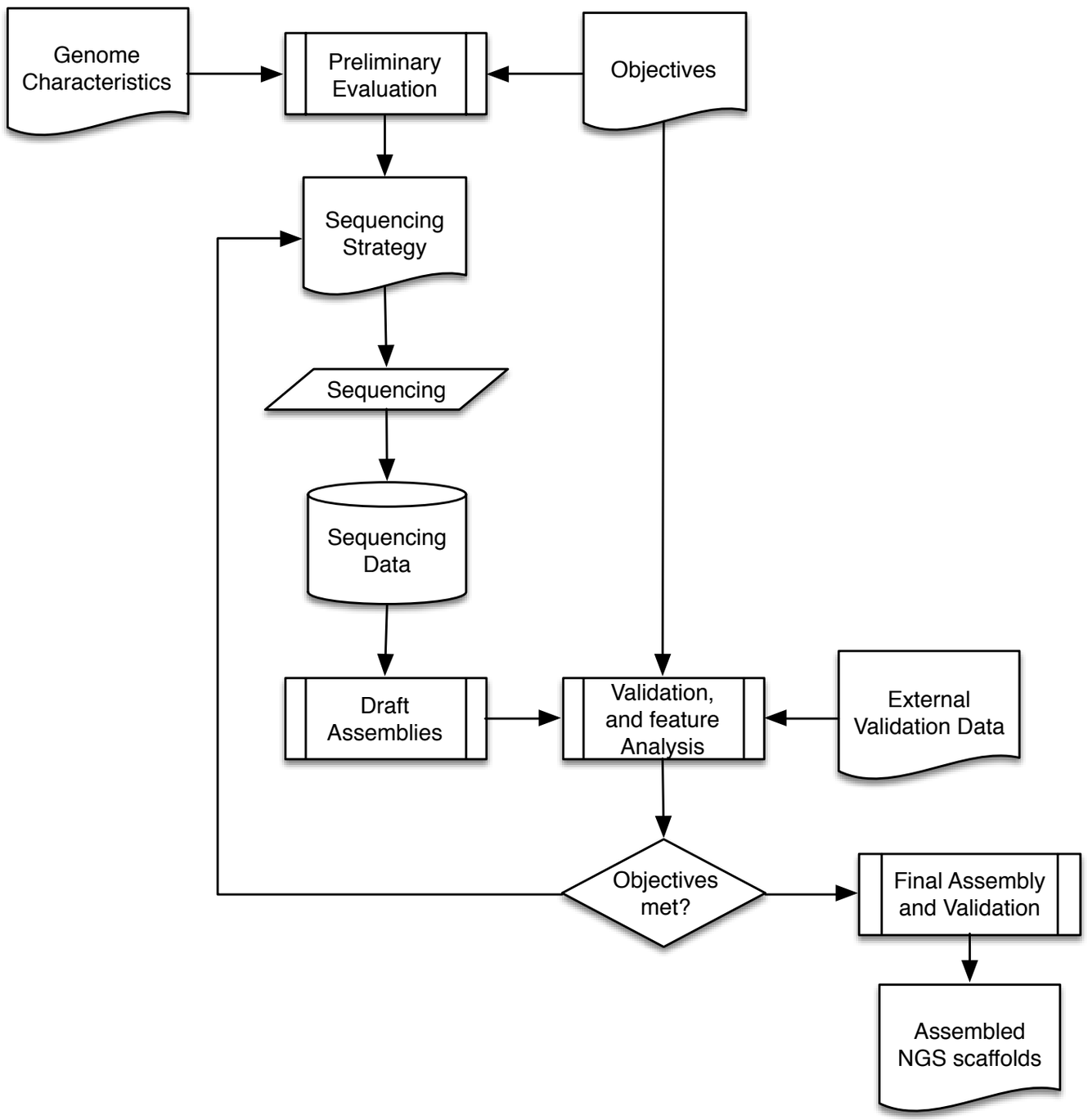
4 - Genome Assembly and Validation (Concepts)

Wednesday afternoon

Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia

TGAC

Assembly project workflow | Prior Knowledge



- Kariotype: Genome size, Ploidy
- Heterozygocity
- GC content
- Contaminants / Symbionts
- Data Sets:
  - Close relatives
  - Genes / ESTs / RNAseq / Markers
- Mithocondria
- Chloroplast

TGAC

Experiment design (you choose the data!)

