riboSeed: it's whats for dinner

Nicholas R Waters,[1,2*] Florence Abram,[1] Ashleigh Holmes,[2] Fiona Brennan,[1,3] and Leighton Pritchard[2]

[1]National University of Ireland, Galway
[2]The James Hutton Institute, Dundee, Scotland
[3]Teagasc, Johnstown Castle, Wexford

[*]To whom hate mail should be addressed; E-mail: n.waters4@nuigalway.ie.

The vast majority of bacterial genome sequencing has been performed using Illumina short reads. Because of the inherent difficulty of resolving repeated regions with short reads alone, only 10% of sequencing projects have resulted in a closed genome. The most ubiquitous repeated regions are those coding for ribosomal operons (rDNAs), which can occur in a bacterial genome between 1 and 15 times. Here, we show that the genomic context in which rDNAs occur is conserved across taxa. We demonstrate that within a single genome, the regions flanking the rDNAs are unique. By utilizing the conserved nature of rDNAs across taxa and the uniqueness of their flanking regions, targeted pseudocontigs can be generated by iteratively assembling reads mapping to a references rDNAs, and these pseudocontigs can be used to assemble across rDNAs. This method, implemented as riboSeed, reduces the number of contigs in the assembly, and when used in conjunction with other genome polishing tools, can result in closure of a genome.

Background

Sequencing bacterial genomes has become much more cost effective and convenient, but the number of complete, closed bacterial genomes remains a small fraction of the total number sequenced (Table 1). The length of short reads is increasing, but even with the advent of new long-read technologies, bacterial assembly remains a major bottleneck [**?**, 17]. Although draft genomes are often of very high quality and suited for many types of analysis, researchers are forced to choose between working with these draft genomes (and the inherent potential loss of data), or spending time and resources polishing the genome with some combination of in silico tools, PCR, optical mapping, resequencing, or hybrid sequencing [17]. Many in silico genome finishing tools are available, and we summarise several of these in Table 2.

Table 1: NCBI Genome Assemblies of Bacteria

Date	Total	Complete genome	Chromosome	Scaffold	Contig
January 4th, 2017	85799	6255	1143	39972	38429
May 17th, 2017	96849	7212	1254	42839	43899

Table 2: Alternative in silico genome polishing tools

Tool	Reference	Method Summary
Boetzer, et als GapFiller	Boetzer2012 [3]	utilize paired end and other short read information to close contig junctions
GapCloser/IMAGE	Luo2012, Tsai2010 [12, 20]	iteratively use reads that are mapped to contigs, to close contig junctions
CloG	Yang2011 [25]	use trimmed de novo contigs in hybrid assembly followed by a stitching algorithm
FGap	Piro2014,Guizelini2016 [10, 18]	use BLAST to find potential gap closures from alternate assemblies, libraries or references.
GFinisher	Guizelini2016 [10]	use GC-skew to refine assemblies
Nadalin, et als GapFiller	Nadalin2012 [16]	local assembler using a hash-based method to produce long-reads from paired end sequencing data, which can then be used in a de novo assembly.
CONTIGuator	Galardini2011 [9]	use contigs from a de novo assembly along with one or more reference sequences to generate a contig map and PCR primer sets to validate in the lab.
Konnector	Vandervalk2015 [21]	use paired end reads to make long reads to be used in a Bloom filter representation of a de Bruijn graph
MapRepeat	Mariano2015 [13]	use a directed scaffolding method to fill in rDNA gaps, but limited to Ion Torrent reads, and affected by inversions between rDNAoperons [14]
GRabB	Brankovics2016 [4]	Selective assembly tandem rDNA clusters and mitochondria

The number of Illumina entries in NCBIs Sequence Read Archive (SRA) [**?**] continues to outnumber all other technologies combined by about an order of magnitude (sup table 3). Draft assemblies from these datasets have systematic problems common to short read datasets, namely gaps in the sequences due to the difficulty of resolving assemblies of repeated regions [19, 23]. Therefore it may be possible to obtain additional sequence information from short read datasets in the SRA, and improve on current assemblies, by improving the ability to resolve assemblies through repeated regions.

