

RESEARCH

Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool

Jennifer M Shelton^{*}, Michelle C Coleman, Nic Herndon, Nanyan Lu and Susan J Brown

^{*}Correspondence:

sheltonj@ksu.edu

Department of Biology, Kansas State University, Manhattan, KS, USA

Full list of author information is available at the end of the article

[†]Equal contributor

Abstract

First part title: Text for this section.

Second part title: Text for this section.

Keywords: Genomic physical map; Bionano; Genome scaffolding; Genome validation

Content

Text and results for this section, as per the individual journal's instructions for authors.

Section title

Text for this section ...

Sub-heading for section

Text for this sub-heading ...

Sub-sub heading for section

Text for this sub-sub-heading ...

Sub-sub-sub heading for section Text for this sub-sub-sub-heading ... In this section we examine the growth rate of the mean of Z_0 , Z_1 and Z_2 . In addition, we examine a common modeling assumption and note the importance of considering the tails of the extinction time T_x in studies of escape dynamics. We will first consider the expected resistant population at vT_x for some $v > 0$, (and temporarily assume $\alpha = 0$)

$$E[Z_1(vT_x)] = E\left[\mu T_x \int_0^{v \wedge 1} Z_0(uT_x) \exp(\lambda_1 T_x(v-u)) du\right].$$

If we assume that sensitive cells follow a deterministic decay $Z_0(t) = xe^{\lambda_0 t}$ and approximate their extinction time as $T_x \approx -\frac{1}{\lambda_0} \log x$, then we can heuristically estimate the expected value as

$$E[Z_1(vT_x)] = \frac{\mu}{r} \log x \int_0^{v \wedge 1} x^{1-u} x^{(\lambda_1/r)(v-u)} du$$

$$\begin{aligned}
&= \frac{\mu}{r} x^{1-\lambda_1/\lambda_0 v} \log x \int_0^{v \wedge 1} x^{-u(1+\lambda_1/r)} du \\
&= \frac{\mu}{\lambda_1 - \lambda_0} x^{1+\lambda_1/rv} \left(1 - \exp \left[-(v \wedge 1) \left(1 + \frac{\lambda_1}{r} \right) \log x \right] \right). \quad (1)
\end{aligned}$$

Thus we observe that this expected value is finite for all $v > 0$ (also see [1, 2, 3, 4, 5]).

Availability and requirements

Assembly scripts

Project name: AssembleIrysCluster.pl

Project home page: AssembleIrysCluster scripts are available on Github at https://github.com/i5K-KINBRE-script-share/Irys-scaffolding/tree/master/KSU_bioinfo_lab/assemble_SGE_cluster

Operating system(s): SGE Linux (Gentoo) cluster

Programming language: Perl, Rscript, Bash

License: AssembleIrysCluster.pl is available free of charge to academic and non-profit institutions.

Any restrictions to use by non-academics: Please contact authors for commercial use.

Dependencies: AssembleIrysCluster.pl requires DRMAA job submission libraries. RefAligner and Assembler are also required and can be provided by request by Bionano Genomics <http://www.bionanogenomics.com/>.

Super scaffolding scripts

Project name: stitch.pl

Project home page: stitch scripts are available on Github at https://github.com/i5K-KINBRE-script-share/Irys-scaffolding/tree/master/KSU_bioinfo_lab/stitch

Operating system(s): MAC and LINUX (tested on Gentoo and Ubuntu)

Programming language: Perl, Bash

License: stitch.pl is available free of charge to academic and non-profit institutions.

Any restrictions to use by non-academics: Please contact authors for commercial use.

Dependencies: stitch.pl requires BioPerl. RefAligner and Assembler are also required between iterations and can be provided by request by Bionano Genomics <http://www.bionanogenomics.com/>.

Map summary scripts

Project name: cmap_stats.pl and xmap_stats.pl

Project home page: all scripts are available on Github at https://github.com/i5K-KINBRE-script-share/Irys-scaffolding/tree/master/KSU_bioinfo_lab/map_editing

Operating system(s): MAC and LINUX (tested on Gentoo and Ubuntu)

Programming language: Perl

License: cmap_stats.pl and xmap_stats.pl are available free of charge to academic and non-profit institutions.

Any restrictions to use by non-academics: Please contact authors for commercial use.

Dependencies: cmap_stats.pl and xmap_stats.pl have no dependencies.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Text for this section ...

Acknowledgements

Text for this section ...

References

1. Koonin, E.V., Altschul, S.F., Bork, P.: Brca1 protein products: functional motifs. *Nat Genet* **13**, 266–267 (1996)
2. Kharitonov, S.A., Barnes, P.J.: Clinical Aspects of Exhaled Nitric Oxide. in press
3. Zvaifler, N.J., Burger, J.A., Marinova-Mutafchieva, L., Taylor, P., Maini, R.N.: Mesenchymal cells, stromal derived factor-1 and rheumatoid arthritis [abstract]. *Arthritis Rheum* **42**, 250 (1999)
4. Jones, X.: Zeolites and synthetic mechanisms. In: Smith, Y. (ed.) *Proceedings of the First National Conference on Porous Sieves: 27-30 June 1996; Baltimore*, pp. 16–27 (1996). Stoneham: Butterworth-Heinemann
5. Margulis, L.: *Origin of Eukaryotic Cells*. Yale University Press, New Haven (1970)

Figures

Figure 1 Data analysis steps. (1) Autonoise converts TIFF images of molecules to (2) BNX text files. (3) Genome FASTA is *in silico* labeled with fa2cmap_multi producing (4) a CMAP. (5) Assemblelryscluster uses *in silico* CMAPs, BNX files and estimated genome size to (6) adjust molecule stretch and set assembly parameters. (7) Assembler produces (8) a BNG CMAP. (9) RefAligner aligns the BNG CMAP to the *in silico* CMAP producing (10) an XMAP. (11) XMAP, *in silico* CMAP and BNG CMAP are used by stitch to produce a super scaffolded FASTA. (13) Until no more super scaffolds are created the super scaffold FASTA is *in silico* labeled with fa2cmap_multi producing (14) a CMAP that is aligned to (9) the BNG CMAP and steps 10-15 are iterated.

