Plant & Food RESEARCH

RANGAHAU AHUMĀRA KAI

Discover. Innovate. Grow.™

Hybrid Genome Assembly: A Practical Guide

Cecilia Deng, Senior Bioinformatician
邓泓

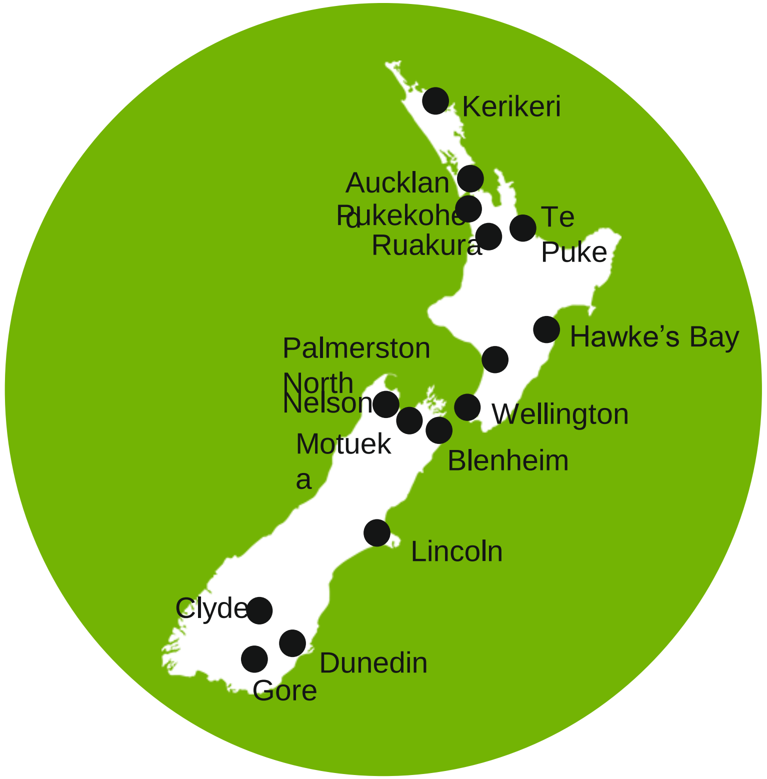
THE SCIENCE OF PREMIUM™

The New Zealand Institute for Plant & Food Research Limited

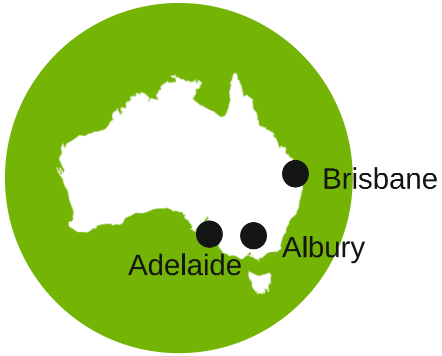
Overview

- Genome Assembly Project
- Sequencing Platforms
- Sequencing Considerations & Experimental Design
- Hybrid Genome Assembly Strategies
- A Complete Genome Assembly Workflow
- Conclusions
- Acknowledgements

Plant & Food Research (PFR)



NEW ZEALAND

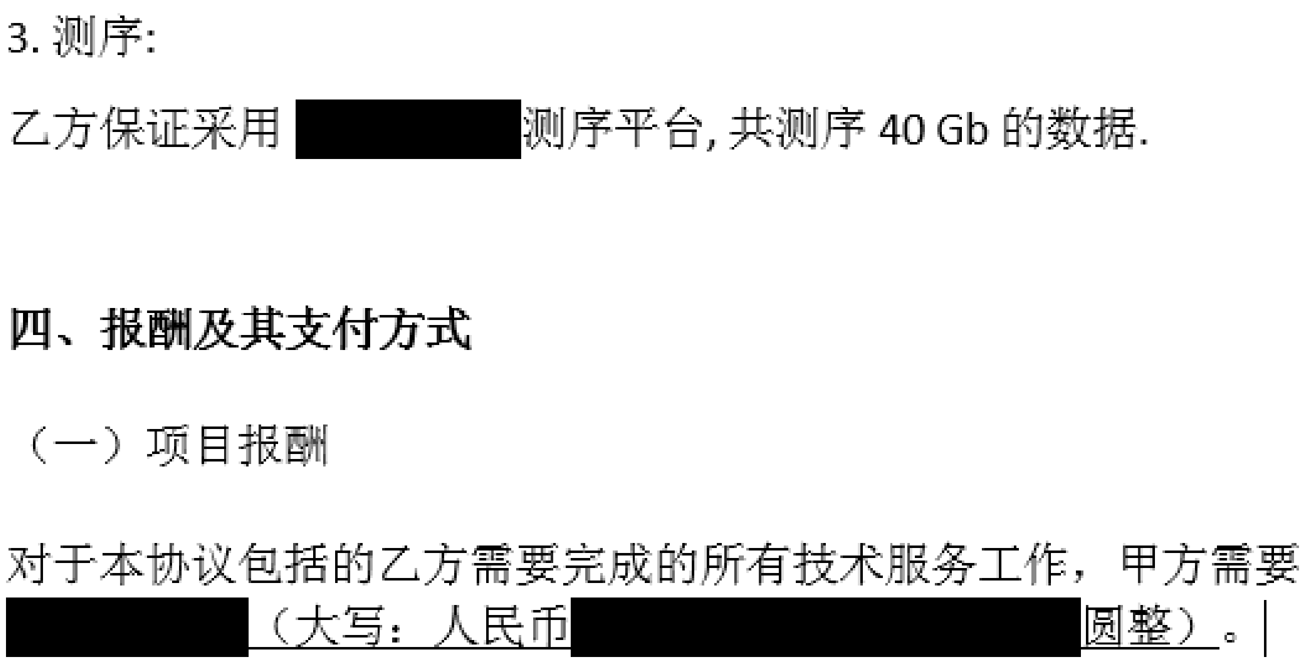


AUSTRALIA

Davis,
California

USA

Sequencing Contract



3. 测序:

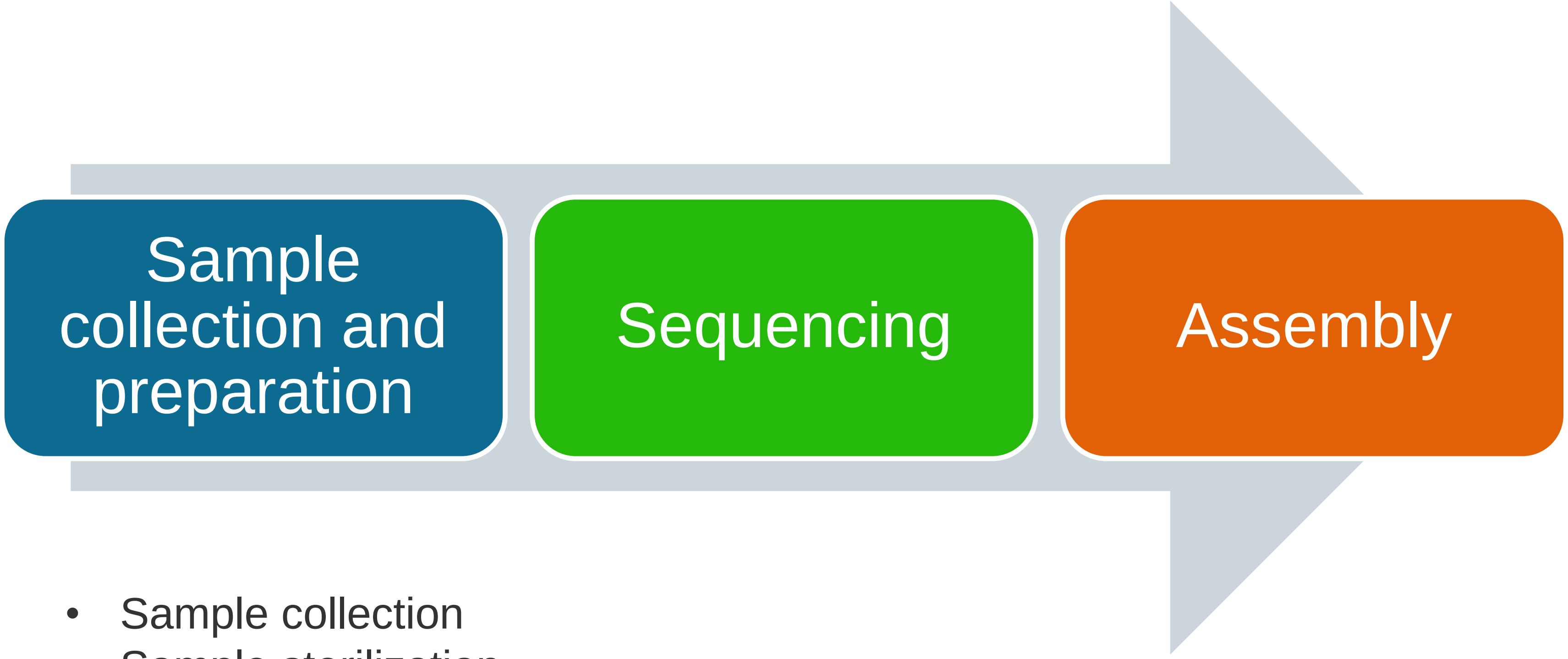
乙方保证采用 [redacted] 测序平台, 共测序 40 Gb 的数据.

四、报酬及其支付方式

（一）项目报酬

对于本协议包括的乙方需要完成的所有技术服务工作，甲方需要按服务内容向乙方支付的技术服务报酬为[redacted]（大写：人民币[redacted]圆整）。

Genome Assembly Project



- Sample collection
- Sample sterilization
- gDNA extraction
- gDNA QC
- Transportaion
- Sequencing library construction
- Library QC

Sequencing Platforms

illumina

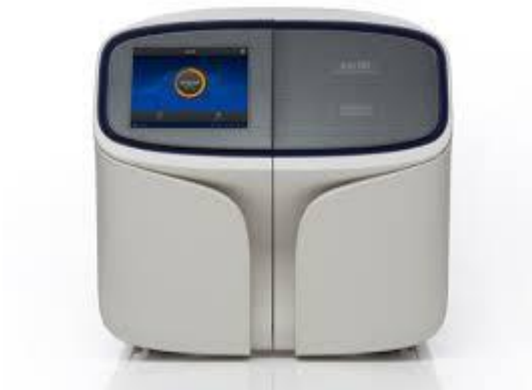
PACBIO

Oxford NANOPORE Technologies

BGISEQ-500

ThermoFisher SCIENTIFIC

Ion Torrent



applied biosystems

Sanger Sequencing

Sequencing Platforms

	NovaSeq 6000	HiSeq X Ten	HiSeq 4000
Output Range	167–6000 Gb	900–1800 Gb	125–1500 Gb
Run Time	19–40 hr	<3 days	3.5 days
Reads per Run	1.4–20 billion	3–6 billion	2.5 - 5 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp
Samples per Run	4–48	8–16	6 - 12
Relative Price per Sample	Higher Cost	Lower Cost	Mid Cost

illumina

Short read:

- High accuracy
- Deep coverage
- Cheap

- Illumina SLR

X Ten is NOT 10X !

