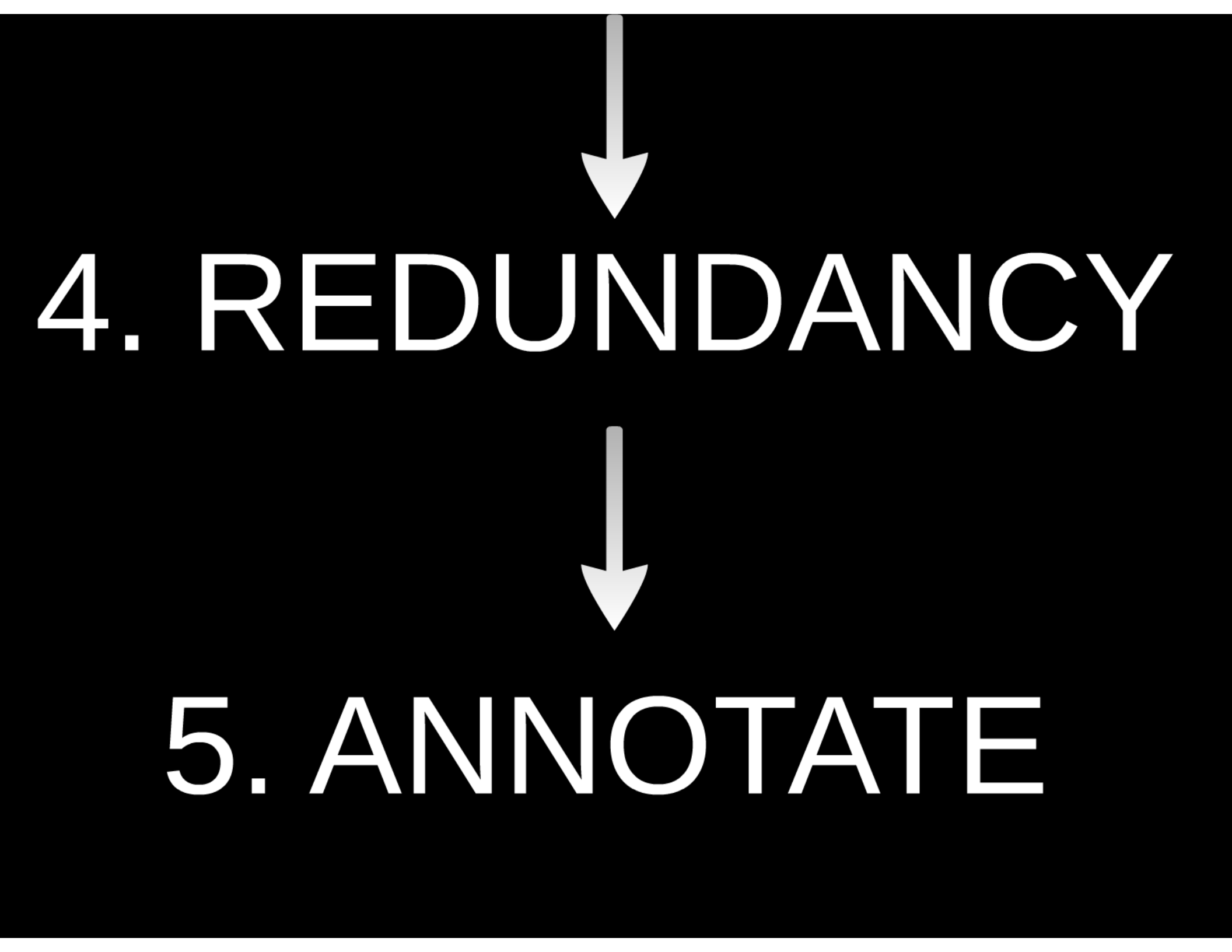
Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

RNA-seq
(100 bp - PE)