- **Know your biological question.**

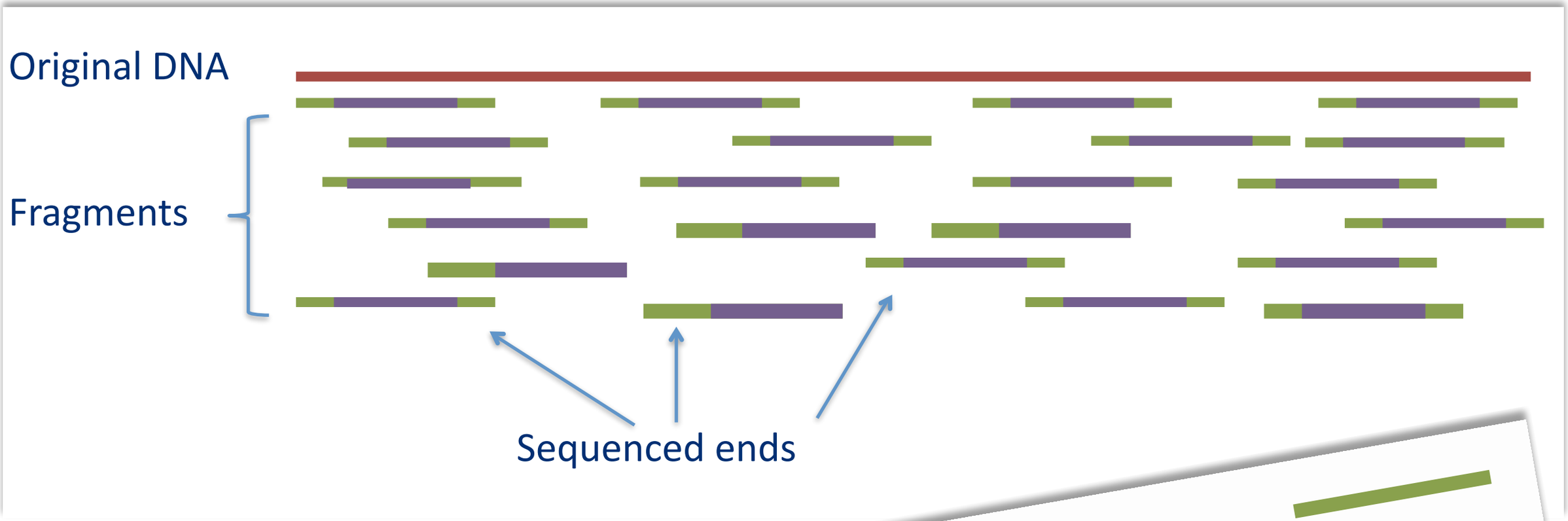
- Plan your data processing (from an information perspective).