The most ubiquitous repeated regions are those coding for ribosomal RNAs. Sequencing of the 16S ribosomal region is widely used to identify bacteria and explore microbial community dynamics [5, 6, 22, 24], as the region is conserved within taxa, yet retains enough variability to act as a bacterial fingerprint to separate clades informatively. However, the 16S, 23S, and 5S ribosomal subunit coding regions (rDNA) are often present multiple times in a single genome, commonly exhibiting polymorphism [**?**, 7, 11, 15]. These long, inexactly repeated regions [1, 2] are problematic for short-read genome assembly. Other large repeated regions also exist, but none as pervasive as rDNAs, as ribosomes are essential for cell function. As rDNAs are frequently used as a sequence marker for taxonomic classification, resolving their genomic context could increase the accuracy of community analysis in cases where polymorphic rDNAs from the same strain result in multiple operational taxonomic units. as ribosomes are essential for cell function. Although a PCR-based method has been developed [8], no effective in silico method established to resolve the repeats introduced by rDNAs when assembling Illumina data. We present here an in silico method, riboSeed, to capitalize on the genomic conservation of these regions within a taxon, to improve resolution of these normally difficult

	$w = 8$			$w = 16$			$w = 32$		
	$t = 0$	$t = 1$	$t = 2$	$t = 0$	$t = 1$	$t = 2$	$t = 0$	$t = 1$	
$dir = 1$									
c	0.0790	0.1692	0.2945	0.3670	0.7187	3.1815	- 1.0032	-1.7104	-
c	- 0.8651	50.0476	5.9384	-9.0714	297.0923	46.2143	4.3590	34.5809	
c	124.2756	- 50.9612	-14.2721	128.2265	-630.5455	-381.0930	-121.0518	-137.1210	-2
$dir = 0$									
c	0.0357	1.2473	0.2119	0.3593	-0.2755	2.1764	-1.2998	-3.8202	
c	-17.9048	-37.1111	8.8591	-30.7381	-9.5952	-3.0000	-11.1631	-5.7108	-
c	105.5518	232.1160	-94.7351	100.2497	141.2778	-259.7326	52.5745	10.1098	-1

Table 3: Caption

regions and provide a means to benefit from unexploited information in the SRA.

riboSeed is most similar to GRabB, the method of Brankovics et. al [4]. for assembling mitochondrial and rDNA regions, as both use targeted assembly. However, GRabB does not make inferences about how many clusters are present in the genome, or take advantage of the genomic context of the rDNA cluster. In riboSeed, genomic context is resolved by exploiting both rDNA regions and the flanking regions, harnessing the unique characteristics of the broader rDNA region to improve bacterial genome assembly.

The success of riboSeed hinges on two core ideas: (1) that although repeated bacterial rDNA coding sequences within a single genome are nearly identical, their flanking regions (ie, the neighboring locations within the genome), are distinct, and (2) that the genomic context of rDNAs in relation to the rest of the genome is conserved within a taxonomic grouping. Conservation of flanking regions may be a byproduct of the evolutionary trend towards conservation of functional ribosomes.

Briefly, riboSeed uses rDNA regions from the closest completely sequenced reference genome to generate pseudocontigs. Pseudocontigs are produced by iteratively mapping short reads to rDNA regions from the reference, and then producing subassemblies from the reads mapping to the rDNA regions, generating what are essentially long reads. These are seeded into the raw reads for the final assembly, which we refer to as de fere novo (meaning 'starting from almost nothing'). While there are usually many differences between the reference genome and that of the sequenced isolate, riboSeed mitigates the influence these differences have on the resulting assembly by utilizing this subassembly procedure, and by only relying on the references rDNA regions as opposed to the rest of the genome. Thus, the characteristics of the rDNA flanking regions allow us to leverage genomic architecture to assemble across rDNAs.