Sequencing Platforms

	Sequel	RS II
Average read length	10 - 15 Kb	10 Kb
Throughput per cell	~5 - 10 Gb	500 Mb ~ 1 Gb
SMRT Cells per run	1 - 16	1 - 16
Movie lengths per SMRT Cell	30 mins - 6 hrs	30 mins - 6 hrs

PACBIO®

- Single Molecular Real Time
- Long read length
- No PCR, less bias
- Higher error rate
- More expensive
- Read length distribution

MinION

SmidgION

ONT

Oxford NANOPORE Technologies

Sequencing Considerations & Experimental Design

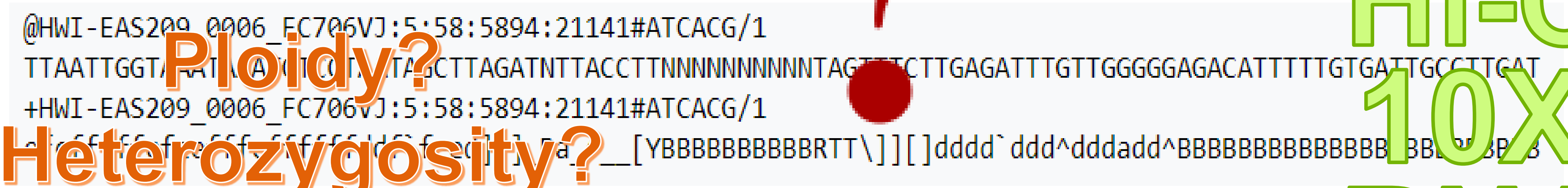
Sequencing

1304KHS-0092/CKVAopDA4_2.fastq.gz

D1B7JACXX_SD1_29_3_GCCAAT_L002_R1_001.fastq.bz2

C3BC5ACXX_NoIndex_L004_R1.fastq.gz

Genome Size?



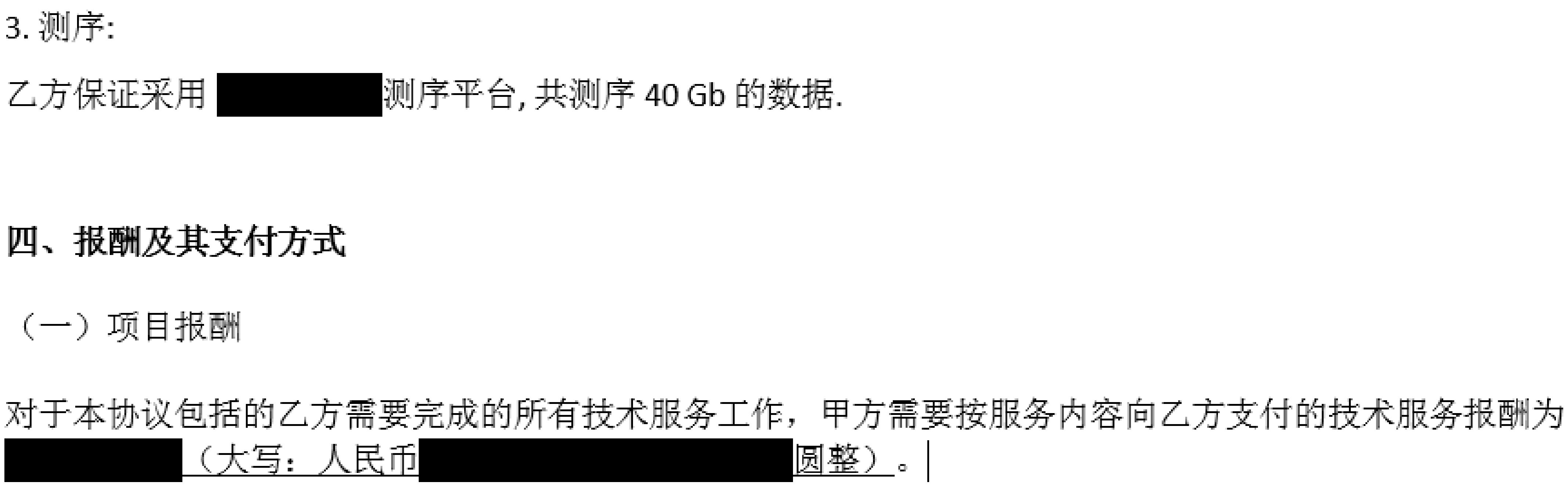
```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGT[illegible]CTTAGATNTTACCTTNNNNNNNNNNTAG[illegible]CTTGAGATTTGTTGGGGGAGACATTTTTGTGA[illegible]GAT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
[illegible]a[illegible]_[YBBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BBBBBBBBBBBBBB[illegible]B
```

Ploidy?

Heterozygosity?

PE?
MP?
Hi-C?
10X?
RNA?
GBS?

Sequencing Considerations & Experimental Design



3. 测序:

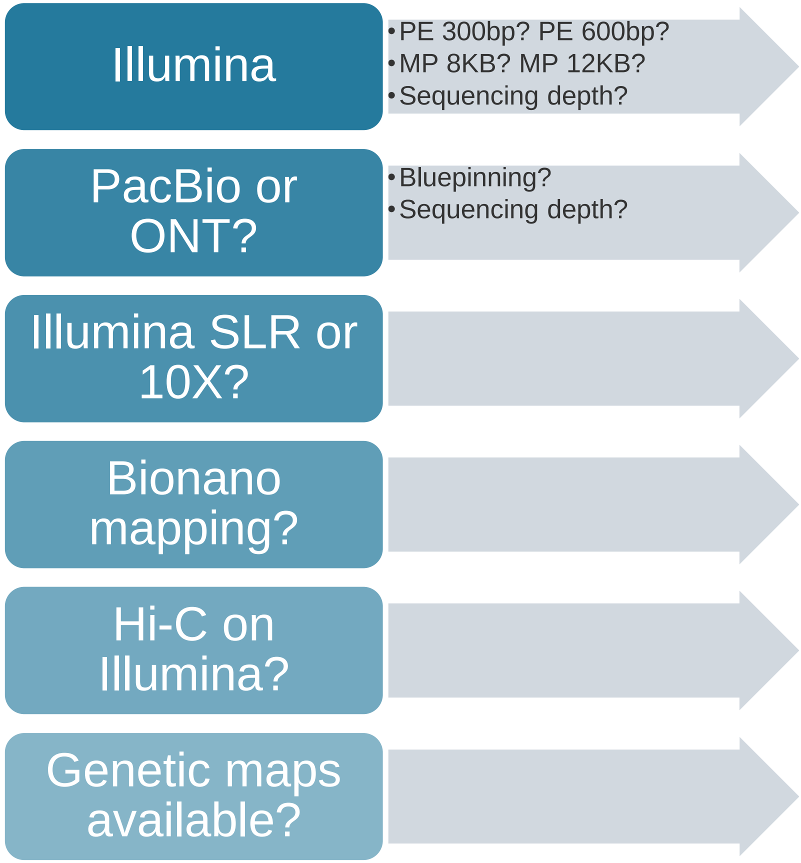
乙方保证采用 [redacted] 测序平台, 共测序40 Gb的数据.

