Annotation of genomes and transcriptomes

Torsten Struck – t.h.struck@nhm.uio.no
NHM UiO

Annotation is not like annotation

Structural annotation:

The process of identifying genes, their intron-exon structures and for transcriptomes their isoforms. Recently also ncRNA like tRNAs, µRNAs, retroposons, line elements and so forth.

Annotation is not like annotation

Structural annotation:

The process of identifying genes, their intron-exon structures and for transcriptomes their isoforms. Recently also ncRNA like tRNAs, µRNAs, retroposons, line elements and so forth.

Functional annotation:

The process of attaching meta-data such as gene names, gene family, gene ontology terms to structural annotations.

Annotation is not like annotation

Structural annotation:
The process of identifying genes, their intron-exon structures and for transcriptomes their isoforms. Recently also ncRNA like tRNAs, µRNAs, Retroposons, Line elements and so forth.

Functional annotation:
The process of attaching meta-data such as gene names, gene family, gene ontology terms to structural annotations.

Only minimaly necessary for phylogenomics (as off here)
Orthology determination

Most important

Most important

Most important

Quality of assembly

Most important

Quality of assembly
Depends on purpose

Quality parameters

contigs & scaffolds
total bp
shortest & longest contigs | scaffolds
average length of contigs | scaffolds
GC%

Quality parameters

contigs & scaffolds
total bp
shortest & longest contigs | scaffolds
average length of contigs | scaffolds
GC%

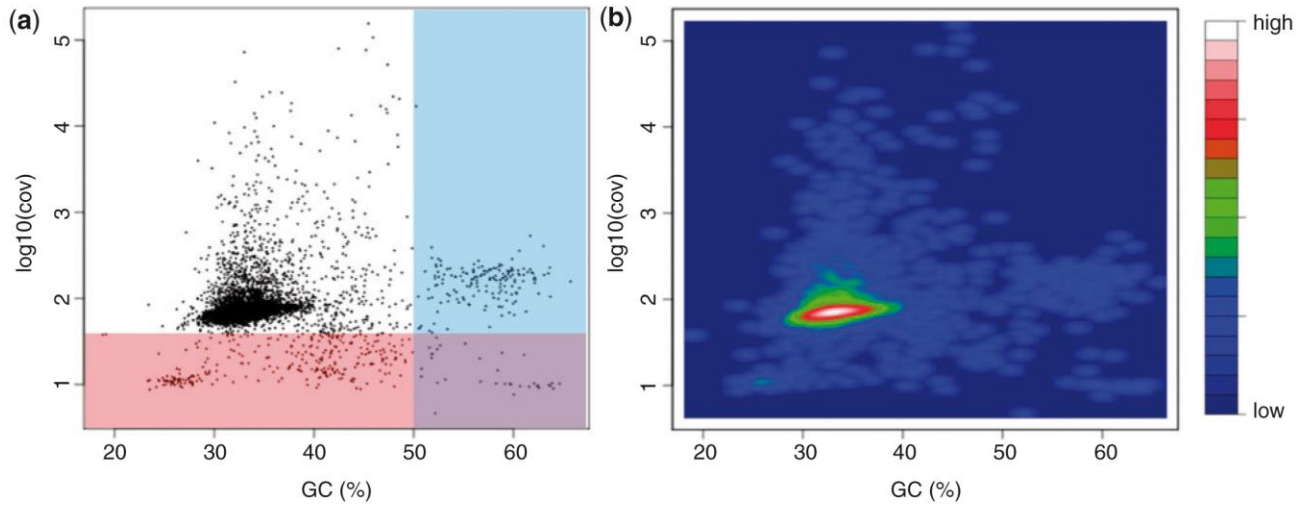


FIG. 1.—Aggregate properties (coverage and GC content) of contigs obtained by Celera 7.0 as (*a*) scatter plot and (*b*) heat map. Colored areas in (*a*) illustrate the aggressive cleaning strategy adopted in this study, that is, the removal of putative nontarget contigs of coverage <40× (red; putative host contamination) and GC content >50% (blue; putative bacterial contamination) (see supplementary file S1, Section A for details, Supplementary Material online).

Hahn et al. (2014) Genome Biology & Evolution

Quality parameters

contigs & scaffolds
total bp
shortest & longest contigs | scaffolds
average length of contigs | scaffolds
GC%

or % non–ACTG
% gaps in scaffolds

Quality parameters

contigs & scaffolds
total bp
shortest & longest contigs | scaffolds
average length of contigs | scaffolds
GC%

or % non–ACTG
% gaps in scaffolds

coverage → total sequenced bp/known genome size
genome coverage → total bp/known genome size
gene coverage → # genes found/# genes tested