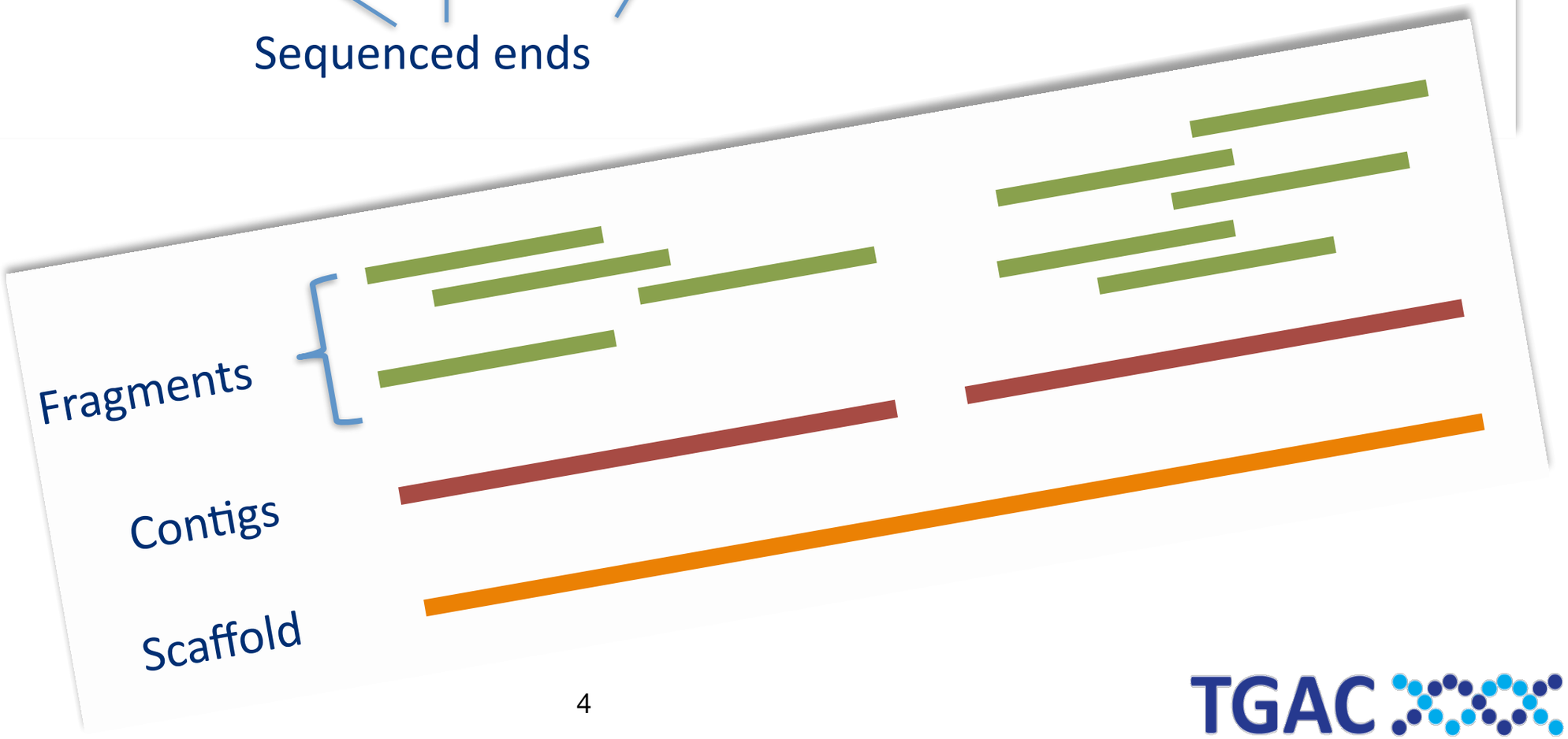
- Decide on conditions and biological/technical replicas.

- Decide on technologies and coverages:
  - How will the typical bias affect your experiment?
  - Is the coverage enough? Significant results?

TGAC

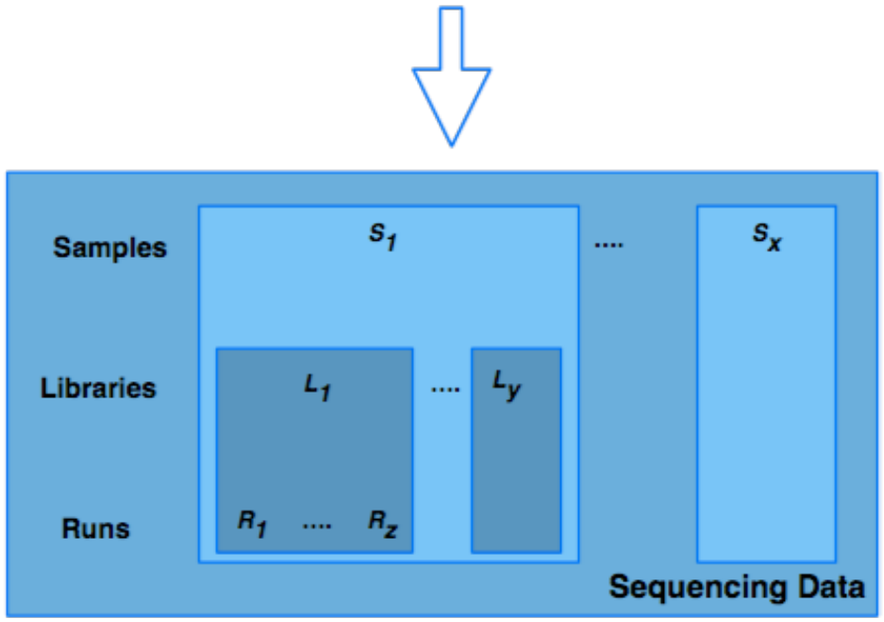
The genome assembly problem (WGS)