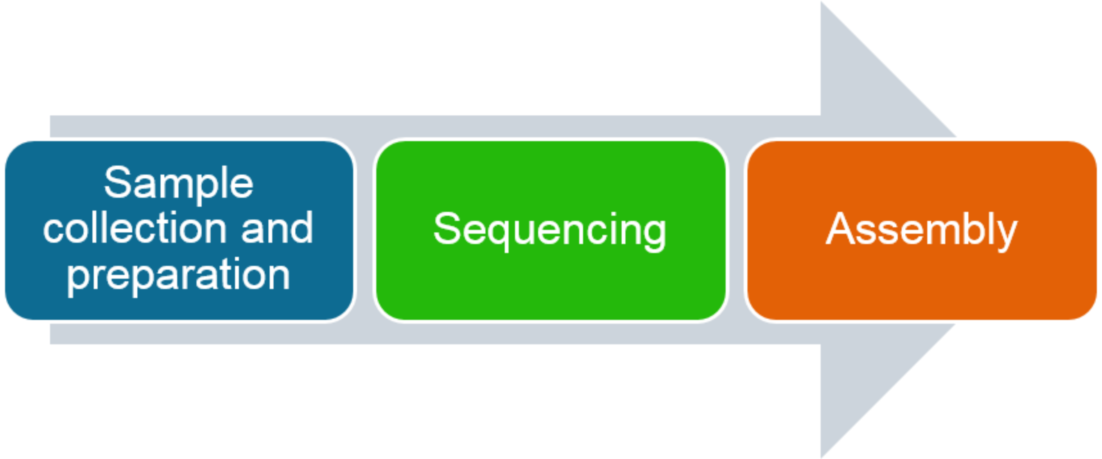
四、报酬及其支付方式

（一）项目报酬

对于本协议包括的乙方需要完成的所有技术服务工作，甲方需要按服务内容向乙方支付的技术服务报酬为[redacted]（大写：人民币[redacted]圆整）。

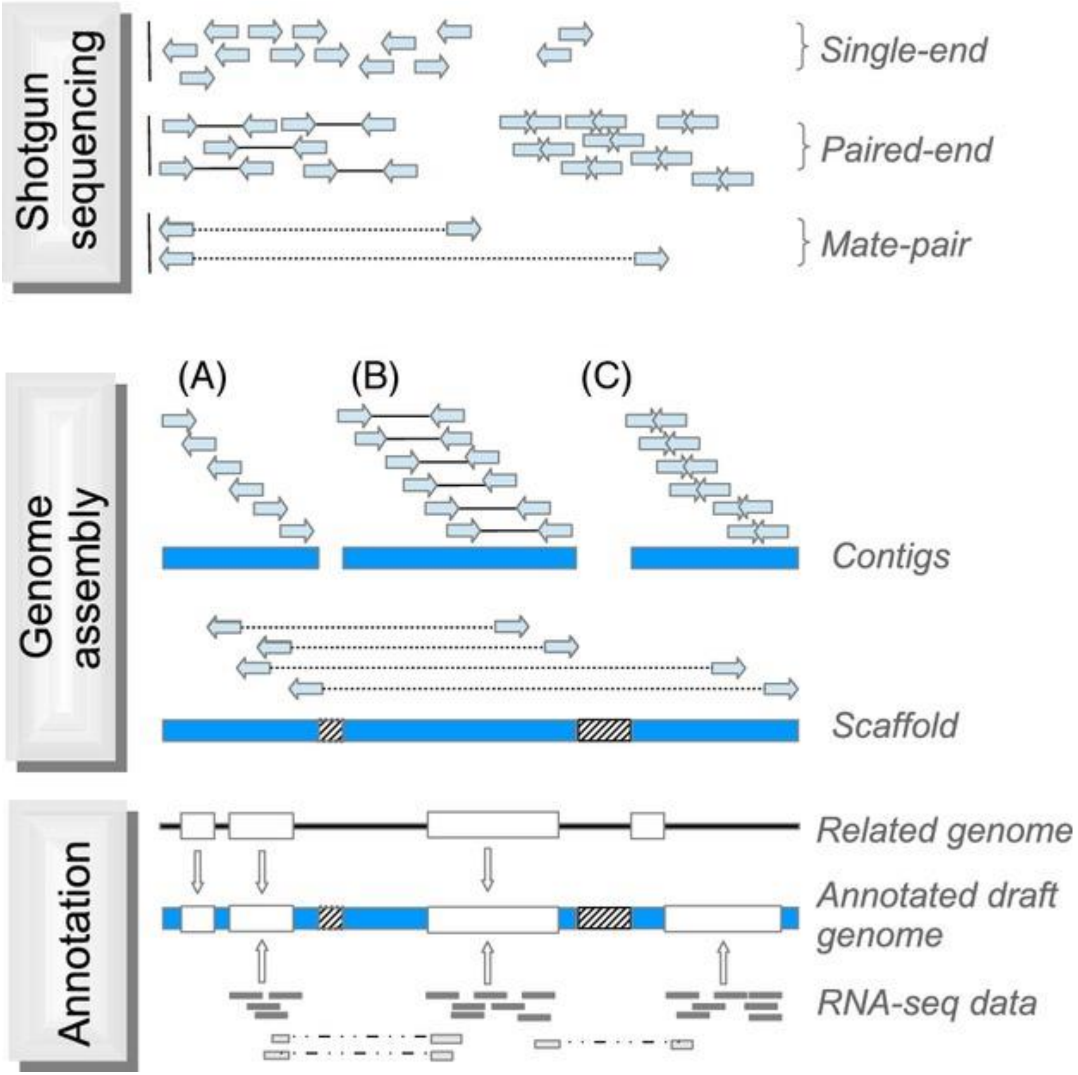
Sequencing Considerations & Experimental Design





Genome Assembly: The Concepts

Assembly

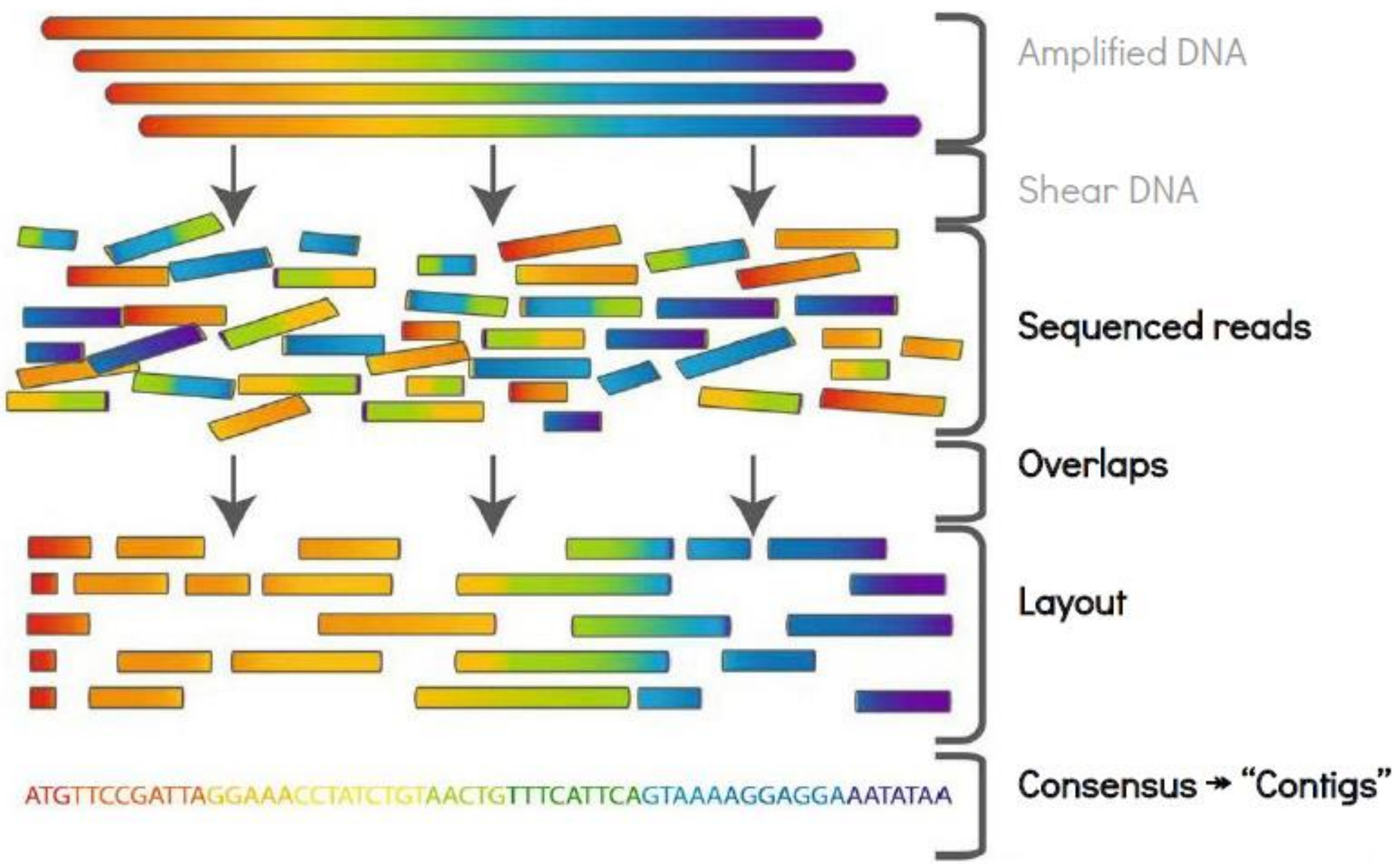


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231593/

Genome Assembly Algorithms: OLC, DBG & SG

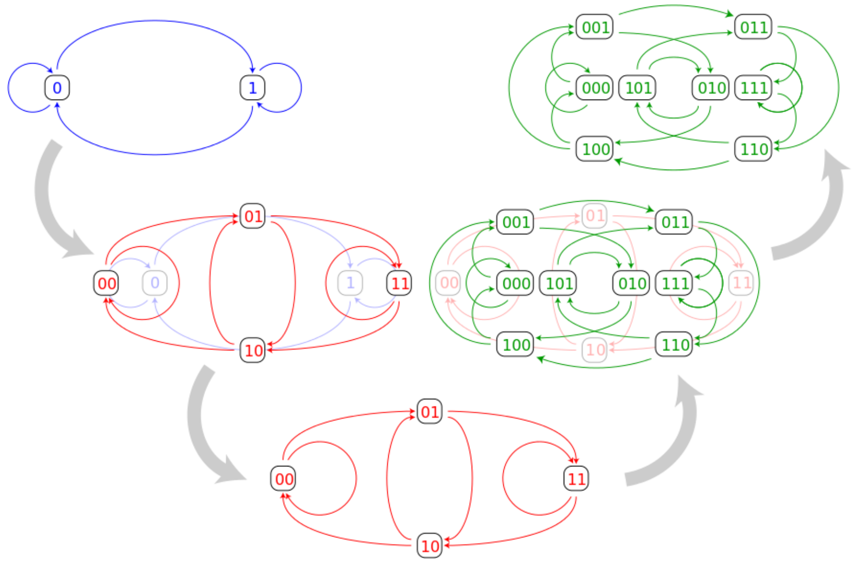
Assembly

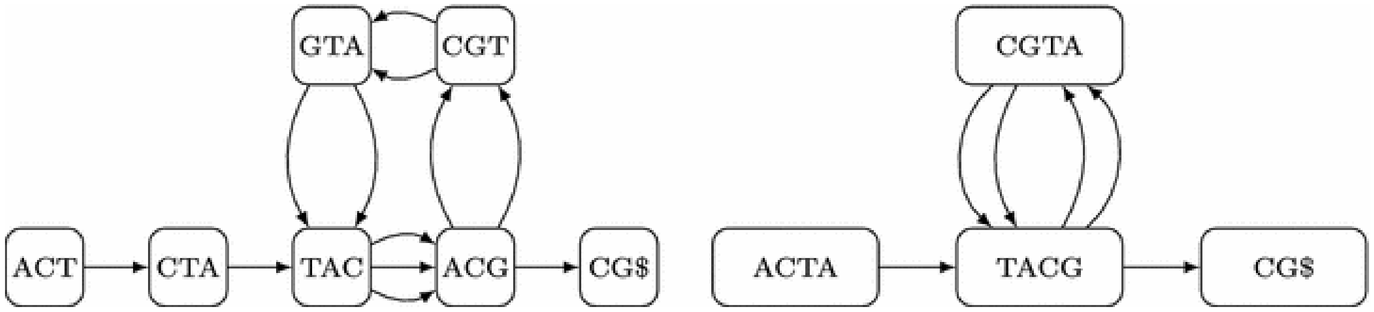
- OLC: Overlap-Layout-Consensus
  - Suitable for long reads
  - Newbler, Celera Assembler, PCAP, etc.