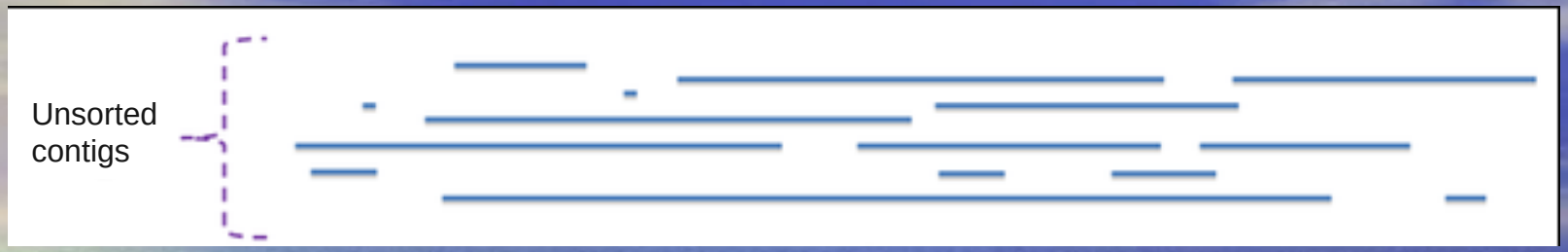
Quality parameters

contigs & scaffolds
total bp
shortest & longest contigs | scaffolds
average length of contigs | scaffolds
GC%

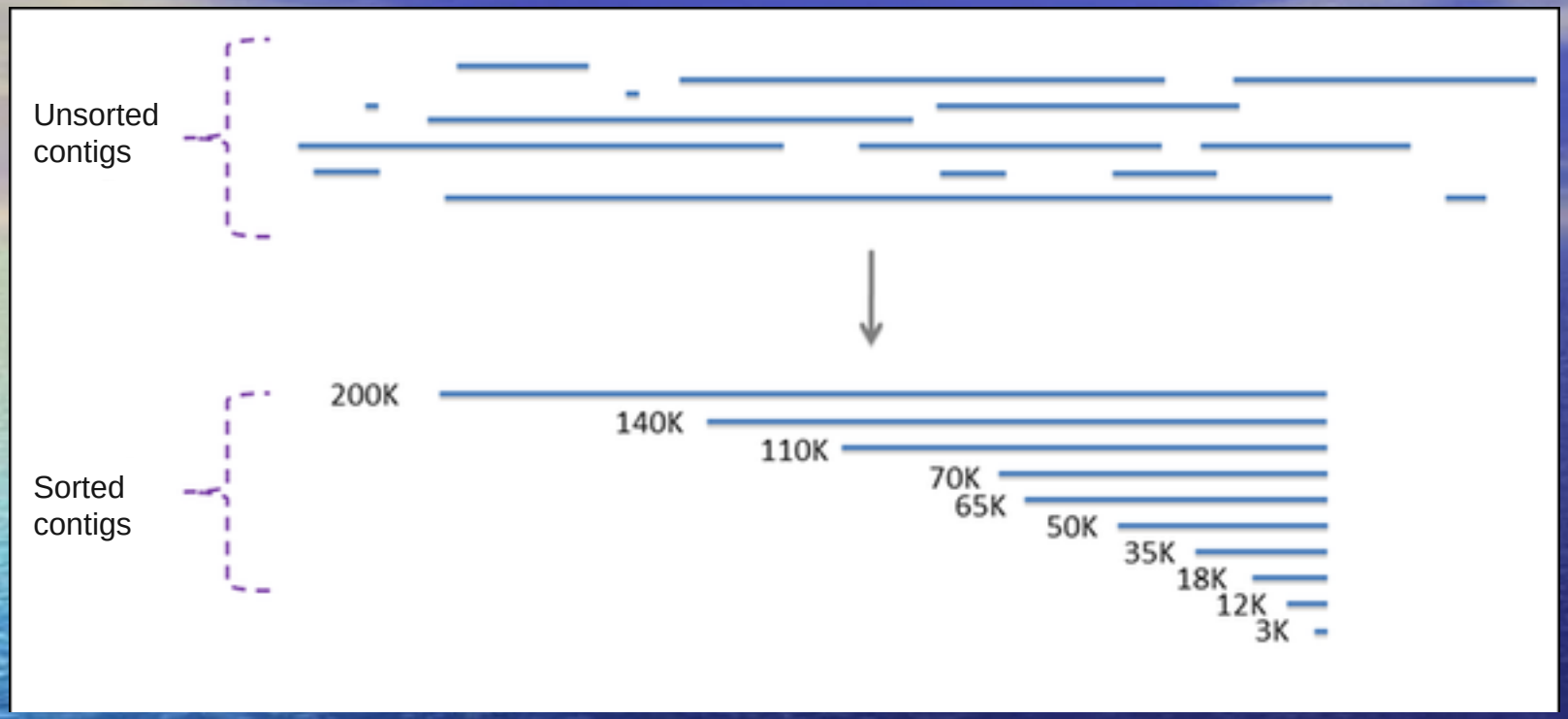
or % non–ACTG
% gaps in scaffolds

coverage → total sequenced bp/known genome size
genome coverage → total bp/known genome size
gene coverage → # genes found/# genes tested
CEGMA → screening set of universal, single-copy genes

