Plant & Food™ Research
Rangahau Ahumāra Kai

Hi-C data QC and genome scaffolding

Chen Wu (PhD), Bioinformatician, PFR

Shane Choi (PhD candidate) Manaaki Whenua Landcare Research

About Shane

- PhD student, approaching submission date
- Co-author rewarewa genome publication

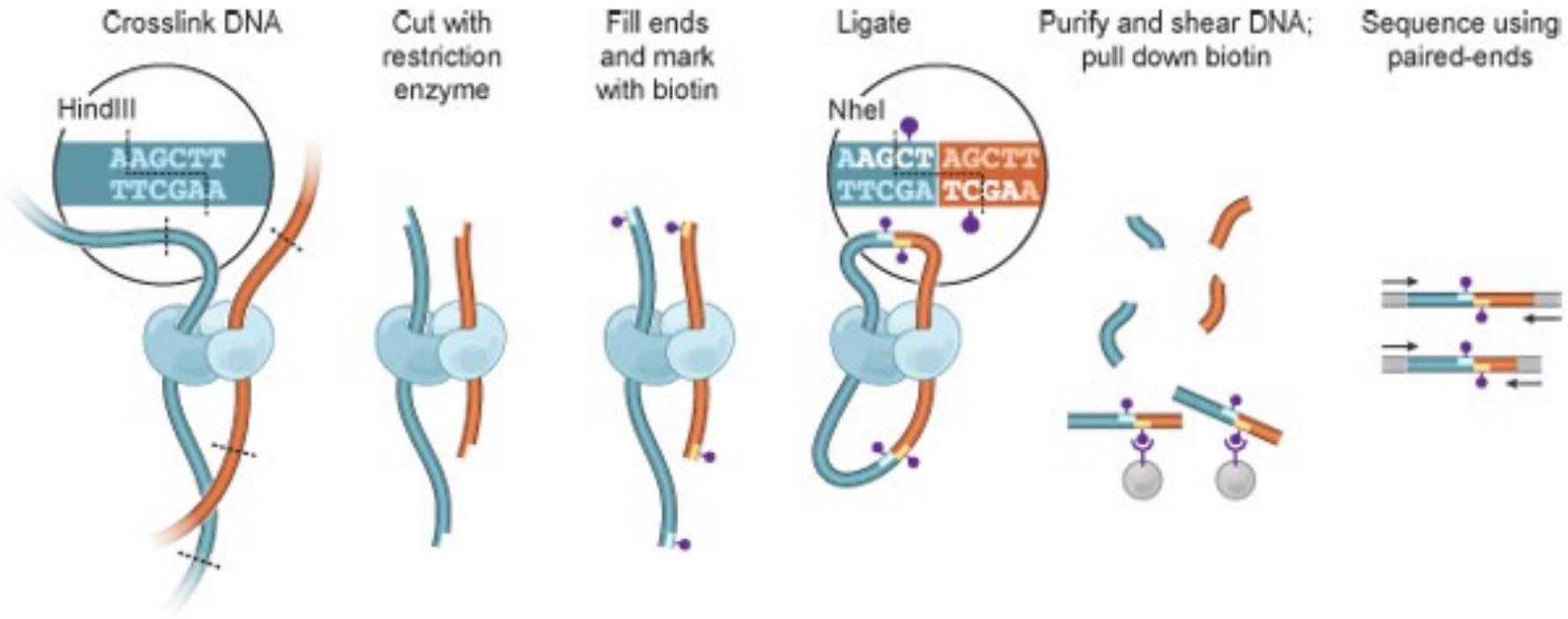
An exploration of assembly strategies and quality metrics on the accuracy of the *Knightia excelsa* (rewarewa) genome