Genome Assembly Algorithms: OLC, DBG & SG

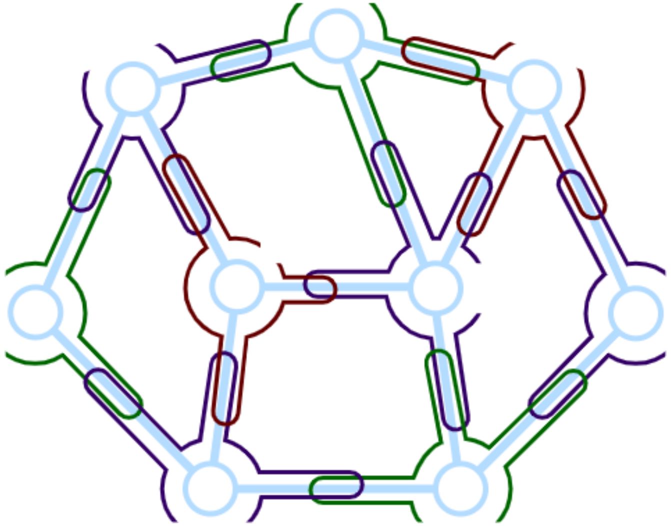
- DBG: De Bruijn Graph
  - An *n*-dimensional **De Bruijn graph** of *m* symbols is a directed graph representing overlaps between sequences of symbols
  - Each read is broken into fixed-size *k*-mers. A graph is directly constructed where each vertex is a *k*-mer and each edge indicates two adjacent *k*-mers overlapping by *k* – 1 letters.
  - Suitable for short reads
  - Velvet, AllPath-LG, ABySS, etc.

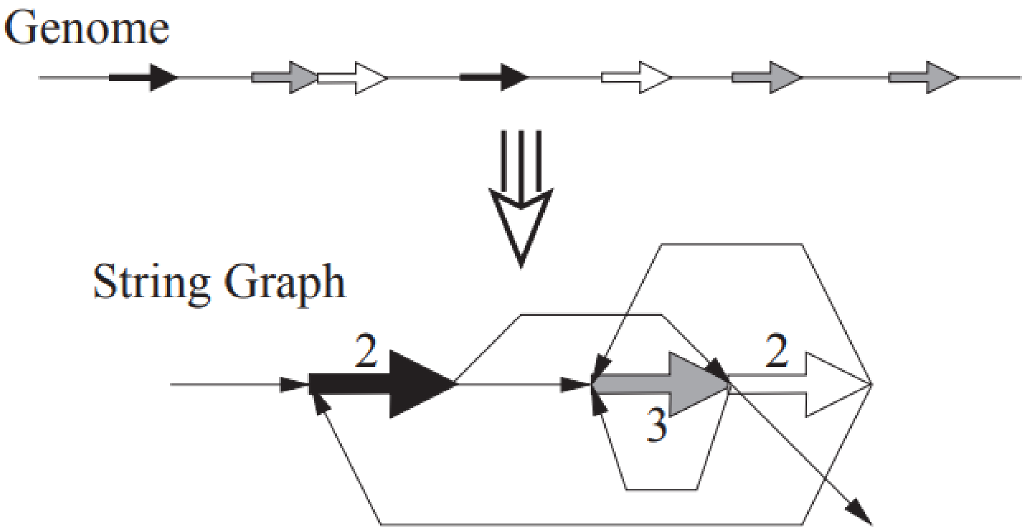




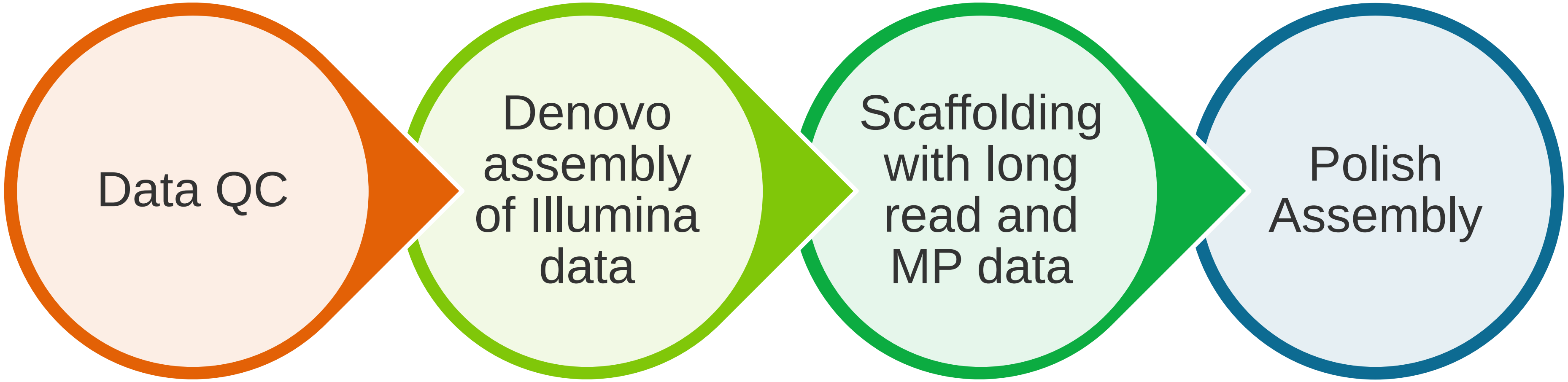
Genome Assembly Algorithms: OLC, DBG & SG

- SG: String Graph
  - In graph theory, a **string graph** is an intersection graph of curves in the plane; each curve is called a "string". Given a graph *G*, *G* is a string graph if and only if there exists a set of curves, or strings, drawn in the plane such that no three strings intersect at a single point and such that the graph having a vertex for each curve and an edge for each intersecting pair of curves is isomorphic to *G*.
  - Falcon





Hybrid Genome Assembly Strategies: Option 1



Data QC

Denovo assembly of Illumina data

Scaffolding with long read and MP data

Polish Assembly

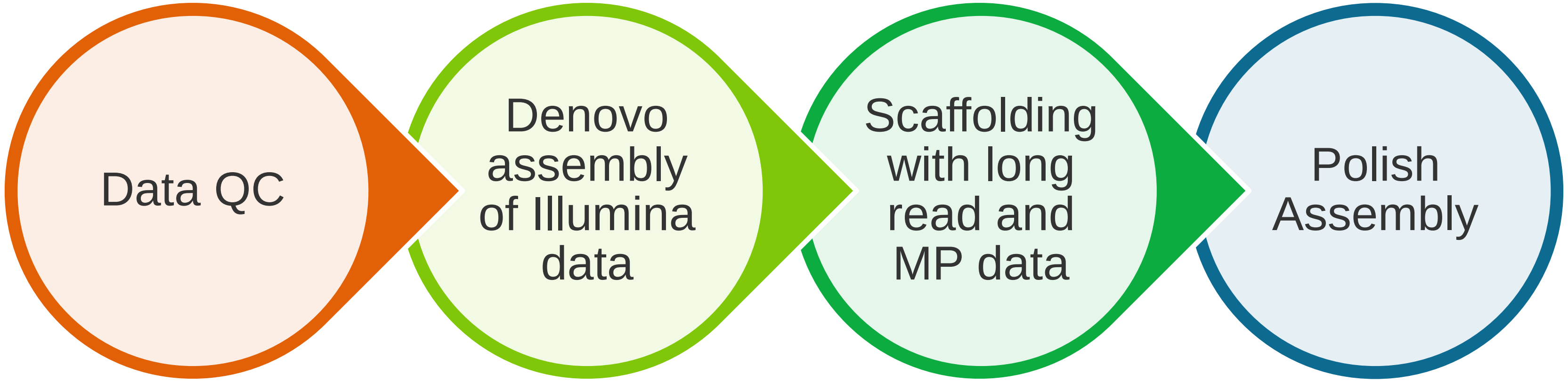
- Data integrity (md5sum)
- FastQC and MultiQC
- Fastp
- Trimmomatic
- Trim Galore!
- ErrorCorrection