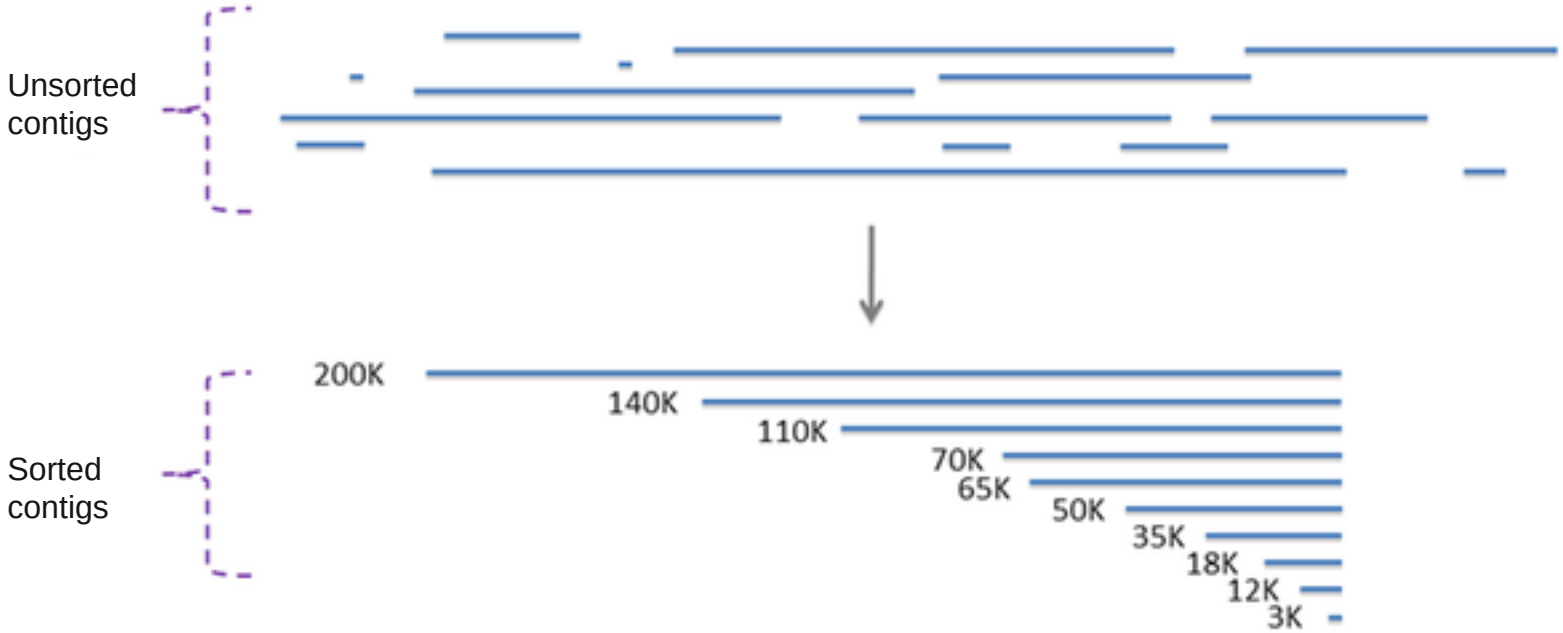
N50 & L50



N50 & L50


Unsorted contigs
Sorted contigs
200K
140K
110K
70K
65K
50K
35K
18K
12K
3K

N50 & L50

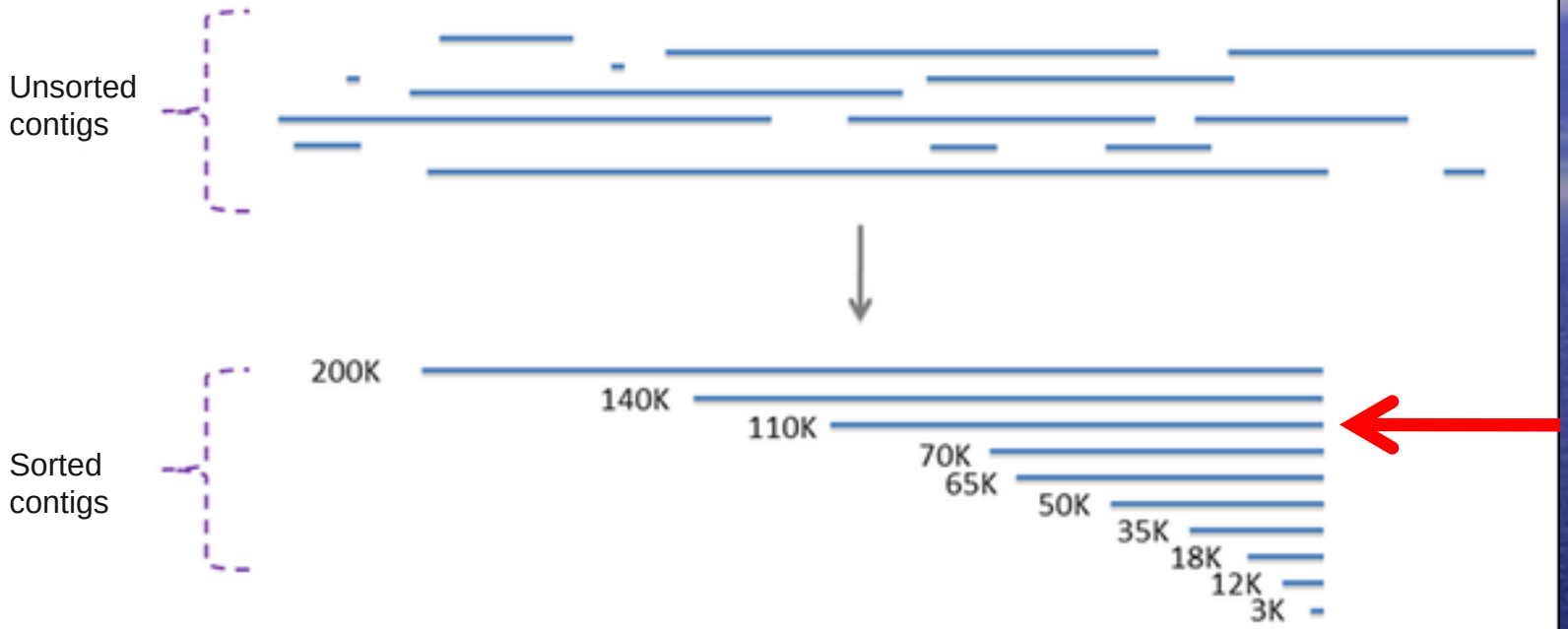


Total contig length= 200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K= 703K