Ann McCartney, Elena Hilario, Seung-Sub Choi, Joseph Guhlin, Jessica M. Prebble, Gary Houliston, Thomas R. Buckley, David Chagné

doi: https://doi.org/10.1101/2020.10.28.358903

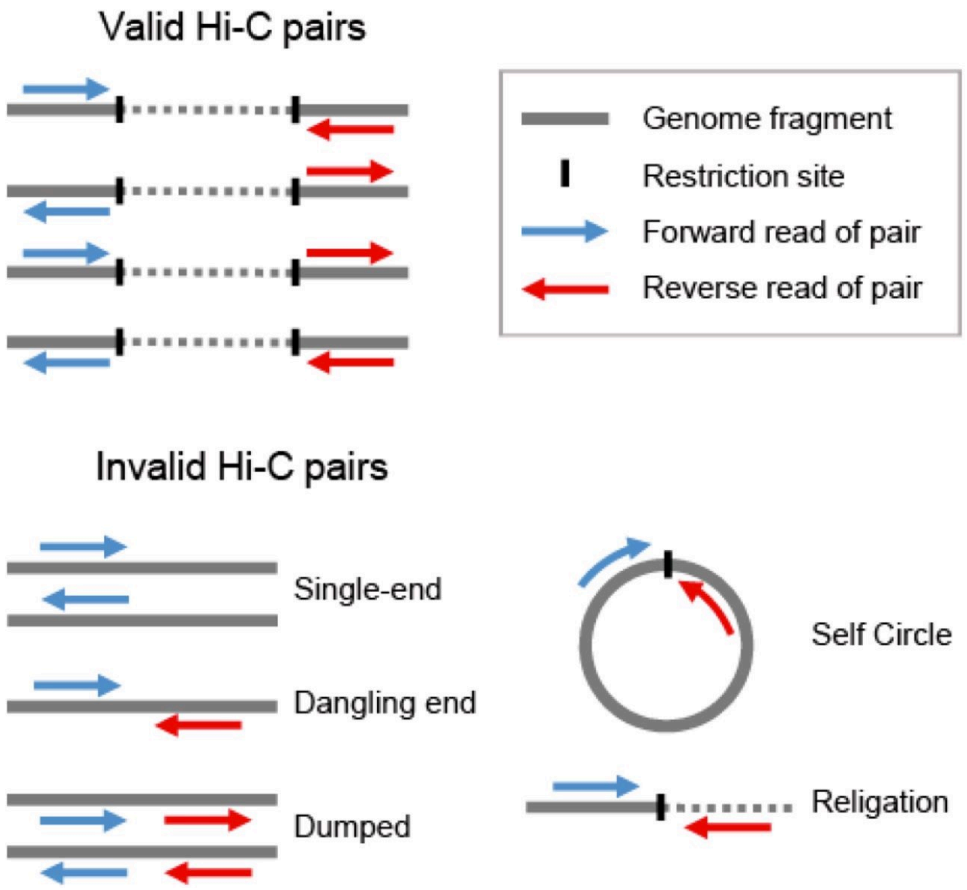
- Specialising in bioinformatics-based research, such as genome assembly, population genetics and currently working on triploid phasing
- Species: stick insects, rewarewa
- Contact: scho249@aucklanduni.ac.nz

Hi-C data



van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., Lander, E. S. Hi-C: A Method to Study the Three-dimensional Architecture of Genomes.. J. Vis. Exp. (39), e1869, doi:10.3791/1869 (2010).