- Velvet
- ALLPATHS-LG
- ABySS
- SOAPdenovo2
- MIRA
- SGA
- And many more

- SSPACE
- SOAPdenovo2
- SOPRA

- PBJelly
- OPERA-LG

Hybrid Genome Assembly Strategies: Option 1

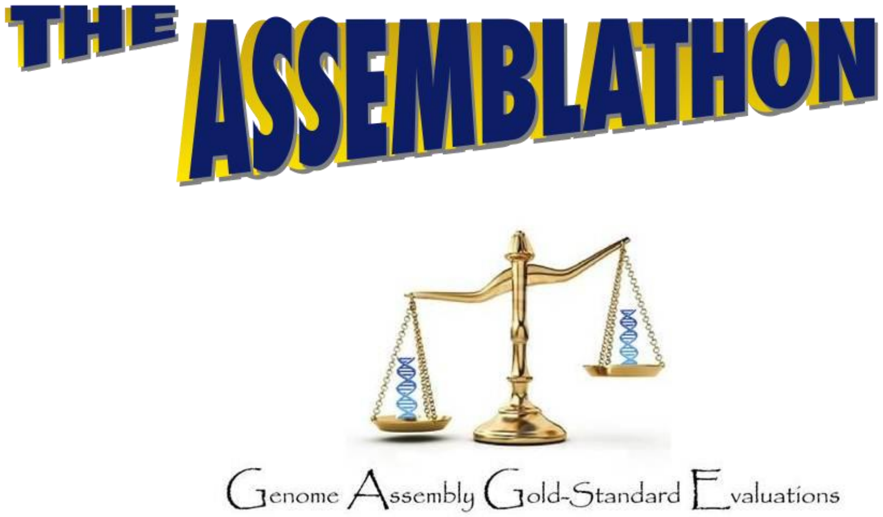


Data QC

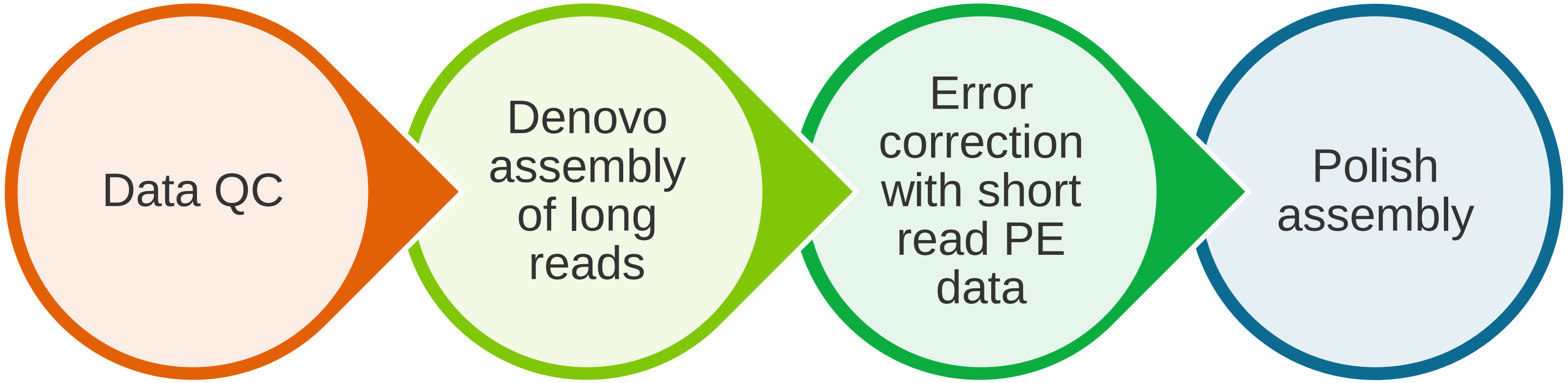
Denovo assembly of Illumina data

Scaffolding with long read and MP data

Polish Assembly



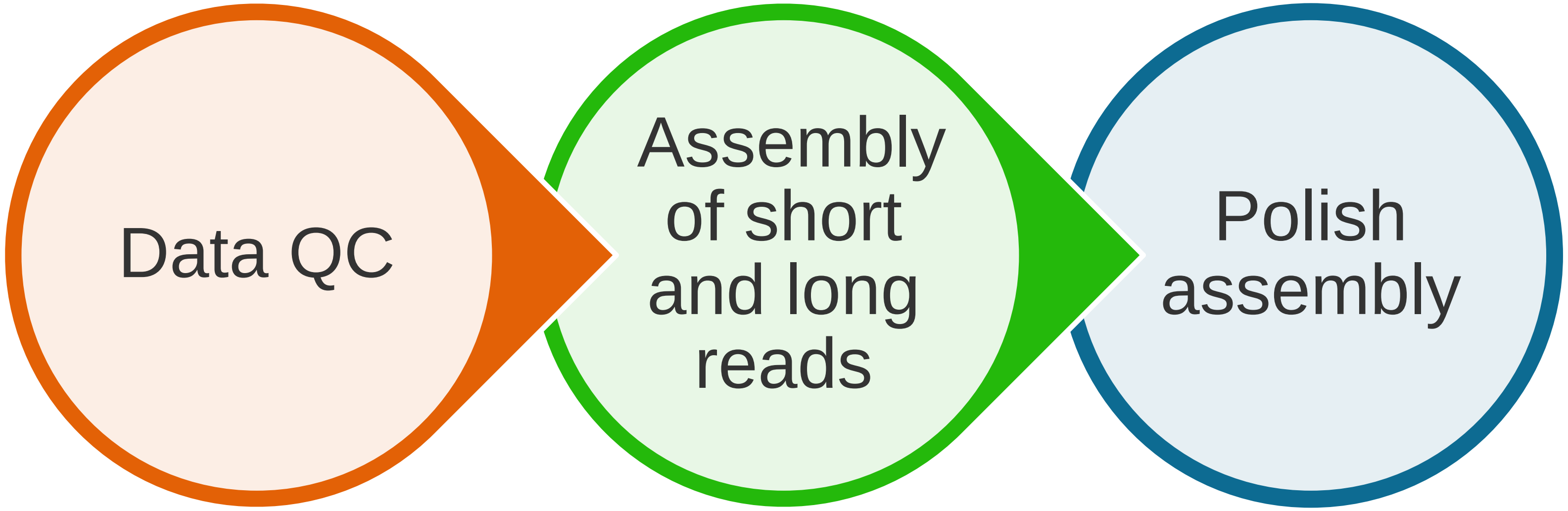
Hybrid Genome Assembly Strategies: Option 2


Data QC
Denovo assembly of long reads
Error correction with short read PE data
Polish assembly

- Canu
- Falcon and Falcon-Unzip
- HGAP4
- Celera Assembler
- mugqic/genpipes/pacbio_assembly

- Arrow/Quiver
- Pilon

Hybrid Genome Assembly Strategies: Option 3



- MaSuRCA
- Spades
- MIRA
- CABOG

Hybrid Genome Assembly

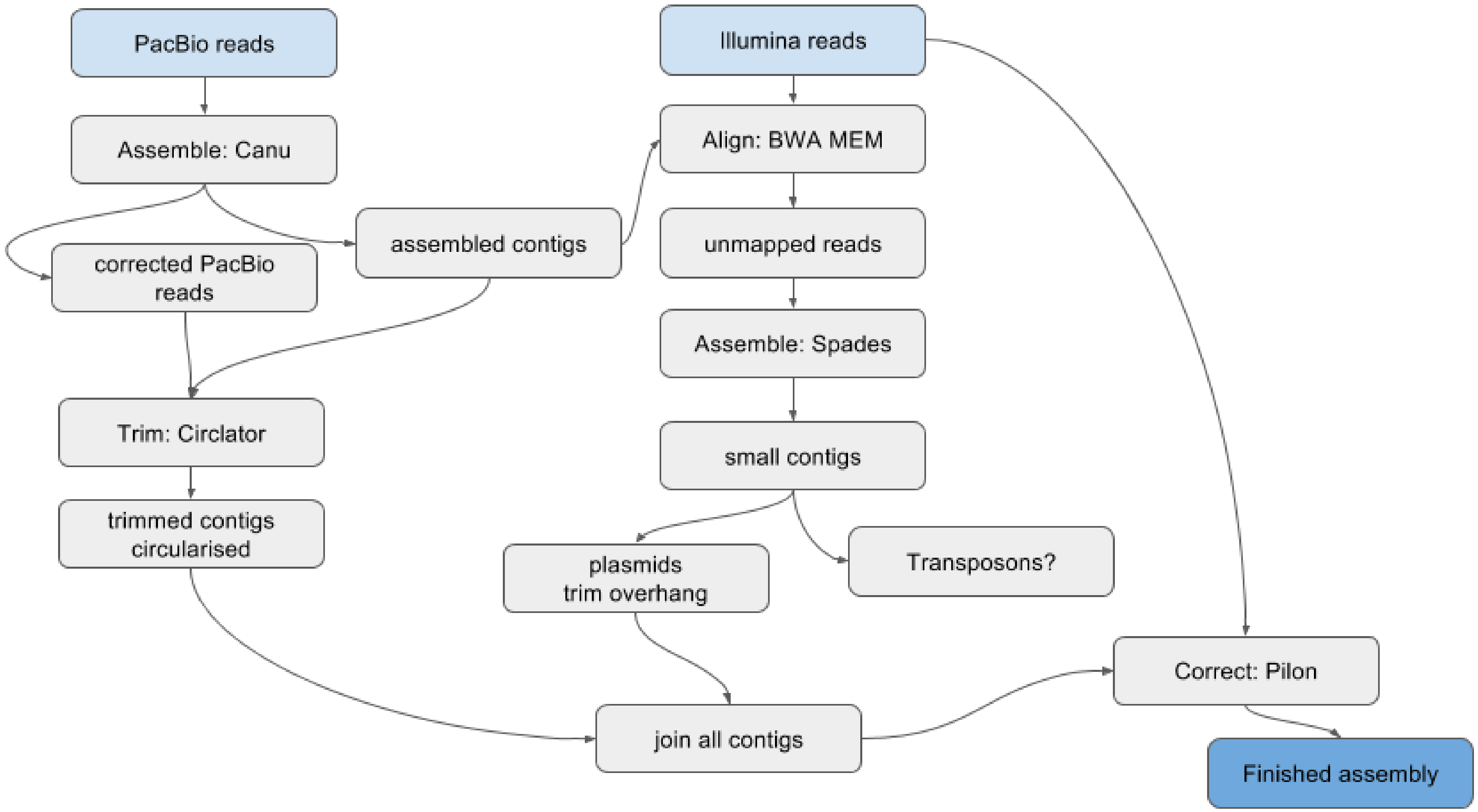
Assembly

Assembler	Algorithm	Input
Arachne	OLC	Sanger
CAP3	OLC	Sanger
TIGR	Greedy	Sanger
Newbler	OLC	454/Roche
Edena	OLC	Illumina
SGA	OLC	Illumina
MaSuRCA	De Bruijn/OLC	Illumina/PacBio
MIRA	De Bruijn/OLC	Illumina/PacBio/454/Sanger
Velvet	De Bruijn	Illumina
ALLPATHS	De Bruijn	Illumina/PacBio
ABySS	De Bruijn	Illumina
SOAPdenovo	De Bruijn	Illumina
Spades	Paired De Bruijn	Illumina/PacBio
CLC	De Bruijn	Illumina/454
CABOG	OLC	Hybrid
Falcon	String graph	PacBio
StriDe	String graph + De Brujin	Illumina

- Every species has it’s own surprises and characters
- Every sequencing chemistry has it’s strengths and weaknesses
- Every assembler has it’s own set of heuristics.

An Example Workflow To Assemble A Bacteria Genome

Command-line assembly



PacBio reads

Assemble: Canu

corrected PacBio reads

assembled contigs

Trim: Circlator

trimmed contigs circularised

Illumina reads

Align: BWA MEM

unmapped reads

Assemble: Spades

small contigs

plasmids
trim overhang

Transposons?