1. CLEANUP (?)


2. NORMALIZE (?)
3. ASSEMBLE

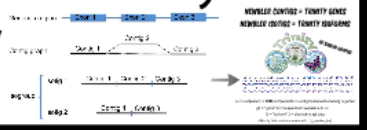

4. REDUNDANCY
5. ANNOTATE

O'Neil and Emrich *BMC Genomics* 2013, **14**:465
http://www.biomedcentral.com/1471-2164/14/465

BMC Genomics

RESEARCH ARTICLE Open Access

Assessing *De Novo* transcriptome assembly metrics for consistency and utility



Shawn T O'Neil[1,2] and Scott J Emrich[2*]

Table 1 Metric trends and consistency

Metric	Trend by sequencing depth	Consistent over sequencing depths	Trend by read length	Consistent over read lengths	Fully consistent metric
Contig count	↗	✓	↘	✓	✗
% of reads used in contigs	↗	✓	↗	✓	✓
BP in contigs	↗	✓	↗	✓	✓
% BP in contigs	↗	✓	↗	✓	✓
Average contig coverage	↗	✗	↗	✗	✗
Average unigene coverage	↗	✓	↗	✓	✓
Contig read count COV	↝	✓	↘	✓	✗
Unigene read count COV	↝	✓	↘	✓	✗
Average contig length	↗	✗	↗	✗	✗
Average unigene length	↗	✓	↗	✓	✓
Contig N50 length	↗	✗	↗	✗	✗
Unigene N50 length	↗	✓	↗	✓	✓
Unique annotations in singletons	↝	✓	↘	✓	✗
Unique annotations in contigs	↗	✓	↗	✓	✓
Unique annotations in unigenes	↗	✗	↝	✓	✗
Average contig OHR	↗	✗	↗	✗	✗
Average unigene OHR	↗	✓	↗	✓	✓
Contig RBH count	↗	✓	↗	✓	✓
Unigene RBH count	↗	✓	↗	✓	✓
% of Annotated contigs with RBHs	↝	✗	↗	✗	✗
% of Annotated unigenes with RBHs	↝	✓	↗	✗	✗
Average contig CF	↘	✓	↘	✓	✓
Average unigene CF	↘	✗	↘	✗	✗
Unique reverse annotations in contigs	↗	✓	↗	✓	✓
Unique reverse annotations in unigenes	↗	✓	↗	✓	✓

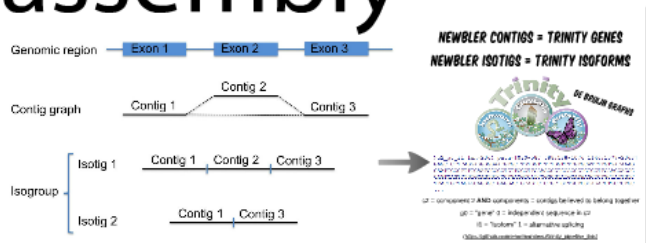
PLUS

O'Neil and Emrich *BMC Genomics* 2013, **14**:465
http://www.biomedcentral.com/1471-2164/14/465