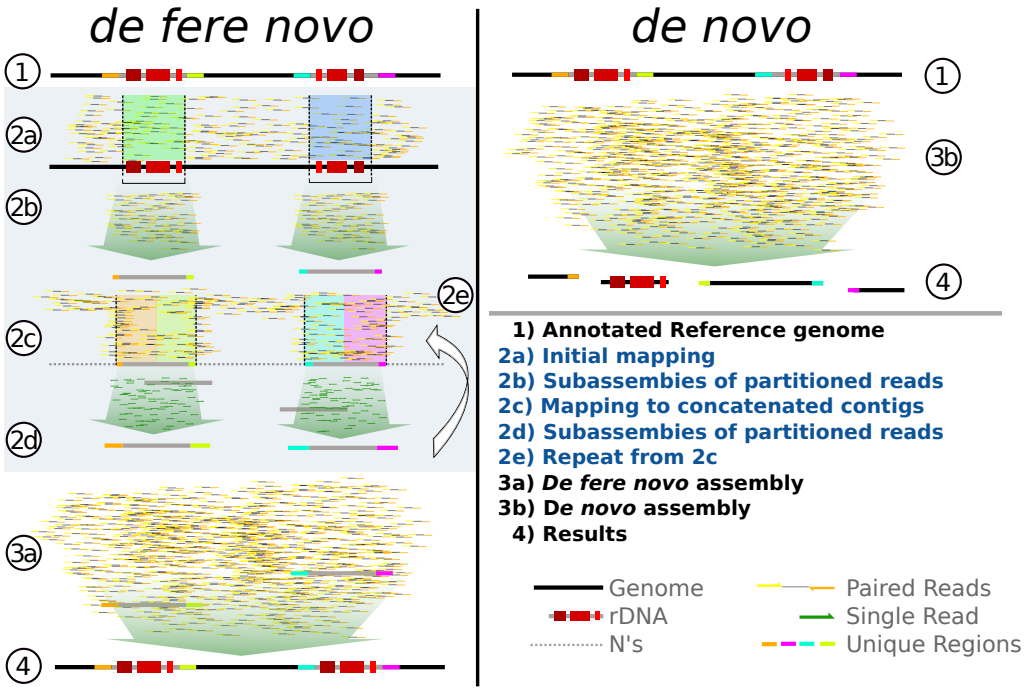

Figure 1: a nice plot

References

[1] Saumya Agrawal and Austen R. D. Ganley. Complete Sequence Construction of the Highly Repetitive Ribosomal RNA Gene Repeats in Eukaryotes Using Whole Genome Sequence Data. In *Methods in molecular biology (Clifton, N.J.)*, volume 1455, pages 161–181. 2016.

[2] Can Alkan, Saba Sajjadian, and Evan E Eichler. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1), 2011.

[3] Marten Boetzer, Walter Pirovano, DR Zerbino, E Birney, JT Simpson, K Wong, SD Jackman, JE Schein, SJ Jones, I Birol, R Li, W Fan, G Tian, H Zhu, L He, J Cai, Q Huang, Q Cai, B Li, Y Bai, Z Zhang, Y Zhang, W Wang, J Li, F Wei, H Li, M Jian, J Li, Z Zhang, R Nielsen, D Li, W Gu, Z Yang, Z Xuan, OA Ryder, FC Leung, Y Zhou, J Cao, X Sun, Y Fu, M Boetzer, CV Henkel, HJ Jansen, D Butler, W Pirovano, A Dayarian, TP Michael, AM Sengupta, IJ Tsai, TD Otto, M Berriman, B Langmead, C Trapnell, M Pop, SL Salzberg, H Li, R Durbin, H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, DR Kelley, MC Schatz, SL Salzberg, SL Salzberg, AM Phillippy, A Zimin, D Puiu, T Magoc, S Koren, TJ Treangen, MC Schatz, AL Delcher, M Roberts, G Marçais, M Pop, JA Yorke, I MacCallum, D Przybylski, S Gnerre, J Burton, I Shlyakhter, A Gnirke, J Malek, K McKernan, S Ranade, TP Shea, L Williams, S Young, C Nusbaum, and DB Jaffe. Toward almost closed genomes with GapFiller. *Genome Biology*, 13(6):R56, 2012.