TGAC

Planning and "informed guesses"



Whole genome information (unknown?)
+Bias Information (unknown)

"known error profile"

short sequences ($10^{\{2-3\}}$)

large data (10^X)

$Y << X$

small data (10^Y)

long sequences (10^{10})

"unknown error profile"

Whole genome information (known?)
+Bias Information (unknown, reduced?)

Assemble and Scaffold

Scaffolds & Contigs

Validate and release

TGAC

The assembly is just a probabilistic model of a genome, condensing the information from the experimental evidence.

All the information is already present in the experimental results.

TGAC

A correct assembly has:

The right *motifs*,

the correct number of times,

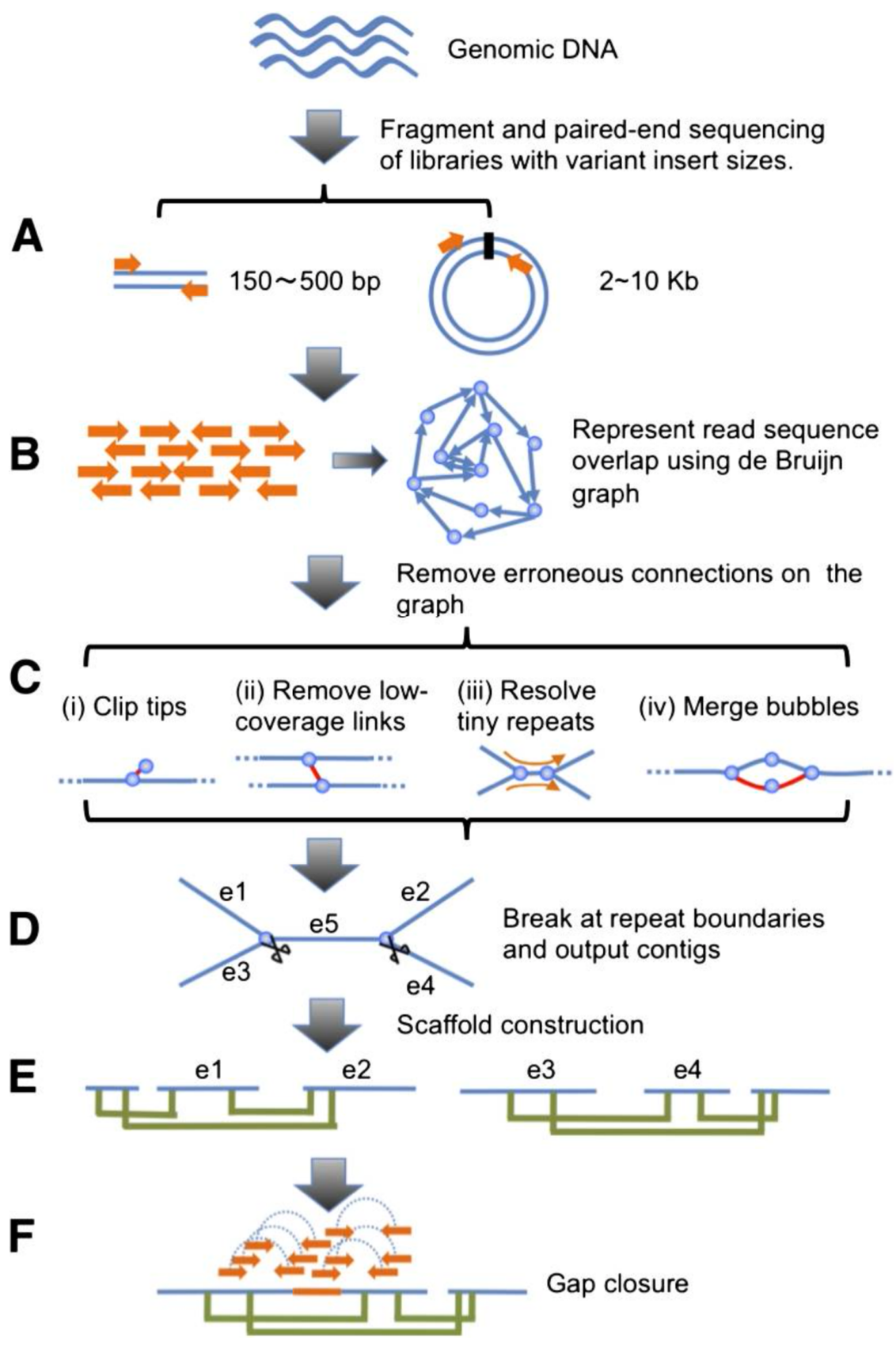
in correct order and position.