50% total contig length= 703K x 50%= 351.5K

200K + 140K + **110K** > 351.5K → N50= 110K

N50 & L50

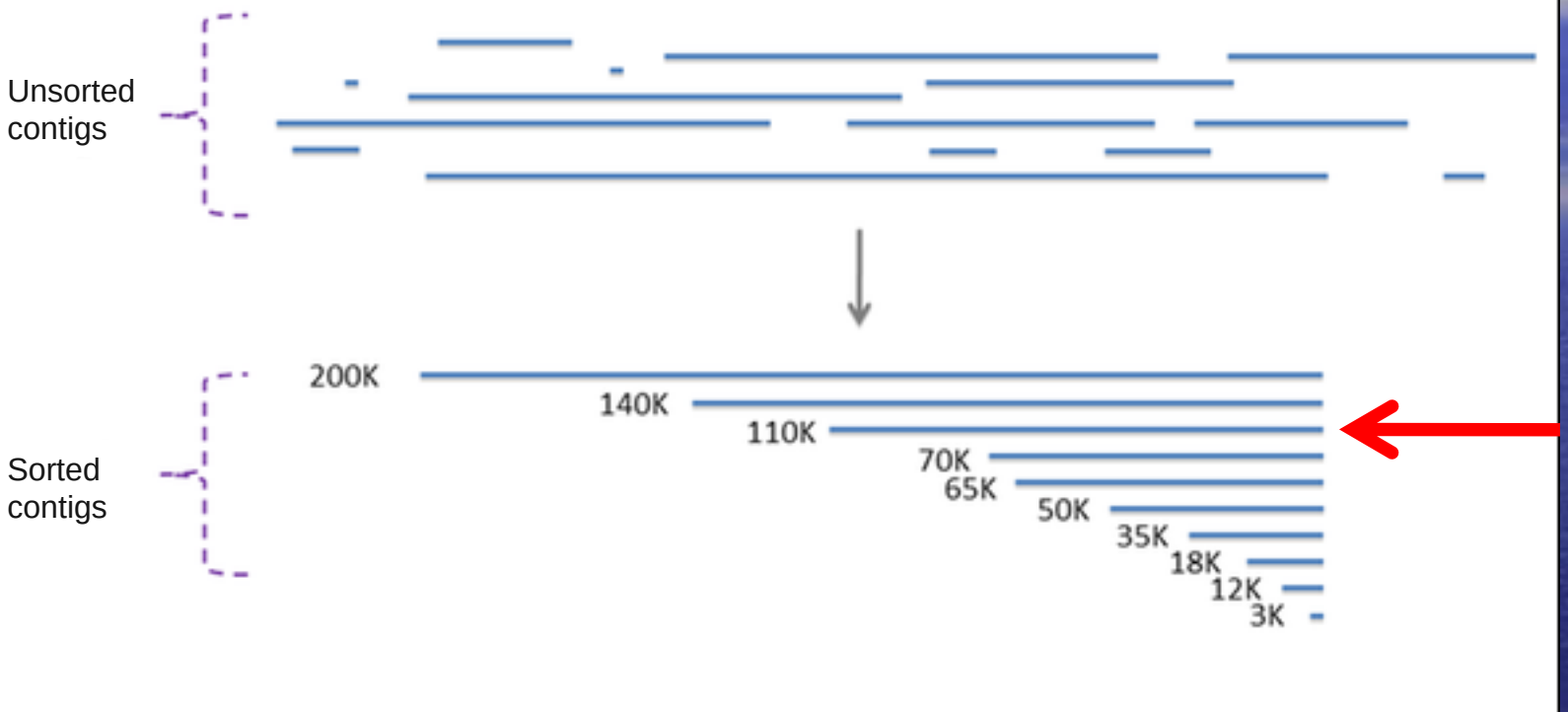


Total contig length= 200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K= 703K