[4] Balázs Brankovics, Hao Zhang, Anne D. van Diepeningen, Theo A. J. van der Lee, Cees Waalwijk, G. Sybren de Hoog, C Hahn, L Bachmann, B Chevreux, RE Green, AS Malaspinas, J Krause, AW Briggs, PLF Johnson, C Uhler, IJ Tsai, TD Otto, M Berriman, D Hernandez, P François, L Farinelli, M Osterås, J Schrenzel, D Hernandez, R Tewhey, JB Veyrieras, L Farinelli, M Østerås, P François, DR Zerbino, E Birney, GSC Slater, E Birney, L Guo, L Han, L Yang, H Zeng, D Fan, Y Zhu, G Fourie, NA van der Merwe, BD Wingfield, M Bogale, B Tudzynski, MJ Wingfield, DM Hillis, and MT Dixon. GRAbB: Selective Assembly of Genomic Regions, a New Niche for Genomic Research. *PLOS Computational Biology*, 12(6):e1004753, jun 2016.

[5] R. J. Case, Y. Boucher, I. Dahllof, C. Holmstrom, W. F. Doolittle, and S. Kjelleberg. Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Applied and Environmental Microbiology*, 73(1):278–288, jan 2007.

[6] Jill E Clarridge and III. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, 17(4):840–62, table of contents, oct 2004.

[7] Tom Coenye and Peter Vandamme. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS microbiology letters*, 228:45–49, 2003.

[8] Alexander W Eastman, Ze-Chun Yuan, Jonathan H Badger, and J Craig Venter. Development and validation of an rDNA operon based primer walking strategy applicable to de novo bacterial genome finishing. *Frontiers in microbiology*, 5, 2015.

[9] Marco Galardini, Emanuele G Biondi, Marco Bazzicalupo, and Alessio Mengoni. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code for Biology and Medicine*, 6(11), 2011.

[10] Dieval Guizelini, Roberto T Raittz, Leonardo M Cruz, Emanuel M Souza, Maria B R Steffens, and Fabio O Pedrosa. GFinisher: a new strategy to refine and finish bacterial genome assemblies. *Nature Scientific Reports*, 6, 2016.

[11] Oksana Lukjancenko, Trudy M Wassenaar, and David W Ussery. Comparison of 61 Sequenced Escherichia coli Genomes. *Microbial Ecology*, 60, 2010.

[12] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, dec 2012.

[13] Diego CB Mariano, Felipe L Pereira, Preetam Ghosh, Debmalya Barh, Henrique CP Figueiredo, Artur Silva, Rommel TJ Ramos, and Vasco AC Azevedo. MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. *Bioinformatician*, 11(6):276–279, 2015.

[14] Diego César Batista Mariano, Thiago De Jesus Sousa, Felipe Luiz Pereira, Flávia Aburjaile, Debmalya Barh, Flávia Rocha, Anne Cybelle Pinto, Syed Shah Hassan, Tessália Diniz, Luerce Saraiva, Fernanda Alves Dorella, Alex Fiorini De Carvalho, Carlos Augusto Gomes Leal, Henrique César, Pereira Figueiredo, Artur Silva, Rommel Thiago, Jucá Ramos, Vasco Ariston, and Carvalho Azevedo. Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. *BMC genomics*, 17, 2016.

[15] Claudia Moreno, Jaime Romero, and Romilio T Espejo. Polymorphism in repeated 16S rRNA genes is a common property of type strains and environmental isolates of the genus Vibrio. *Microbiology*, 148:1233–1239, 2002.

[16] Francesca Nadalin, Francesco Vezzi, and Alberto Policriti. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC bioinformatics*, 13:12–14, 2012.