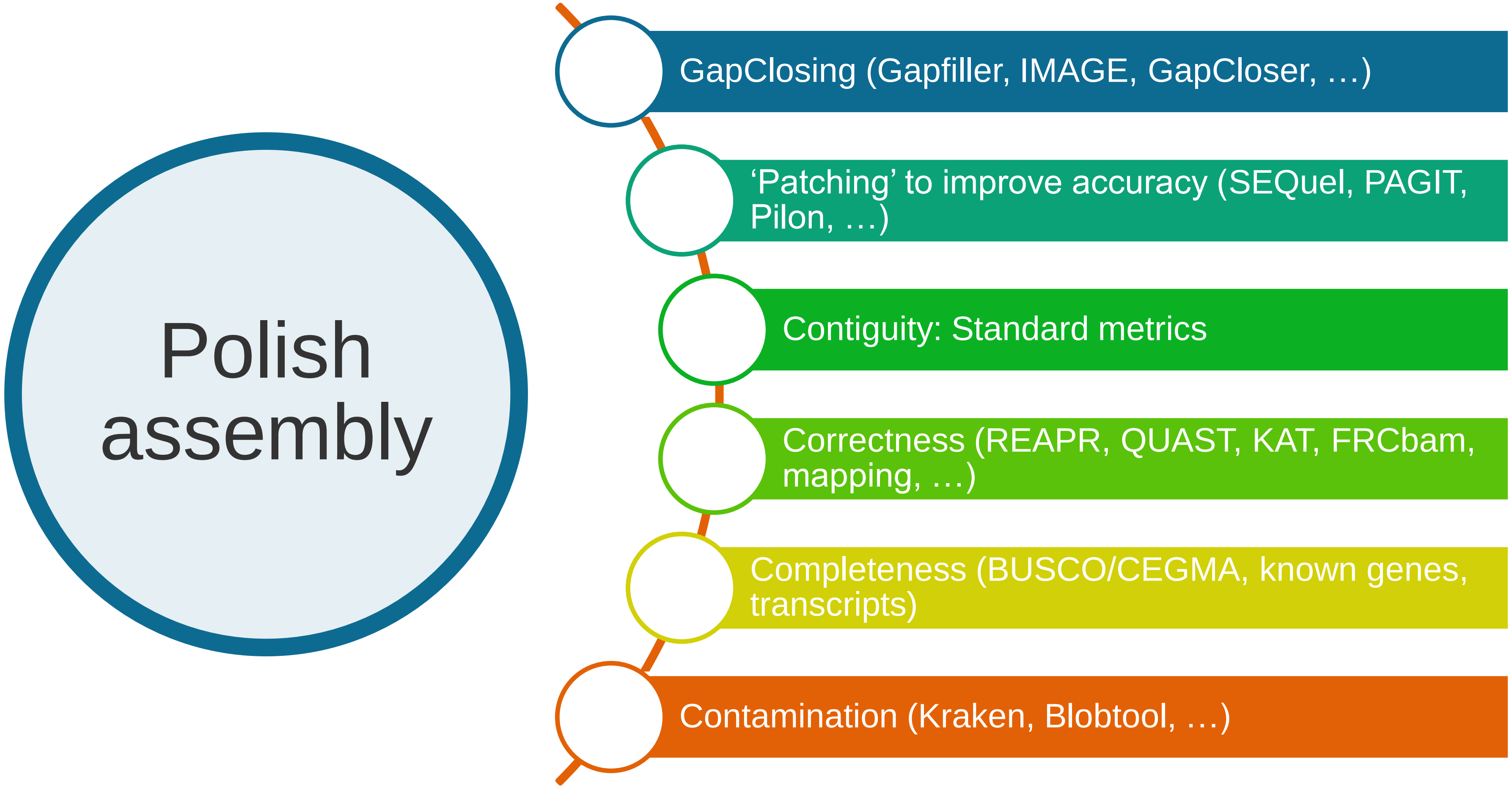
join all contigs

Correct: Pilon

Finished assembly

http://sepsis-omics.github.io/tutorials/modules/cmdline_assembly/

Assembly Assessment and Improvement At Scaffolds Level

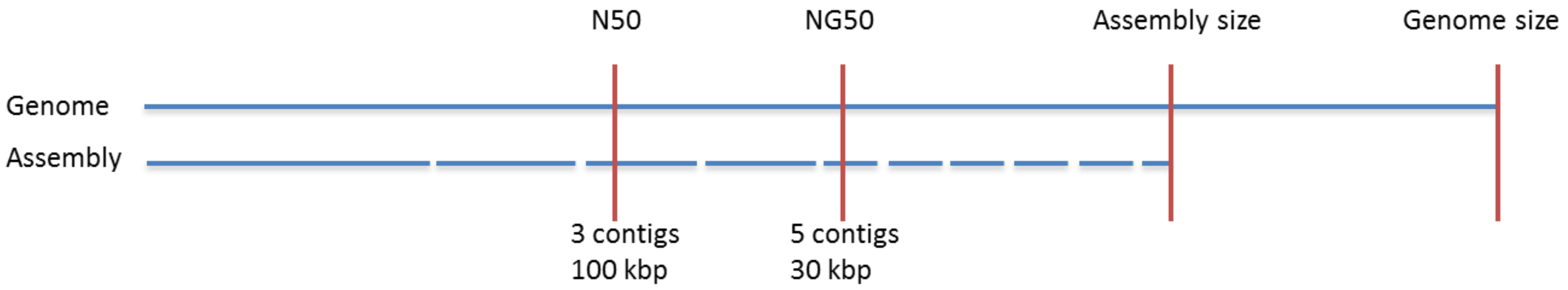


Polish assembly

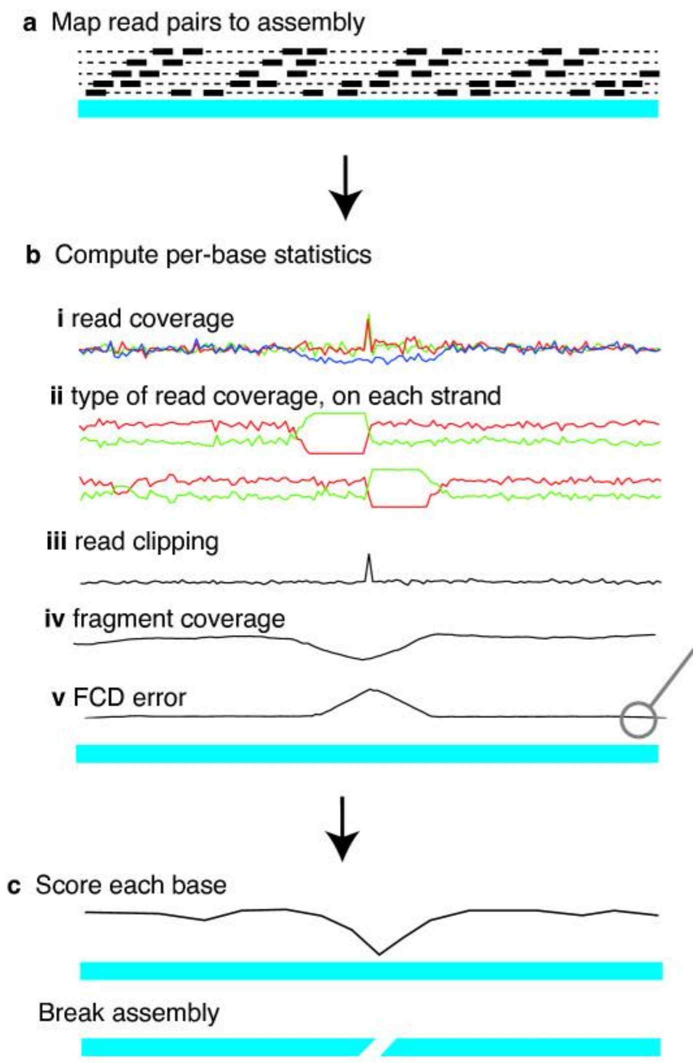
- GapClosing (Gapfiller, IMAGE, GapCloser, …)
- ‘Patching’ to improve accuracy (SEQuel, PAGIT, Pilon, …)
- Contiguity: Standard metrics
- Correctness (REAPR, QUAST, KAT, FRCbam, mapping, …)
- Completeness (BUSCO/CEGMA, known genes, transcripts)
- Contamination (Kraken, Blobtool, …)

Assembly Assessment: Standard Metrics

- Standard metrics
  - Assembled size, # of contigs, # of scaffolds, N50, size of the longests contig, size of the longest scaffold, etc.
- N50: The length of the longest sequence such that the sum of sequences longer than it reaches half of the assembled size
- NG50: The length of the longest sequence such that the sum of sequences longer than it reaches half of the genome size