BMC Genomics

RESEARCH ARTICLE Open Access

Assessing *De Novo* transcriptome assembly metrics for consistency and utility

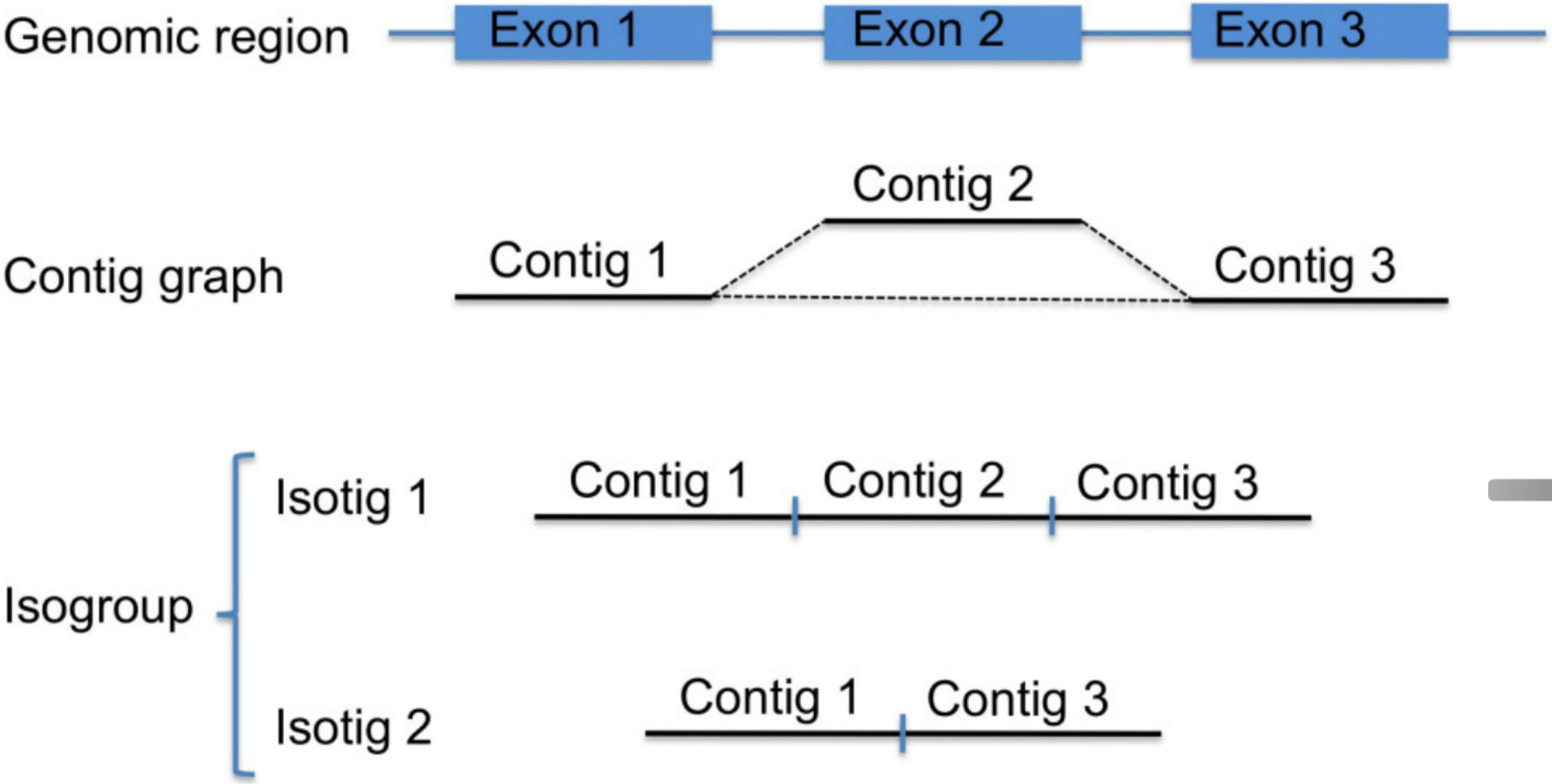


Shawn T O'Neil[1,2] and Scott J Emrich[2*]

Metric trends and consistency

Metric	Trend by sequencing depth	Consistent over sequencing depths	Trend by read length	Consistent over read lengths	Fully con met
Contig count	↗	✓	↘	✓	✗
of reads used in contigs = # BASES IN "GENES" / # TOTAL ASSEMBLED BASES	↗	✓	↗	✓	✓

assembly


Genomic region
Exon 1
Exon 2
Exon 3
Contig graph
Contig 2
Contig 1
Contig 3
Isogroup
Isotig 1
Contig 1
Contig 2
Contig 3
Isotig 2
Contig 1
Contig 3