None of which is assessed by length stats.

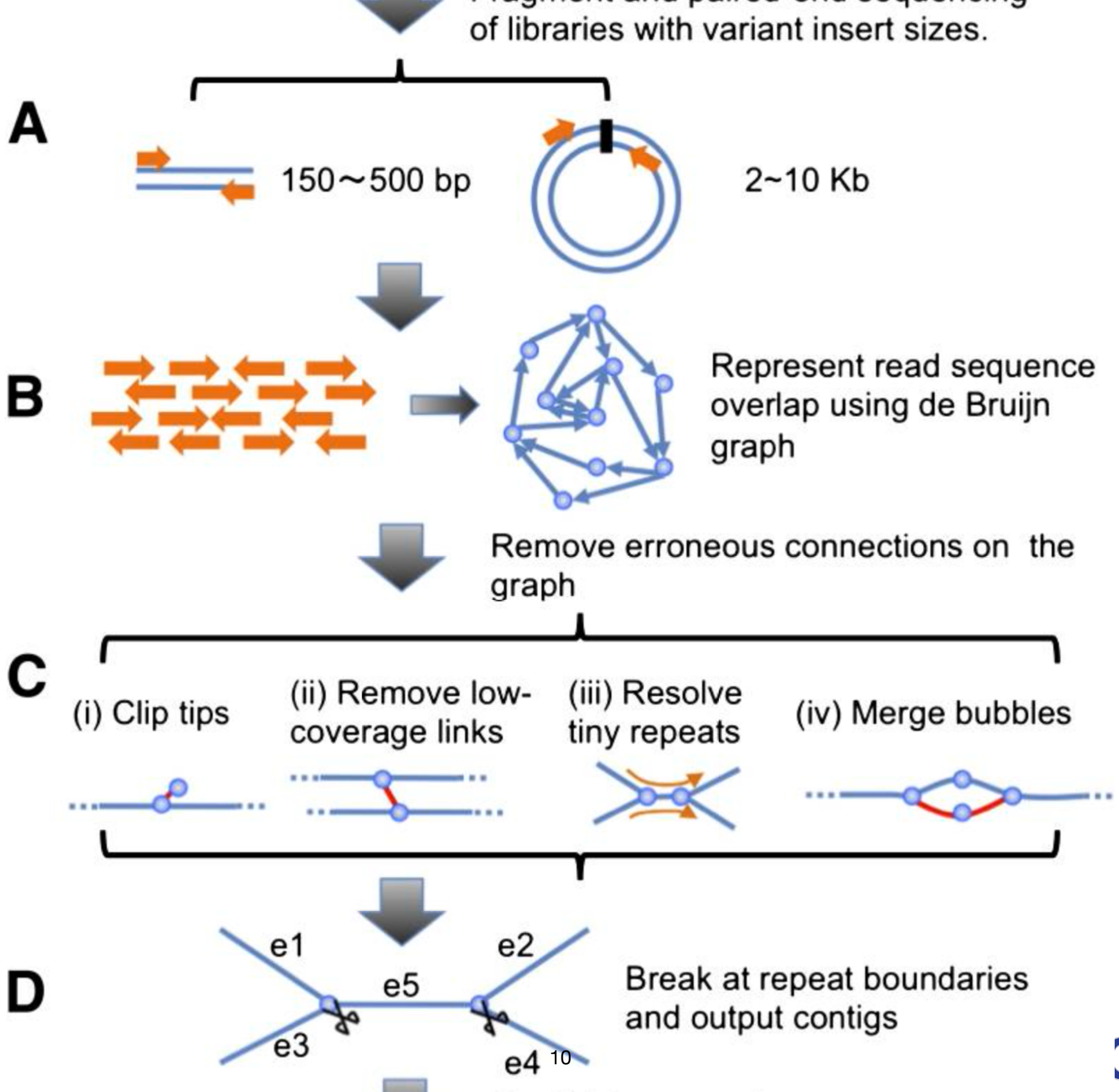
TGAC

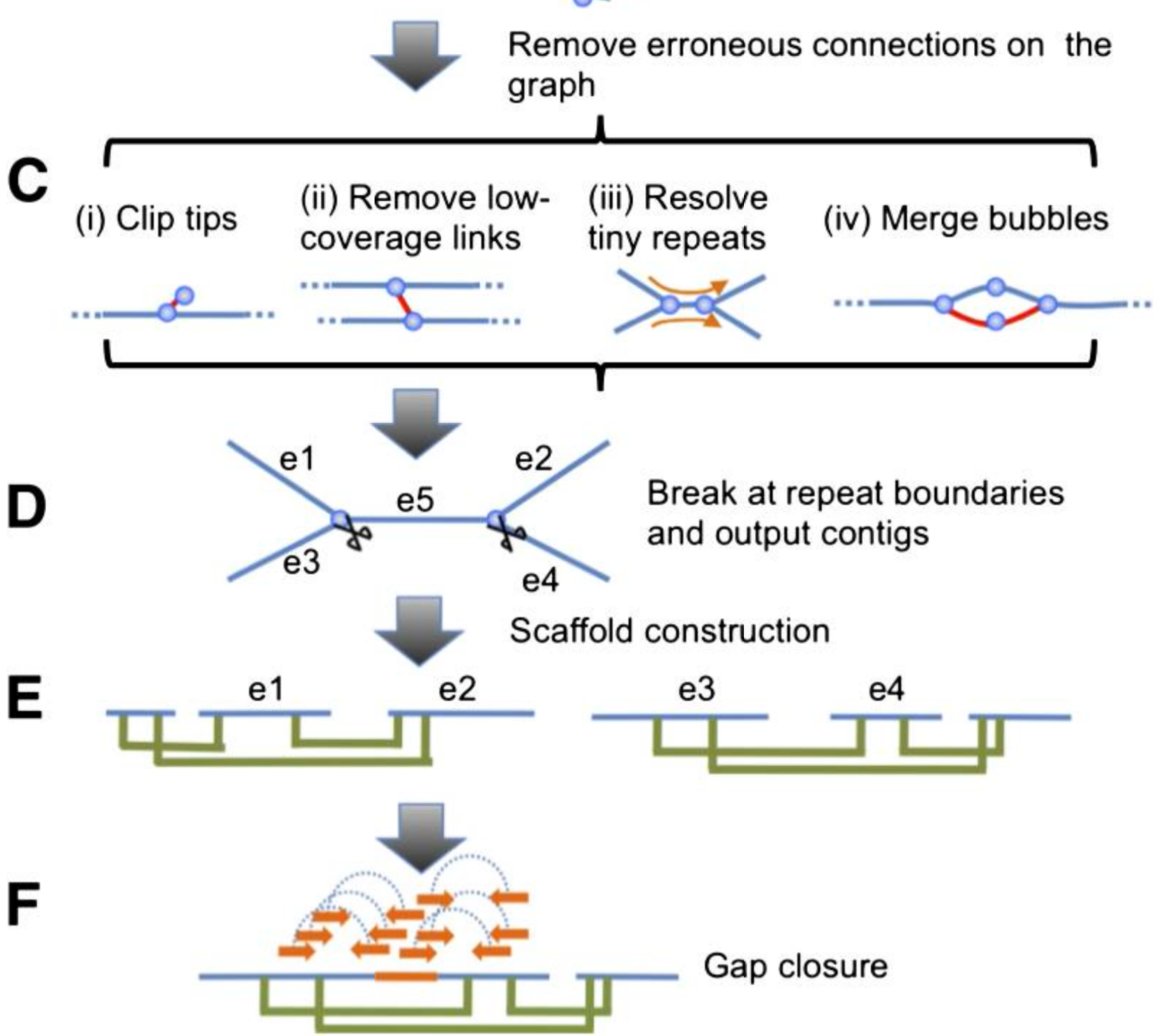
A modern assembler

Using SOAPdenovo2 as an example


Genomic DNA
Fragment and paired-end sequencing of libraries with variant insert sizes.
A
150~500 bp
2~10 Kb
B
Represent read sequence overlap using de Bruijn graph
Remove erroneous connections on the graph
C
(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles