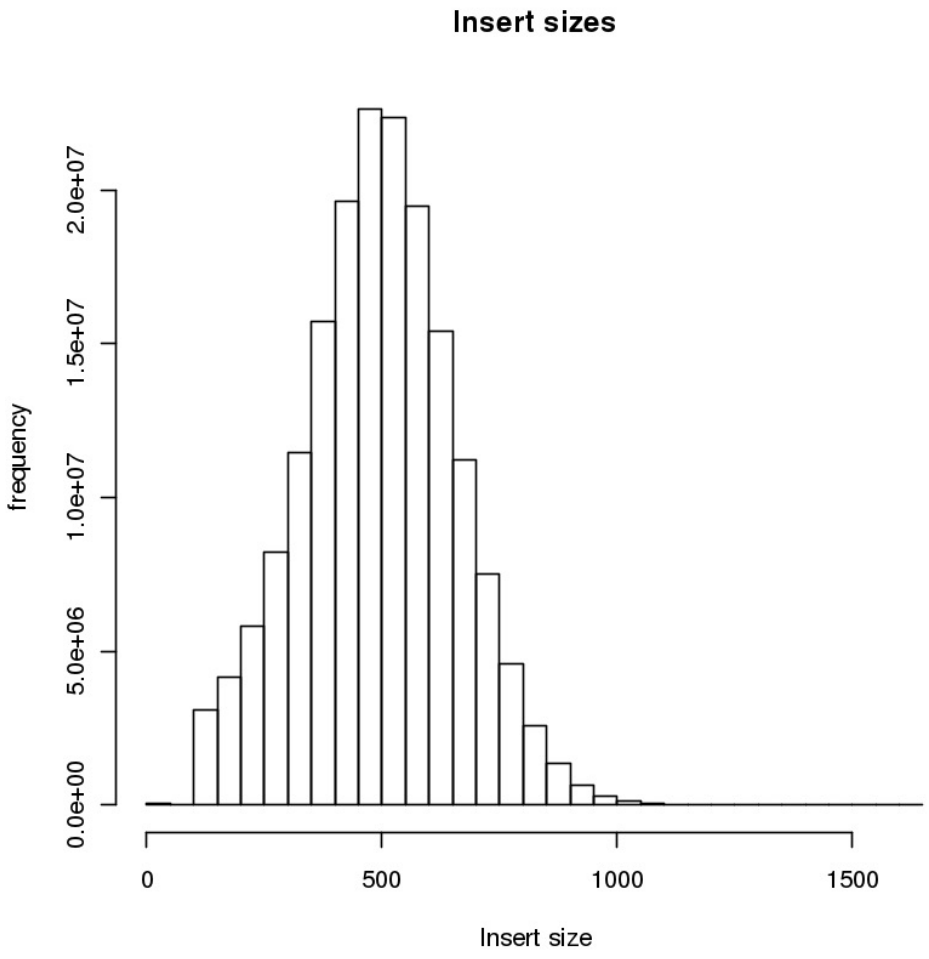
What is good HiC data?



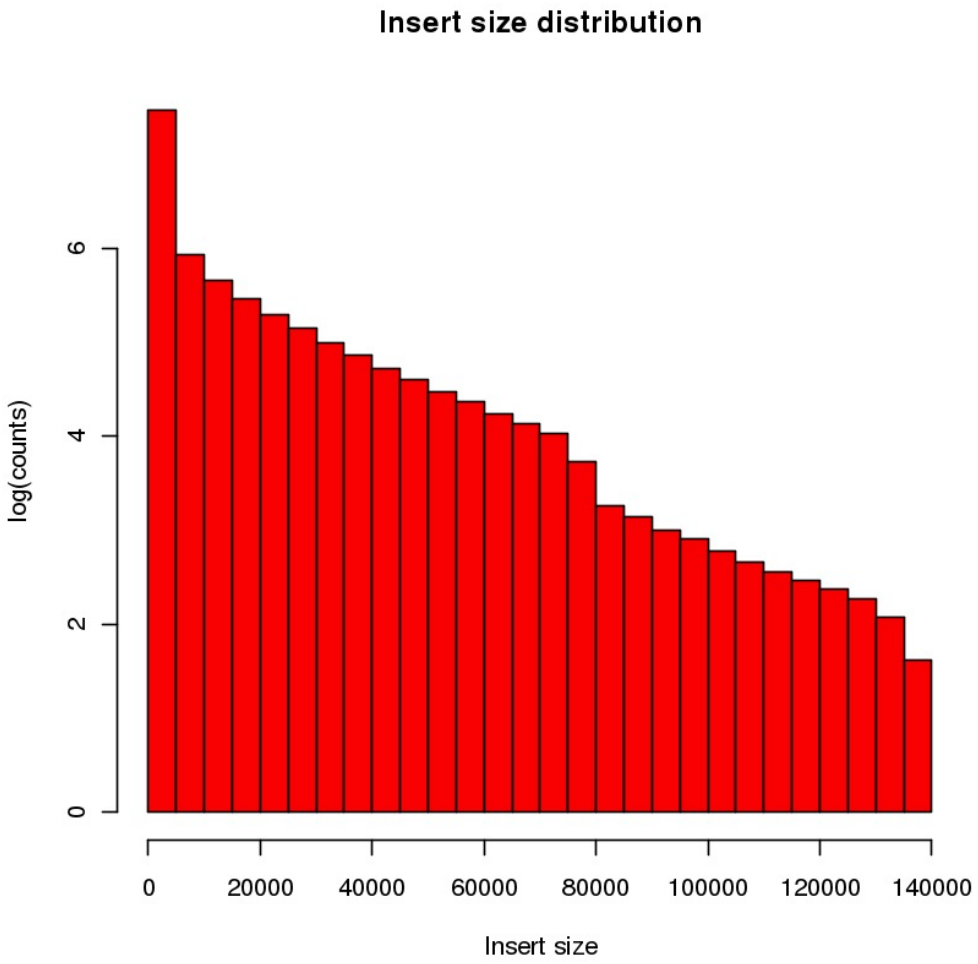
Kadota, M, etc, Multifaceted Hi-C benchmarking: what makes a difference in chromosome-scale genome scaffolding?, 2020, GigaScience