DE BRUIJN GRAPHS

```
>c2_g0_i1 len=2364 path=[0:0-587 588:588-1076 1146:1077-2363]
GAGCTCTTCAGGAGGGGGAATGTGCTTGTGGTTTTTGGTCTTGTGCATTTTGTGACAAAG
GAATTCCCTTTTGAATCGCGCTGTTCCCTTGAAACCCTGGAGCCTCTGGTTCAAGCAGCG
CAGTCAGTCTGTGCAGTGTCCCTGACGTCATCCGGCGTATGCATAAGCTCTGCTATTGTC
TTACCGCTAGAGCAGGGCTGAGGACTGCAGTCTCTGCTGCTGCTCGCAGACCTGCCCTGC
...
```

c2 = component 2 **AND** components = contigs believed to belong together

g0 = "gene" 0 = independent sequence in c2

i1 = "isoform" 1 = alternative splicing

(https://github.com/enormandeau/trinity_pipeline_ibis)

NEWBLER CONTIGS = TRINITY GENES

NEWBLER ISOTIGS = TRINITY ISOFORMS

DE BRUIJN GRAPHS

```
>c2_g0_i1 len=2364 path=[0:0-587 588:588-1076 1146:1077-2363]
GAGCTCTTCAGGAGGGGGAATGTGCTTGTGGTTTTTGGTCTTGTGCATTTTGTGACAAAG
GAATTCCCTTTTGAATCGCGCTGTTCCCTTGAAACCCTGGAGCCTCTGGTTCAAGCAGCG
CAGTCAGTCTGTGCAGTGTCCCTGACGTCATCCGGCGTATGCATAAGCTCTGCTATTGTC
TTACCGCTAGAGCAGGGCTGAGGACTGCAGTCTCTGCTGCTGCTCGCAGACCTGCCCTGC
...
```

c2 = component 2 **AND** components = contigs believed to belong together

g0 = "gene" 0 = independent sequence in c2

i1 = "isoform" 1 = alternative splicing

(https://github.com/enormandeau/trinity_pipeline_ibis)

Table 1 Metric trends and consistency

Metric	Trend by sequencing depth	Consistent over sequencing depths	Trend by read length	Consistent over read lengths	Fully consistent metric
Contig count	↗	✓	↘	✓	✗
% of reads used in contigs = # BASES IN "GENES" / # TOTAL ASSEMBLED BASES	↗	✓	↗	✓	✓
BP in contigs	↗	✓	↗	✓	✓
% BP in contigs	↗	✓	↗	✓	✓
Average contig coverage	↗	✗	↗	✗	✗
Average unigene coverage	↗	✓	↗	✓	✓
Contig read count COV	↝	✓	↘	✓	✗
Unigene read count COV	↝	✓	↘	✓	✗
Average contig length	↗	✗	↗	✗	✗
Average unigene length	↗	✓	↗	✓	✓
Contig N50 length	↗	✗	↗	✗	✗
Unigene N50 length	↗	✓	↗	✓	✓
Unique annotations in singletons	↝	✓	↘	✓	✗
Unique annotations in contigs	↗	✓	↗	✓	✓
Unique annotations in unigenes	↗	✗	↝	✓	✗
Average contig OHR	↗	✗	↗	✗	✗
Average unigene OHR	↗	✓	↗	✓	✓
Contig RBH count	↗	✓	↗	✓	✓
Unigene RBH count	↗	✓	↗	✓	✓
% of Annotated contigs with RBHs	↝	✗	↗	✗	✗
% of Annotated unigenes with RBHs	↝	✓	↗	✗	✗
Average contig CF	↘	✓	↘	✓	✓
Average unigene CF	↘	✗	↘	✗	✗
Unique reverse annotations in contigs	↗	✓	↗	✓	✓
Unique reverse annotations in unigenes	↗	✓	↗	✓	✓

Contig count

% of reads used in contigs

BP in contigs

% BP in contigs

$$= \frac{\text{\# BASES IN "GENES"}}{\text{\# TOTAL ASSEMBLED BASES}}$$

Contig count

% of reads used in contigs

BP in contigs

% BP in contigs

Average contig coverage

Average unigene coverage

Contig read count COV

Unigene read count COV

Average contig length

Average unigene length

Contig N50 length

Unigene N50 length

= LENGTH AT WI
THIS LENGTH
CONTAIN AT LEA
ASSEMBL

ique annotations in singleto

= LENGTH AT WHICH CONTIGS OF THIS LENGTH AND MORE CONTAIN AT LEAST 50% OF THE ASSEMBLED BASES.