50% total contig length= 703K x 50%= 351.5K

200K + 140K + **110K** > 351.5K → N50= 110K

N50 & L50



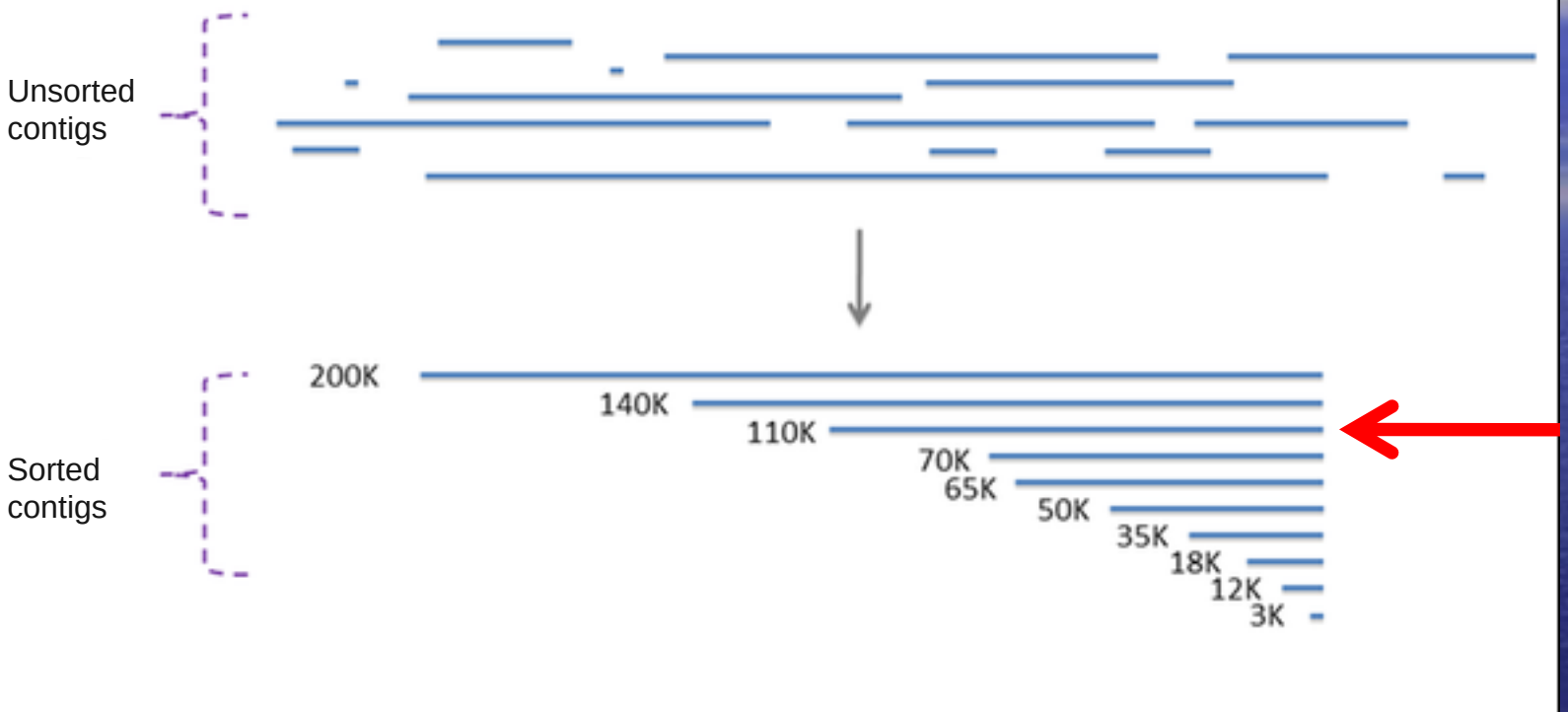
Total contig length= 200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K= 703K