Figure 2 Assembly workflow for assemble_SGE_cluster.pl. (A) The Irys instrument produces tiff files that are converted into BNX text files. (B) Each chip produces one BNX file for each of two flowcells. (C) BNX files are split by scan and aligned to the sequence reference. Stretch (bases per pixel) is recalculated from the alignment. (D) Quality check graphs are created for each pre-adjusted flowcell BNX. (E) Adjusted flowcell BNXs are merged. (F) The first assemblies are run with a variety of p-value thresholds. (G) The best of the first assemblies (red oval) is chosen and a version of this assembly is produced with a variety of minimum molecule length filters.

Figure 3 Steps of the stitch.pl algorithm. BNG CMAPs (blue) are shown aligned to *in silico* CMAPs (green). Alignments are indicated with grey lines. CMAP orientation for *in silico* CMAPs is indicated with a "+" or "-" for positive or negative orientation respectively. (A) The *in silico* CMAP is aligned as the reference. (B) The alignment is inverted and used as input for stitch.pl. (C) The alignments are filtered based on alignment length (purple) relative to total possible alignment length (black) and confidence. Here assuming all alignments have a high confidence score and the minimum percent aligned is 30% two alignments fail for aligning over less than 30% of the potential alignment length for that alignment. (D) Filtering produces an XMAP of high quality alignments with short (local) alignments removed. (E) High quality scaffolding alignments are filtered for longest and highest confidence alignment for each *in silico* CMAP. Third alignment (unshaded) is filtered because the second alignment is the longest alignment for *in silico* CMAP 2. (F) Passing alignments are used to super scaffold (captured gaps indicated in dark green). (G) Stitch is iterated and additional super scaffolding alignments are found using second best scaffolding alignments. (H) Iteration takes advantage of cases where *in silico* CMAPs scaffold BNG CMAPs as *in silico* CMAP 2 does. Stitch is run iteratively until super scaffolding alignments are found.

Figure 4 Putative haplotypes assembled as BNG CMAPs. (A) Two BNG CMAPs (blue with molecule coverage shown in dark blue) align to the *in silico* CMAP of scaffold 131 (green with contigs overlaid as translucent colored squares). (B and C) Both BNG CMAPs are shown (blue) with molecule pileups (yellow). Both BNG CMAPs have similar label patterns except within the lower coverage region indicated with a black square.

Figure 5 Histogram of gap lengths in Tcas5.1. Positive and negative gaps lengths for Tcas5.1 added to the automated output of stitch.pl based on filtered scaffolding alignments. The majority of gap lengths added by stitch.pl, 66, were positive (red). The remaining 26 gaps had negative lengths (purple).

Figure 6 Extremely small negative gap length for in silico CMAP of scaffold 81. Two XMAP alignments for *in silico* CMAP of scaffold 81 are shown. BNG CMAPs (blue with molecule coverage shown in dark blue) align to the *in silico* CMAPs of scaffolds (green with contigs overlaid as translucent colored squares). Scaffolds 79-83 were placed within ChLG 5 and scaffolds 99-103 were placed with ChLG 7 by the *Tribolium* genetic map. (A) Half of the *in silico* CMAP of scaffold 81 aligns with its assigned ChLG (black arrow). (B) The other half aligns with ChLG 7 (red arrow) producing a negative gap length smaller than -20 (kb). The alignment that places scaffold 81 with ChLG 7 disagrees with the genetic map and was manually rejected for Tcas5.2.

Table 1 Assembly Results. Assembly metrics for Tcas5.0 (the starting scaffolded FASTA), the *in silico* CMAP, the BNG CMAP of assembled molecules and the final super scaffolded FASTA (Tcas5.2) produced using stitch.pl for the *Tribolium* genome.

	N50 (Mb)	Number	Cumulative Length (Mb)
Genome FASTA	1.16	2240	160.74
<i>in silico</i> CMAP	1.20	223	152.53
BNG CMAP	1.35	216	200.47
Super scaffold FASTA	4.46	2150	165.92

Table 2 Alignment of BNG assembly to reference genome. Breadth of alignment coverage (non-redundant alignment), length of total alignment (including redundant alignments) and percent of CMAP covered (non-redundantly) were calculated for the *in silico* CMAP and the BNG CMAP of the *Tribolium* genome the using xmap_stats.pl.

	Breadth of alignment coverage (Mb)	Length of total alignment (Mb)	Percent of CMAP aligned
<i>in silico</i> CMAP from FASTA	124.04	132.40	81
BNG CMAP	131.64	132.34	67

Table 3 Chromosome linkage groups before and after super scaffolding. The number of scaffolds and contigs in the Tcas5.0 ChLG bins and the number of super scaffolds, scaffolds and contigs in the Tcas5.2 ChLG bins. The number of scaffolds or contigs that were unplaced in Tcas5.0 and placed with a ChLG in Tcas5.2 is also listed.

Chromosome linkage group (ChLG)	Tcas5.0 scaffolds	Tcas5.2 super scaffolds	Unplaced scaffolds added in Tcas5.2
X	13	2	2
2	18	10	1
3	29	20	4
4	6	2	2
5	17	4	1
6	12	6	6
7	15	6	0
8	14	8	1
9	21	9	0
10	12	11	2
Total	157	78	19

Tables

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.