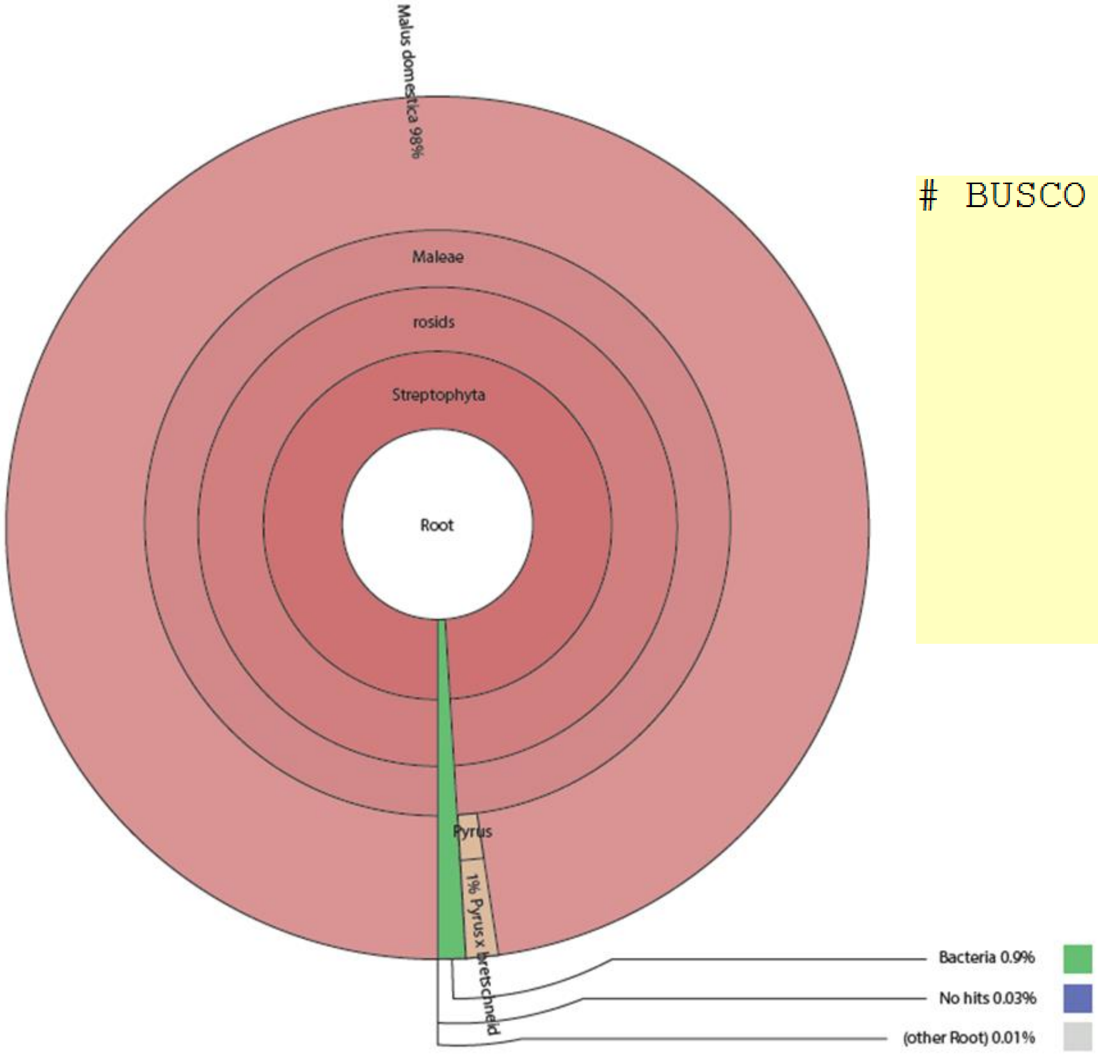


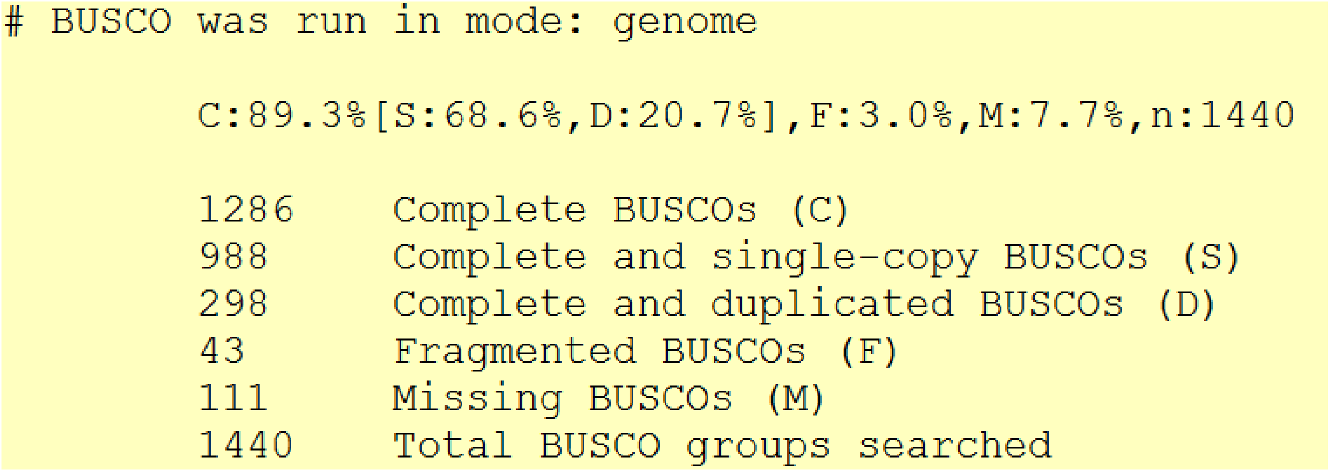
Assembly Assessment: Correctness Check with REAPR


a Map read pairs to assembly
b Compute per-base statistics
i read coverage
ii type of read coverage, on each strand
iii read clipping
iv fragment coverage
v FCD error
c Score each base
Break assembly

- Uses the same principle of feature response curve (FRC)
- Captures trade-off between quality and contiguity
- Identifies erroneous positions
- Breaks sequences at suspicious positions

Assembly Assessment: Completeness and Contamination Check

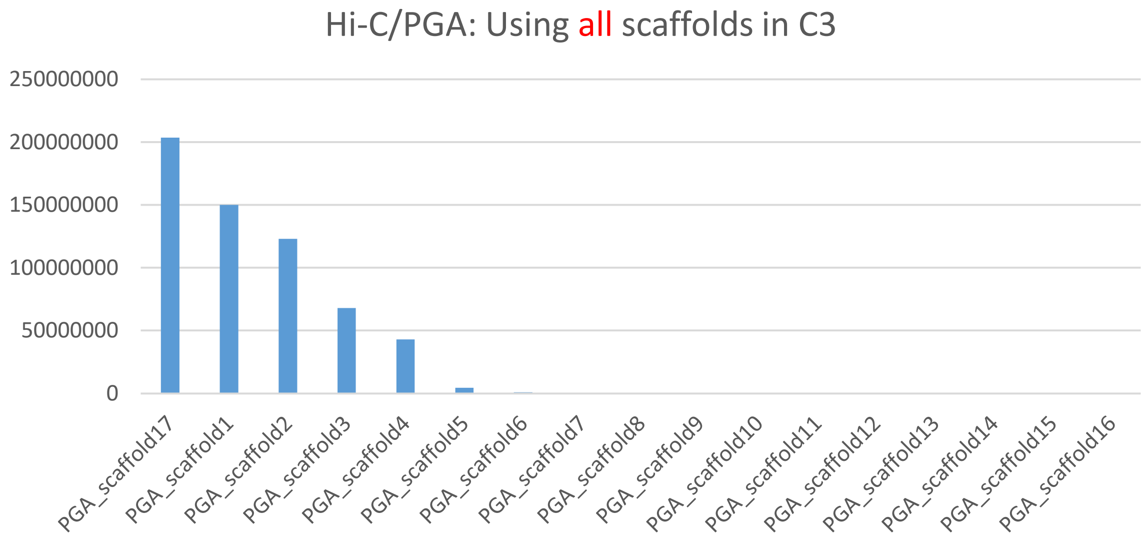


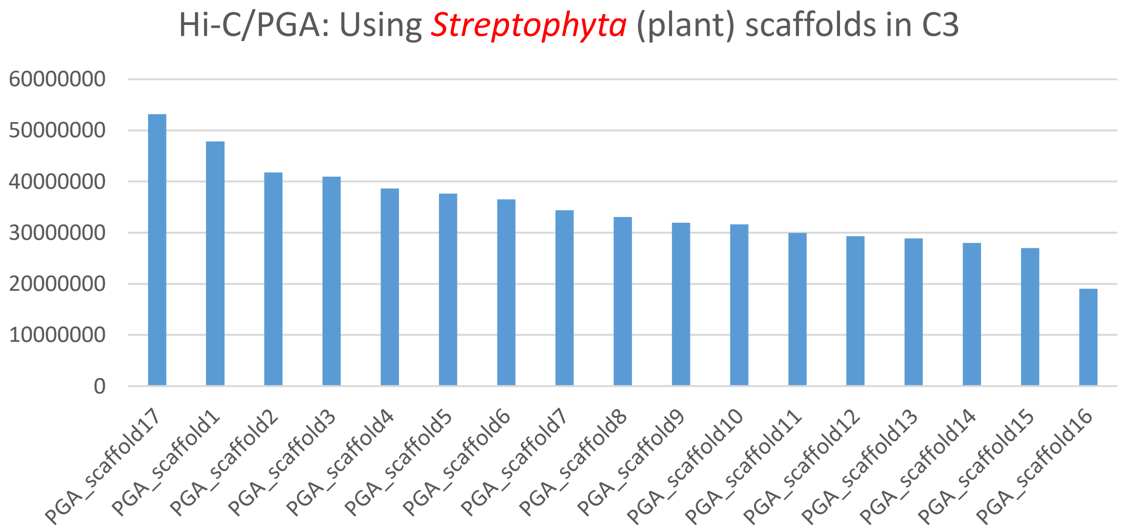


```
# BUSCO was run in mode: genome

        C:89.3%[S:68.6%,D:20.7%],F:3.0%,M:7.7%,n:1440

        1286    Complete BUSCOs (C)
        988     Complete and single-copy BUSCOs (S)
        298     Complete and duplicated BUSCOs (D)
        43      Fragmented BUSCOs (F)
        111     Missing BUSCOs (M)
        1440    Total BUSCO groups searched
```

Figure 4: 'Royal Gala' assembly C3: Scaffold classification and contamination check.

Assembly Assessment: Completeness and Contamination Check

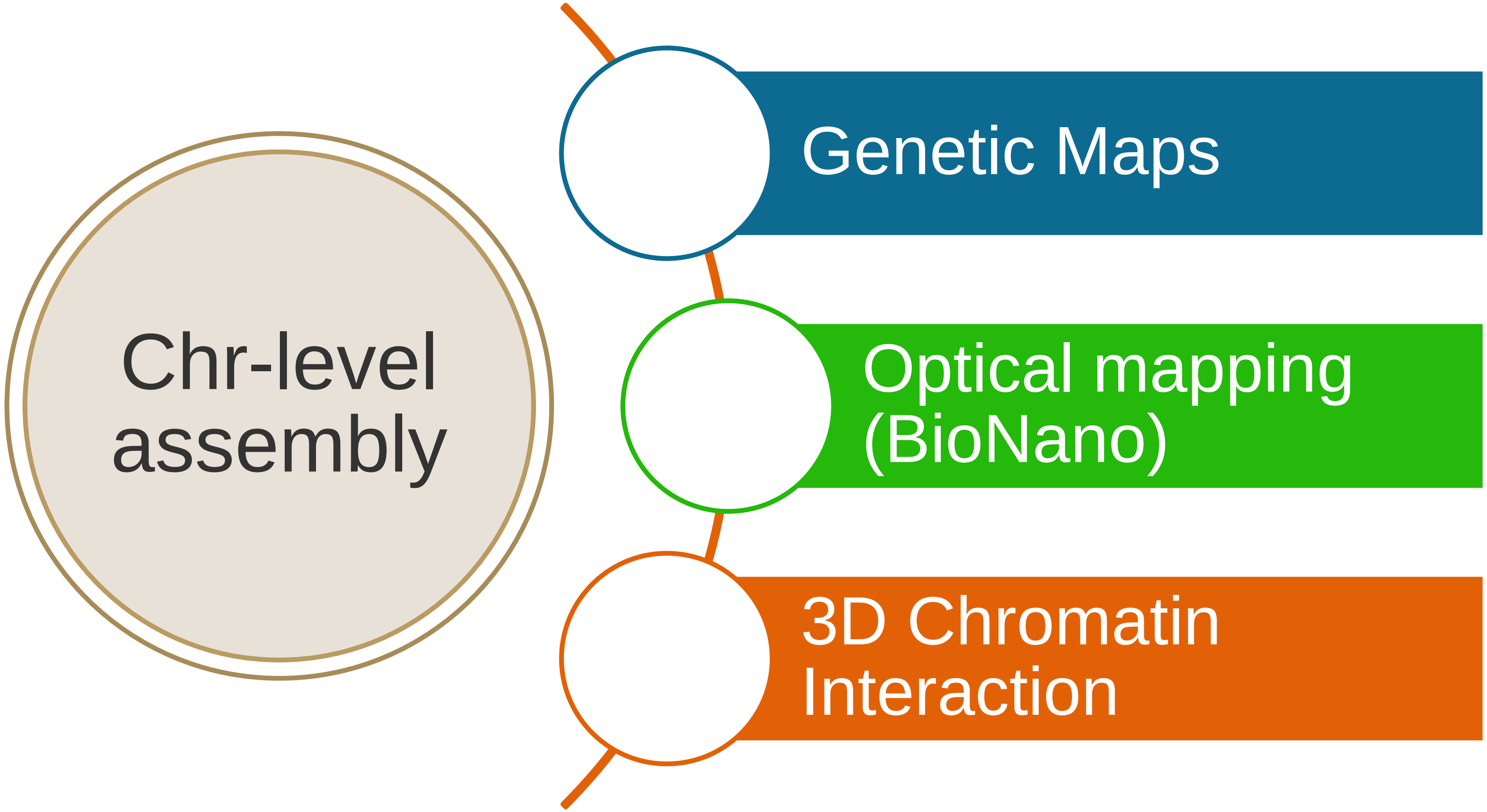




Assembly Assessment: More Examples of Contamination Check

- Example 2: Puccinia co.
- Example 3: Puccinia tr.