[17] Niranjan Nagarajan, Christopher Cook, MariaPia Di Bonaventura, Hong Ge, Allen Richards, Kimberly A Bishop-Lilly, Robert Desalle, Timothy D Read, Mihai Pop, J Parkhill, C Fraser, J Eisen, K Nelson, IT Paulsen, SL Salzberg, E Branscomb, P Predki, H Tettelin, D Radune, S Kasif, H Khouri, SL Salzberg, Mariapia Di MD Bonaventura, Robert Desalle, Mihai Pop, Niranjan Nagarajan, DH Figurski, DH Fine, JB Kaplan, PJ Planet, O Khairat, P Chen, Christopher Cook, A Stewart, Niranjan Nagarajan, D Sommer, Mihai Pop, B Thomason, M Kiley, S Lentz, N Nolan, S Sozhamannan, A Sulakvelidze, A Mateczun, L Du, M Zwick, Timothy D Read, Niranjan Nagarajan, Timothy D Read, Mihai Pop, MJ Chaisson, PA Pevzner, E Myers, G Sutton, A Delcher, I Dew, D Faulo, M Flanigan, S Kravitz, C Mobarry, K Reinert, K Remington, E Anson, S Andersson, A Zomorodipour, J Andersson, T Sicheritz-Ponten, U Alsmark, R Podowski, A Naslund, A Eriksson, H Winkler, C Kurland, K Jo, D Dhingra, T Odijk, J de Pablo, M Graham, R Runnheim, D Forrest, D Schwartz, Z Mulyukov, PA Pevzner, Mihai Pop, D Sommer, A Delcher, SL Salzberg, Mihai Pop, DC Richter, SC Schuster, DH Huson, F Zhao, F Zhao, T Li, DA Bryant, Niranjan Nagarajan, Mihai Pop, J Miller, A Delcher, S Koren, E Venter, B Walenz, A Brownley, J Johnson, K Li, C Mobarry, G Sutton, W Jeck, J Reinhardt, D Baltrus, MT Hickenbotham, V Magrini, ER Mardis, JL Dangl, CD Jones, B Chevreux, T Wetter, S Suhai, A Samad, E Huff, W Cai, D Schwartz, P Latreille, S Norton, B Goldman, J Henkhaus, N Miller, B Barbazuk, HB Bode, C Darby, Z Du, S Forst, S Gaudriault, B Goodner, H Goodrich-Blair, S Slater, M Waterman, SL Salzberg, D Sommer, D Puiu, V Lee, Hong Ge, Allen Richards, Kimberly A Bishop-Lilly, Robert Desalle, Timothy D Read, and Mihai Pop. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics*, 11(1):242, 2010.

[18] Vitor C Piro, Helisson Faoro, Vinicius A Weiss, Maria Br Steffens, Fabio O Pedrosa, Emanuel M Souza, and Roberto T Raittz. FGAP: an automated gap closing tool. *BMC Research Notes*, 7, 2014.

[19] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1), 2013.

[20] Isheng J Tsai, Thomas D Otto, and Matthew Berriman. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, 11, 2010.

[21] Benjamin P Vandervalk, Chen Yang, Zhuyi Xue, Karthika Raghavan, Justin Chu, Hamid Mohamadi, Shaun D Jackman, Readman Chiu, René L Warren, and Inanç Birol. Konnector v2.0: pseudo-long reads from paired-end sequencing data. *From IEEE International Conference on Bioinformatics and Biomedicine*, pages 2–5, 2015.

[22] W G Weisburg, S M Barns, D A Pelletier, and D J Lane. 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology*, 173(2):697–703, jan 1991.

[23] N. Whiteford, Niall Haslam, Gerald Weber, Adam Prügel-Bennett, Jonathan W. Essex, Peter L. Roach, Mark Bradley, and Cameron Neylon. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, 33(19):e171–e171, oct 2005.

[24] Carl R Woese, Otto Kandlert, and Mark L Wheelis. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Evolution*, 87:4576–4579, 1990.

[25] Xing Yang, Daniel Medvin, Giri Narasimhan, Deborah Yoder-Himes, and Stephen Lory. CloG: a pipeline for closing gaps in a draft assembly using short reads. *2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pages 202–207, 2011.