50% total contig length= 703K x 50%= 351.5K

200K + 140K + **110K** > 351.5K → N50= 110K

L50 = # contigs accounting for more than 50% of the assembly

N50 & L50



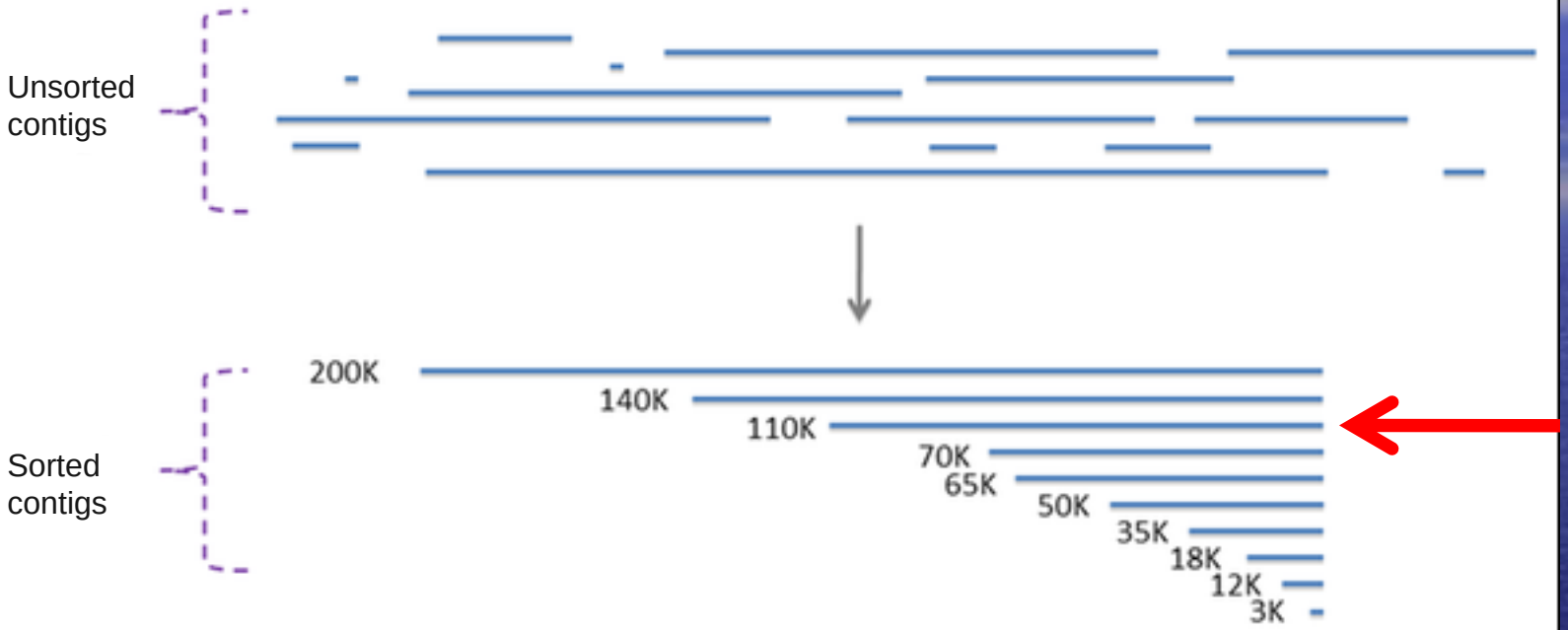
Total contig length= 200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K= 703K

50% total contig length= 703K x 50%= 351.5K

200K + 140K + **110K** > 351.5K → N50= 110K

L50 = # contigs accounting for more than 50% of the assembly
L50 = 3

N50 & L50



Total contig length= 200K + 140K + 110K + 70K + 65K + 50K + 35K + 18K + 12K + 3K= 703K

50% total contig length= 703K x 50%= 351.5K

200K + 140K + **110K** > 351.5K → N50= 110K

L50 = # contigs accounting for more than 50% of the assembly
L50 = 3
Same principals apply to scaffolds

While there are correlations in assembly statistics...

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	↗	✗	↗	✗	✗

O'Neil & Emrich (2013) *BMC Genomics*

Metrics are considered consistent if they consistently ranked perfectly assemblies as better; fully consistent metrics are consistent metrics with similar monotonic trends by sequencing depth and read length for perfect assemblies.

…what represents a "good" assembly is debatable

(what is "truth" when it's unknown to start with?)

Hunt *et al. Genome Biology* 2013, **14**:R47
http://genomebiology.com/2013/14/5/R47

Genome Biology

SOFTWARE — Open Access

REAPR: a universal tool for genome assembly evaluation

Martin Hunt[1], Taisei Kikuchi[1,2], Mandy Sanders[1], Chris Newbol

BIOINFORMATICS APPLICATIONS NOTE
Vol. 29 no. 8 2013, pages 1072–1075
doi:10.1093/bioinformatics/btt086

Genome analysis

Advance Access publication February 19, 2013

QUAST: quality assessment tool for genome assemblies

Alexey Gurevich[1,*], Vladislav Saveliev[1], Nikolay Vyahhi[1] and Glenn Tesler[2]
Academic University, Russian Academy of Sciences, St. Petersburg
s, University of California, San Diego, La Jolla, CA 92093-0112, USA

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*]