D
e1
e2
e5
e3
e4
Break at repeat boundaries and output contigs
Scaffold construction
E
e1
e2
e3
e4
F
Gap closure

TGAC


of libraries with variant insert sizes.
A
150～500 bp
2~10 Kb
B
Represent read sequence overlap using de Bruijn graph
Remove erroneous connections on the graph
C
(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles
D
e1
e2
e5
e3
e4
Break at repeat boundaries and output contigs


Remove erroneous connections on the graph
C
(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles
D
e1
e2
e5
e3
e4
Break at repeat boundaries and output contigs
Scaffold construction
E
e1
e2
e3
e4
F
Gap closure

TGAC

Assembly validation

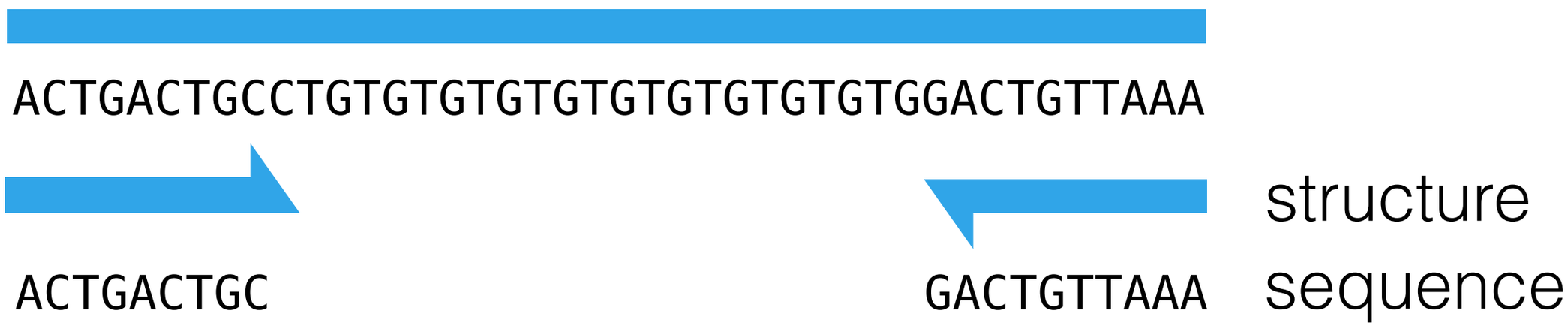
Using biological knowledge to figure out what are…

The right *motifs*,

the correct number of times,

in correct order and position.