Assessment on insert size (early days)

- Myrtle rust trail data



- Kiwifruit trail data



PhaseGenomics QC method

- A simple QC method (hic_qc.py) for Hi-C libraries to assess long-distance contacts
  - Github: https://github.com/phasegenomics/hic_qc
  - Usage: https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html
- Modifications on aligners
  - Reducing ambiguous mapping
  - No assumptions on shotgun or mate pair libraries – rescue more reads
  - Flag PCR duplicates

PHASE GENOMICS

Hi-C Library QC Report

Genome Scaffolding Sufficiency

Label	Library statistics	Expected values
Subjective Hi-C library judgment	SUFFICIENT	See Judgment
Same strand high-quality* (HQ) read pairs (RPs)	2.18%	> 1.5%
Informative RPs**	9.93%	> 5.0%

*High quality (HQ) read pairs have minimum mapping quality >= 20, maximum edit distance <= 5, and are not duplicates.
**Informative read pairs are read pairs which have MAPQ > 0, are not PCR duplicates, and map to different contigs or >10 Kbp apart.

Metrics Demonstrating Strong Proximity Signal

Label	Library statistics	Expected values
Fraction of HQ RPs >10KB apart (CTGs >10KB)*	2.32%	> 3.0%
Fraction of HQ RPs Intercontig on CTGs >10KB**	10.40%	> 2.5%
Clustering usable HQ reads per contig (CTGs >5KB)***	1292.83	> 600.0

*The proportion of *read pairs that span at least 10kbp*, out of all read pairs that map (a) with high-quality, (b) to the same contig, (c) where that contig is at least 10kbp long.
**The proportion of *read pairs mapping to two different contigs each greater than 10kbp*, out of all read pairs that map with high-quality.
***The average number of usable high-quality read pairs per contig, for contigs greater than 5kbp. Read pairs are "usable" if they map (a) with high-quality, (b) to different contigs, (c) where each of those contigs are greater than 5kbp and (c) both mappings are high-quality.

See below for information on differences between Phase Genomics Hi-C libraries and traditional Hi-C libraries.

Noninformative Read Pair Breakdown

Label	Library statistics	Expected values
Noninformative RPs*	35.84%	<= 50.0%
Duplicate reads	0.00%	< 40.0%
Zero map distance read pairs	2.14%	<= 20.0%
Zero MAPQ reads	26.79%	<= 20.0%
Unmapped reads	0.00%	<= 10.0%

*Note that the sum of informative and noninformative read pairs is not 100% because read pairs with mapping distance between 1 and 10 Kbp are not classified as either informative or noninformative.
Because noninformative reads can belong to more than one category, these numbers may sum to a value larger than the overall noninformative read pair amount at the top of the report.

See below for information on differences between Phase Genomics Hi-C libraries and traditional Hi-C libraries.

Other factors affect Hi-C scaffolding

- The choice of restriction enzyme(s)
  - Generate frequent RE cutting sites – single or multi REs?
- Length of contigs/scaffolds of assembly
  - Assembly statistics – N50, BUSCO, etc…..
- Remove haplotigs and contaminations from your assembly prior to scaffolding
- The choice of scaffolding program and its usage
  - **ALLHIC** – polyploid solution if there are allelic contig information generated in advance
  - **SALSA** – can adapt to multi REs and correct mis-joint scaffolds
  - **HiC-Suite** – Ali's scaffolding strategy based on visualising Hi-C interactions (PFR developed)
  - LACHESIS – was popular, but have not been updated for years

Plant & Food™ Research
Rangahau Ahumāra Kai

Thank you

plantandfood.co.nz

chen.wu@plantandfood.co.nz