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

RESEARCH — Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

EBP standard based on vertebrate genomes

1) Species with sufficient DNA and tissue:
   Minimum reference standard of **6.C.Q40**
   - N50 contig > 1 Mb
   - N50 scaffolding = chromosomal scale
   - error rate < 1/10,000 (i.e., Q40)
   - Additional criteria:
     - < 5% false duplications
     - > 90% kmer completeness
     - > 90% sequence assigned to candidate chromosomal sequences
     - > 90% single copy conserved genes (e.g. BUSCO)
     - > 90% transcripts from the same organism mappable

2) Challenging species with limited DNA or material (<~100ng DNA)
   Minimum reference standard of **4.5.Q40**
   - N50 contig > 10kb
   - N50 scaffold > 100kb
   - error rate < 1/10,000 (i.e., Q40)

BLAST

Target specific contigs & “fish” them out

– BLAST and its variants (Altschul et al. (1990) J. Mol. Biol.)

– Cited 50K+ times

BLAST

Target specific contigs & “fish” them out

– BLAST and its variants (Altschul et al. (1990) J. Mol. Biol.)

– Cited 50K+ times

Query/Queries

– What is being compared to the DB sequences

- Nucleotides or amino acids coming from assemblies

Subject/Subjects

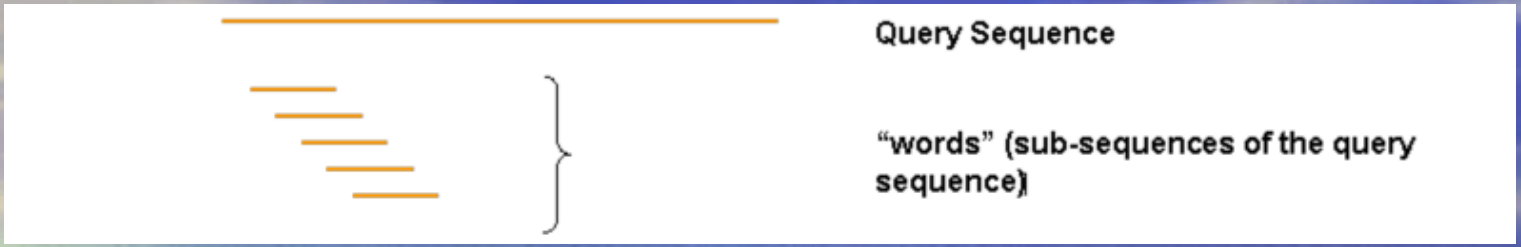
– DB sequences being compared to

- Nucleotides or amino acids potentially coming from other sources

BLAST – how it works

Query Sequence

BLAST – how it works



BLAST – how it works

The BLAST Search Algorithm

query word (*W* = 3)