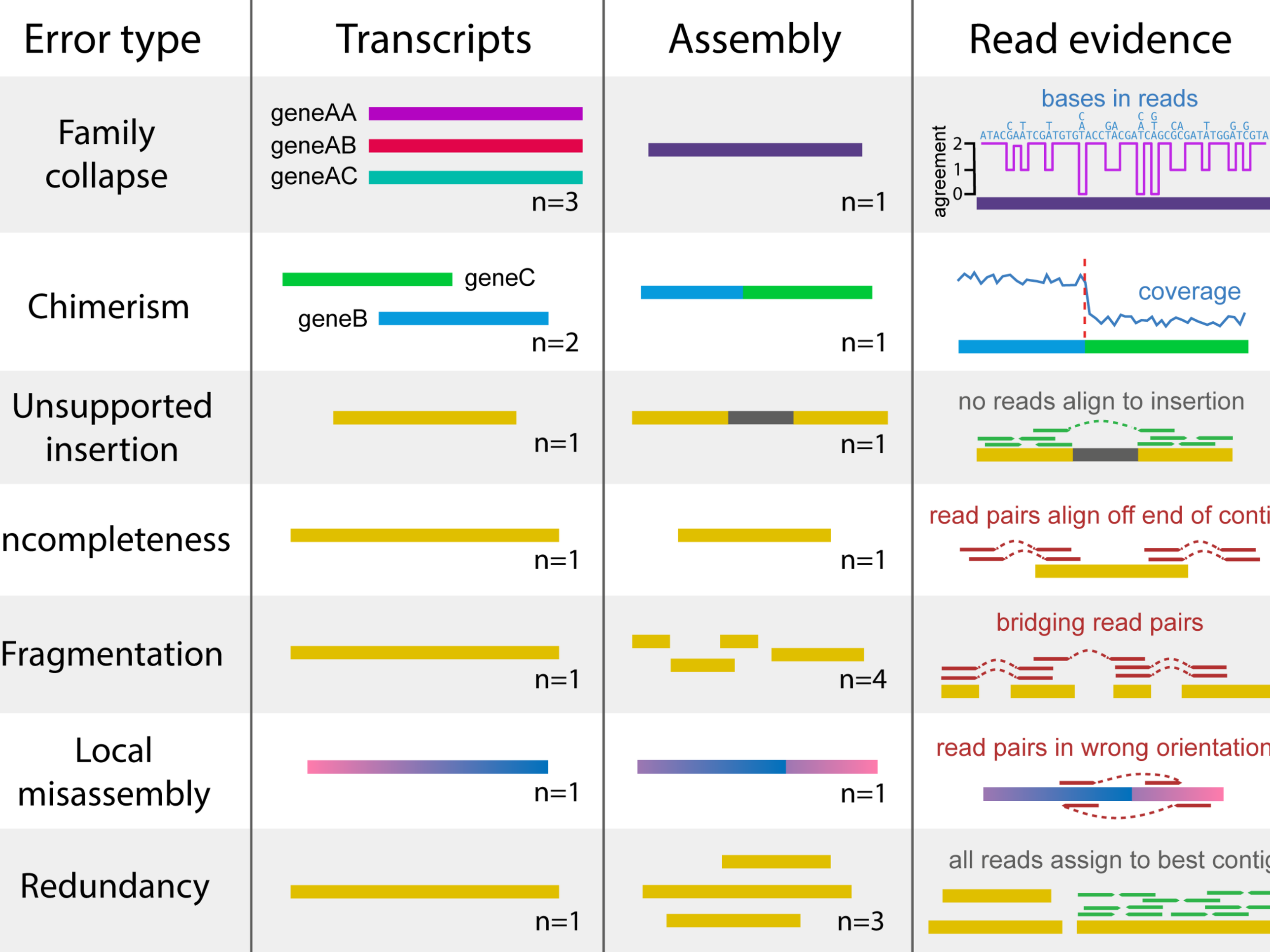
Direct experimental evidence: the reads



The right *motifs*,

the correct number of times,

in correct order and position.


Error type
Transcripts
Assembly
Read evidence
Family collapse
geneAA
geneAB
geneAC
n=3
n=1
bases in reads
ATACGAATCGATGTGTACCTACGATCAGCGCGATATGGATCGTA
agreement
2
1
0
Chimerism
geneC
geneB
n=2
n=1
coverage
Unsupported insertion
n=1
n=1
no reads align to insertion
ncompleteness
n=1
n=1
read pairs align off end of conti
Fragmentation
n=1
n=4
bridging read pairs
Local misassembly
n=1
n=1
read pairs in wrong orientation
Redundancy
n=1
n=3
all reads assign to best contig

Direct experimental evidence: other evidence

- Genome size, ploidy
- GC content
- Symbionts
- Plastids
- ESTs, cDNAs, peptides, genome walking

The right *motifs,*

the correct number of times,

in correct order and position.

Indirect experimental evidence: genomes in general

- Genes! They have structure
- Repeats
- Chromosome macrostructure
  - (circular?, number, telomeres, …)

The right *motifs*,

the correct number of times,

in correct order and position.

Indirect experimental evidence: other species

- Close relatives: proteins, transcripts, genomes
- Distant relatives: single-copy genes,
  phylogeny, HGT

The right *motifs*,

the correct number of times,

in correct order and position.

Questions?