Average contig length

Average unigene length

Contig N50 length

Unigene N50 length

= LENGTH AT WI
THIS LENGTH
CONTAIN AT LEA
ASSEMBL

ique annotations in singleto

Unique annotations in singletons

Unique annotations in contigs

Unique annotations in unigenes

Average contig OHR

Average unigene OHR

Contig RBH count

ORTHOLOGOUS HIT RATIO (OHR)

$$\frac{\#\ \text{BASES IN ALIGNED REGION}}{\text{LENGTH OF BEST HIT}}$$

(TRANSCRIPT COMPLETENESS)

Table 1 Metric trends and consistency

Metric	Trend by sequencing depth	Consistent over sequencing depths	Trend by read length	Consistent over read lengths	Fully consistent metric
Contig count	↗	✓	↘	✓	✗
% of reads used in contigs = # BASES IN "GENES" / # TOTAL ASSEMBLED BASES	↗	✓	↗	✓	✓
BP in contigs	↗	✓	↗	✓	✓
% BP in contigs	↗	✓	↗	✓	✓
Average contig coverage	↗	✗	↗	✗	✗
Average unigene coverage	↗	✓	↗	✓	✓
Contig read count COV	↝	✓	↘	✓	✗
Unigene read count COV	↝	✓	↘	✓	✗
Average contig length	↗	✗	↗	✗	✗
Average unigene length	↗	✓	↗	✓	✓
Contig N50 length	↗	✗	↗	✗	✗
Unigene N50 length	↗	✓	↗	✓	✓
Unique annotations in singletons	↝	✓	↘	✓	✗
Unique annotations in contigs	↗	✓	↗	✓	✓
Unique annotations in unigenes	↗	✗	↝	✓	✗
Average contig OHR	↗	✗	↗	✗	✗
Average unigene OHR	↗	✓	↗	✓	✓
Contig RBH count	↗	✓	↗	✓	✓
Unigene RBH count	↗	✓	↗	✓	✓
% of Annotated contigs with RBHs	↝	✗	↗	✗	✗
% of Annotated unigenes with RBHs	↝	✓	↗	✗	✗
Average contig CF	↘	✓	↘	✓	✓
Average unigene CF	↘	✗	↘	✗	✗
Unique reverse annotations in contigs	↗	✓	↗	✓	✓
Unique reverse annotations in unigenes	↗	✓	↗	✓	✓

Li *et al. Genome Biology* (2014) 15:553
DOI 10.1186/s13059-014-0553-5

Genome Biology

METHOD Open Access

Evaluation of *de novo* transcriptome assemblies from RNA-Seq data

Bo Li[1†], Nathanael Fillmore[2†], Yongsheng Bai[3], Mike Collins[4], James A Thomson[4,5,6], Ron Stewart[4] and Colin N Dewey[2,7*]

http://deweylab.biostat.wisc.edu/detonate/rsem-eval.html

"reference-free de novo transcriptome assembly evaluator"

RESULTS SO FAR...

Assembly number	Trim quality (PHRED)	Normalized	Number of transcripts	Number of genes	unigene N50
1	20	Y	169 089	129 917	1655
2	10	Y	168 649	129 956	1640
3	5	Y	168 702	129 894	1664
4	2	Y	168 797	129 739	1660
5	20	N	174 531	144 326	1435
6	10	N	173 373	145 460	1425
7	5	N	175 339	145 410	1426
8	2	N	175 279	145 407	1427

reads = 19 370 403

assembled bases = 4 682 256 300