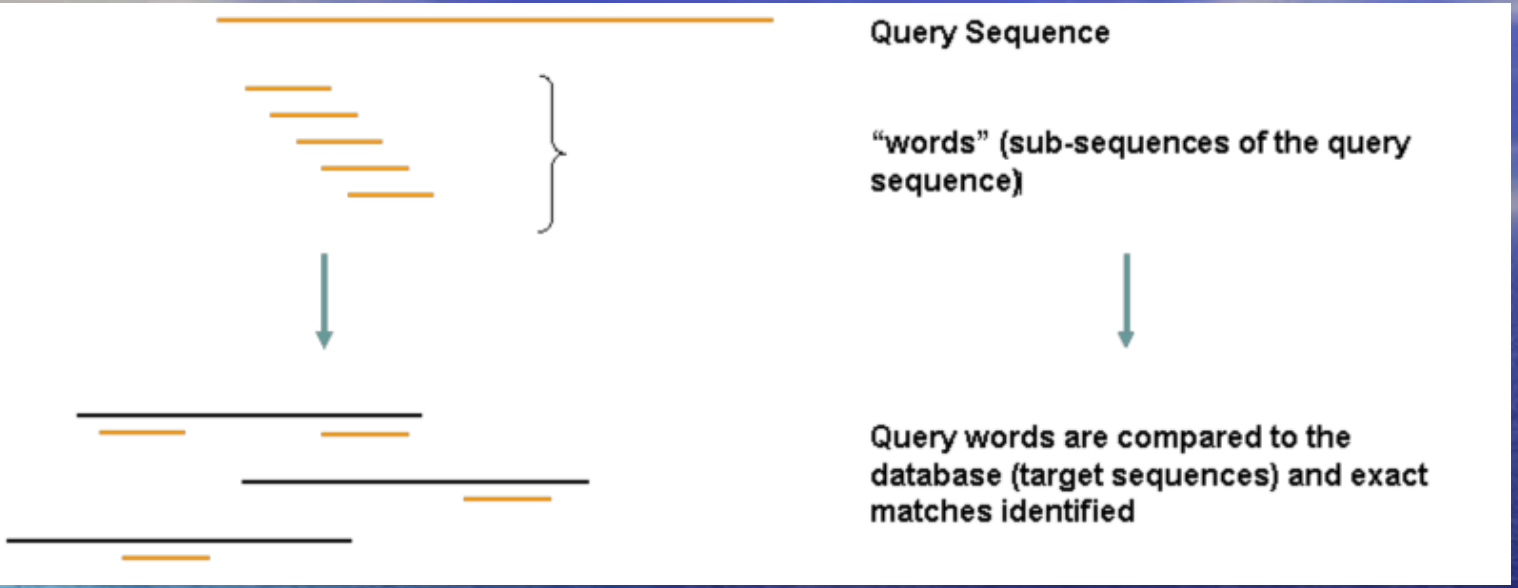
```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

neighborhood words

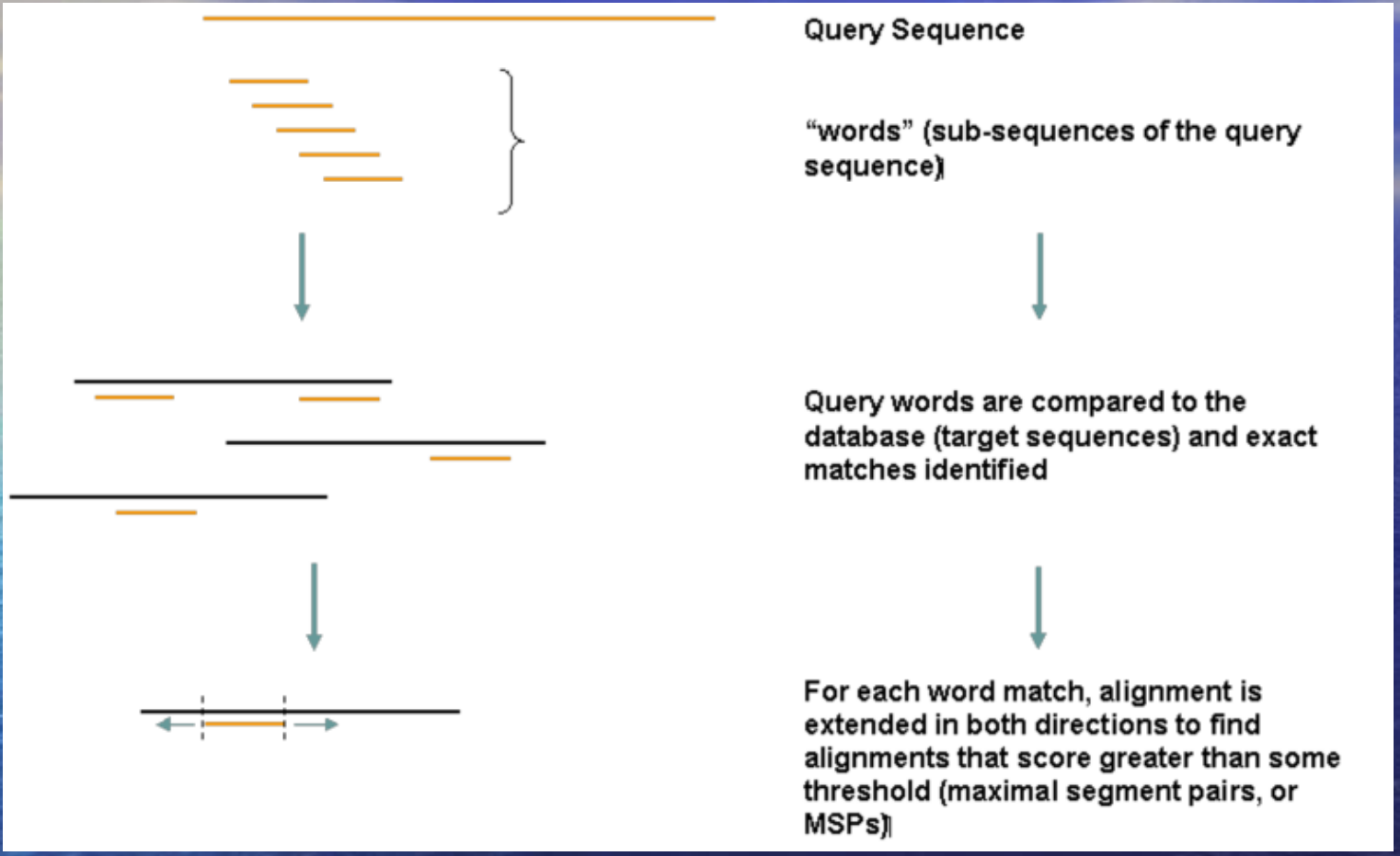
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

neighborhood score threshold (*T* = 13)

BLAST – how it works



BLAST – how it works



BLAST – how it works

The BLAST Search Algorithm

query word (*W* = 3)

```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

neighborhood score threshold (*T* = 13)

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

BLAST – on the command line

Tools in the BLAST+ CLI suite

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

BLAST – on the command line

Tools in the BLAST+ CLI suite

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

makeblastdb = to make BLAST DBs from subject sequences