Assembly Improvement To Pseudo Chromosome Level



Chr-level assembly

Genetic Maps

Optical mapping (BioNano)

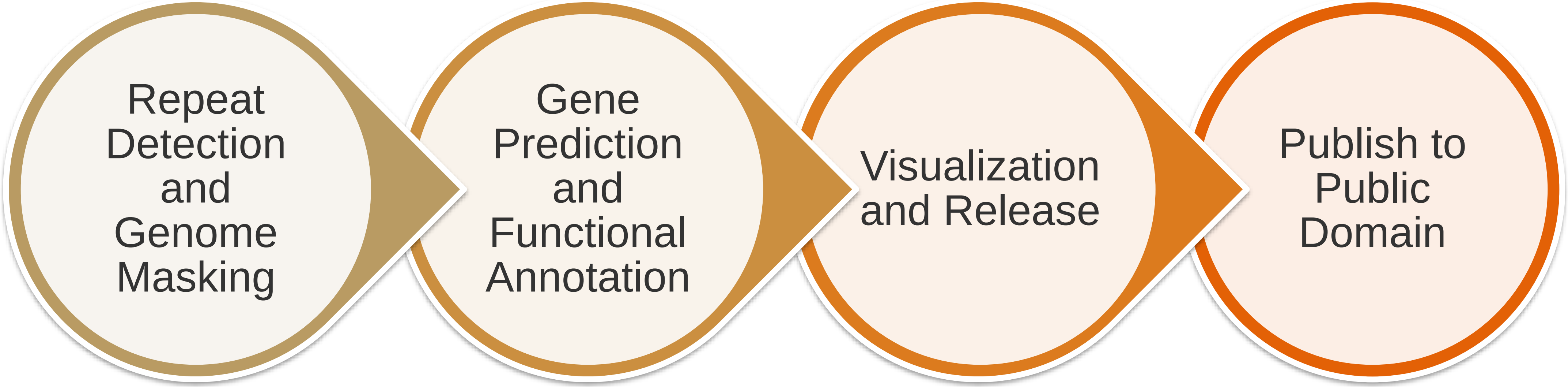
3D Chromatin Interaction

Chromonomer

BioNanoAnalyst

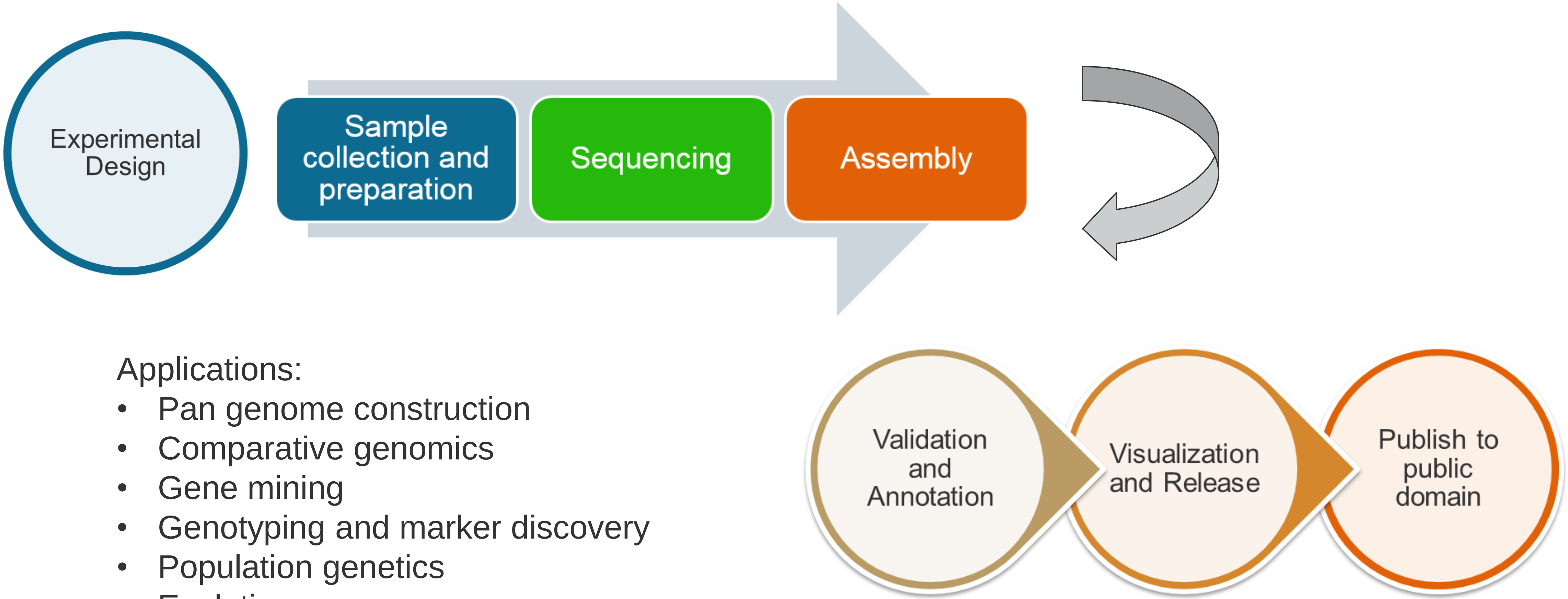
HIC -PGA, Dovetail Chicago HiRise, SALSA, 3D-DNA, LACHESIS

Genome Assembly Post-Processing


Repeat Detection and Genome Masking
Gene Prediction and Functional Annotation
Visualization and Release
Publish to Public Domain

Genome Annotation

A Complete Genome Project Workflow


Experimental Design
Sample collection and preparation
Sequencing
Assembly
Validation and Annotation
Visualization and Release
Publish to public domain

Applications:

- Pan genome construction
- Comparative genomics
- Gene mining
- Genotyping and marker discovery
- Population genetics
- Evolution
- GWAS
- GS

Conclusions

- Genome assembly can be complicated
- Experimental design is critical
- 4Cs: Contiguity/Correctness/Completeness/Contamination
- Assemblies are not perfect
  - Species specific difficulties (repeat, polymorphism, ploidy)
  - Sequencing chemistry
  - Regions not clone/sequence/assemble well
  - Software heuristics
- Know when to stop!

Time frame

Funds

Quality

Biointormatics landscape changes fast

Acknowledgements

- Pipfruit Breeding Team
- Mapping and Markers Team
- Molecular Biology
- Bioinformatics Team
- Flavour Team
- Kiwifruit Breeding Team

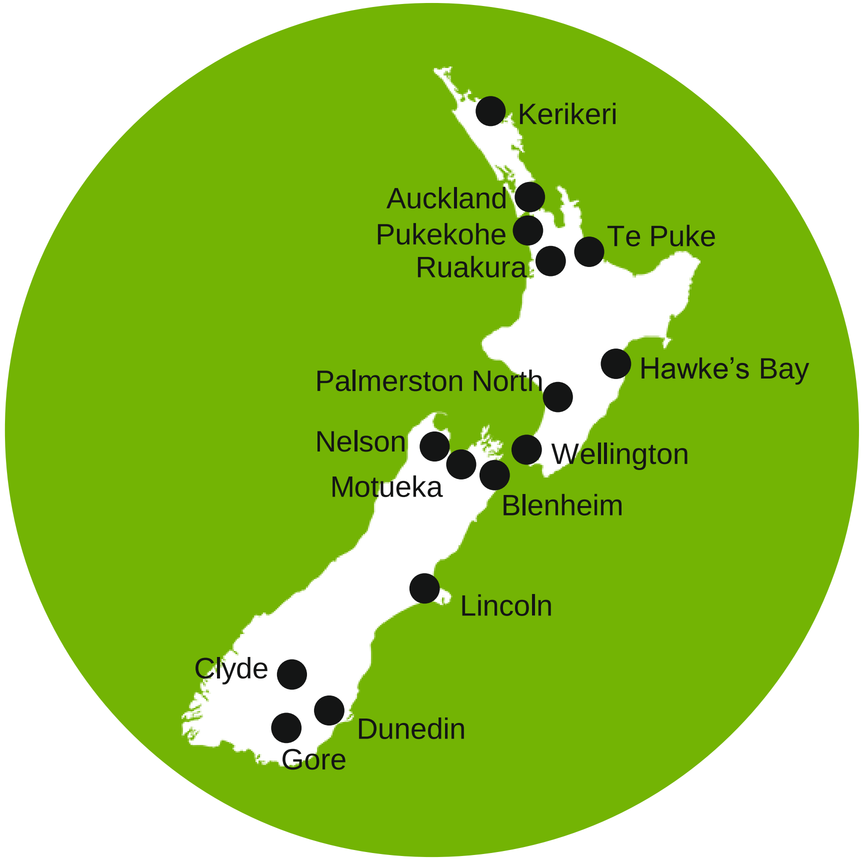
Ministry of Business,
Innovation & Employment

Zespri®
Kiwifruit

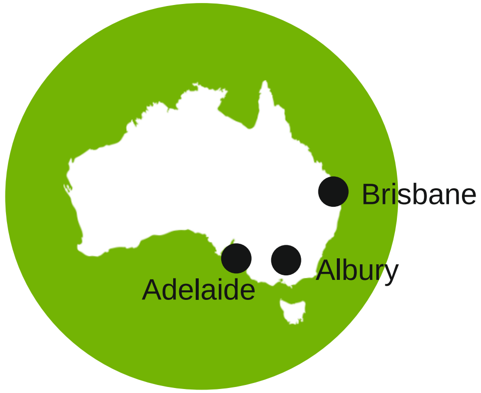
Prevar
PREMIUM APPLE & PEAR PRODUCTS

SlipStream
Automation

Plant & Food Research (PFR)



NEW ZEALAND



AUSTRALIA

Davis,
California

USA

Thank You!

Plant & Food
RESEARCH
RANGAHAU AHUMĀRA KAI

plantandfood.co.nz

Cecilia.Deng@plantandfood.co.nz

OUR SCIENCE IS GROWING FUTURES™

The New Zealand Institute for Plant & Food Research Limited