**Sequence analysis**

**Roary: Rapid large-scale prokaryote pan genome analysis**

Andrew J. Page\(^1\)*, Carla A. Cummins\(^1\), Martin Hunt\(^1\), Vanessa K. Wong\(^1\)\(^2\), Sandra Reuter\(^2\), Matthew T. G. Holden\(^3\), Maria Fookes\(^1\), Jacqueline A. Keane\(^1\) and Julian Parkhill\(^1\)

\(^1\)Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

\(^2\)Department of Medicine, University of Cambridge, UK.

\(^3\)School of Medicine, North Haugh, St Andrews, UK.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

---

**ABSTRACT**

**Summary:** A typical prokaryote population sequencing study can now consist of hundreds or thousands of isolates. Interrogating these datasets can provide detailed insights into the genetic structure of prokaryotic genomes. We introduce Roary, a tool that rapidly builds large-scale pan genomes, identifying the core and dispensable accessory genes. Roary makes construction of the pan genome of thousands of prokaryote samples possible on a standard desktop without compromising on the accuracy of results. Using a single CPU Roary can produce a pan genome consisting of 1000 isolates in 4.5 hours using 13 GB of RAM, with further speedups possible using multiple processors.

**Availability and implementation:** Roary is implemented in Perl and is freely available under an open source GPLv3 license from http://sanger-pathogens.github.io/Roary

**Contact:** roary@sanger.ac.uk

**Supplementary information:** Supplementary data are available at Bioinformatics online.

---

**1 INTRODUCTION**

The term microbial pan genome was first used in 2005 (Medini et al., 2005) to describe the union of genes shared by genomes of interest (Vernikos et al., 2014). Since then, availability of microbial sequencing data has grown exponentially. Aligning whole genome sequenced isolates to a single reference genome can fail to incorporate non-reference sequences. By using *de novo* assemblies, non-reference sequence can also be analysed. Microbial organisms can rapidly acquire genes from other organisms that can increase virulence or promote antimicrobial drug resistance (Medini et al., 2005). Gaining a better picture of the conserved genes of an organism, and the dispensable accessory genome, can lead to a better understanding of key processes such as selection and evolution. The construction of a pan genome is NP-hard (Nguyen et al., 2014) with additional difficulties from real data due to contamination, fragmented assemblies and poor annotation. Therefore any approach must employ heuristics in order to produce a pan genome (reviewed in (Vernikos et al., 2014)). The most complete standalone pan genome tools are: PanOCT (Fouts et al., 2012), which uses conserved gene neighborhood in addition to homology to accurately place proteins into orthologous clusters; and PGAP which takes annotated assemblies, performs an all-against-all BLAST, clusters the results, and produces a pan genome (Zhao et al., 2012).

PanOCT and PGAP require an all-against-all comparison using BLAST, which means the running time grows non-linearly with the size of input data and is computationally infeasible with large datasets. They also have non-linear memory requirements, quickly exceeding the RAM available in high performance servers for large datasets. We have developed a method to generate the pan genome of a set of related prokaryotic isolates. It works with thousands of isolates in a computationally feasible time, beginning with annotated fragmented *de novo* assemblies. We address the computational issues by performing a rapid clustering of highly similar sequences, which can reduce the running time of BLAST substantially, and carefully manage RAM usage so that it increases linearly, both of which make it possible to analyse datasets with thousands of samples using commonly available computing hardware.

---

**2 DESCRIPTION**

The input to Roary is one annotated assembly per sample in GFF3 format (Stein, 2013), such as that produced by Prokka (Seemann, 2014), where all samples are from the same species. Coding regions are extracted from the input and converted to protein sequences, filtered to remove partial sequences and iteratively pre-clustered with CD-HIT (Fu et al., 2012). This results in a substantially reduced set of protein sequences. An all-against-all comparison is performed with BLASTP on the reduced sequences with a user defined percentage sequence identity (default 98%). Low complexity regions are masked out (Camacho et al., 2008). Sequences are then clustered with MCL (Enright et al., 2002) and finally the pre-clustering results from CD-HIT are merged together with the results of MCL. Using conserved gene neighbourhood information, homologous groups containing paralogs are split into groups of true orthologs. A graph is constructed of the relationships of the clusters based on the order of occurrence in the input sequences, allowing for the clusters to be ordered and thus providing context for each gene. Full details of the method and outputs are provided in the Supplementary Material.

---

\(^*\)To whom correspondence should be addressed.
Roary scales consistently as more samples are added, making it feasible to process large data sets on a standard desktop computer within a few hours. We conducted similar experiments with more diverse datasets including *S. pneumonia*, *S. aureus*, and *Y. enterocolitica* and the results show/exhibit similar speed-ups as shown in Sup Figs 7-8.

Table 2. Comparison of pan genome applications using real *S. typhi* data (ERP001718).

<table>
<thead>
<tr>
<th>Samples</th>
<th>Core*</th>
<th>Total</th>
<th>RAM (mb)</th>
<th>Wall time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>PGAP</td>
<td>4545</td>
<td>4929</td>
<td>569</td>
</tr>
<tr>
<td></td>
<td>PanOCT</td>
<td>4544</td>
<td>4936</td>
<td>663</td>
</tr>
<tr>
<td></td>
<td>Roary</td>
<td>4459</td>
<td>4871</td>
<td>179</td>
</tr>
<tr>
<td>24</td>
<td>PGAP</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PanOCT</td>
<td>4522</td>
<td>4991</td>
<td>5313</td>
</tr>
<tr>
<td></td>
<td>Roary</td>
<td>4436</td>
<td>4941</td>
<td>349</td>
</tr>
<tr>
<td>1000</td>
<td>PGAP</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PanOCT</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Roary</td>
<td>4016</td>
<td>9201</td>
<td>13752</td>
</tr>
</tbody>
</table>

* Core is defined as a gene being in at least 99% of samples, which allows for some assembly errors in very large datasets. Where there are no results, the application failed to complete within 2 days or used more than 60GB of RAM. The first column is the number of unique *Salmonella Typhi* genomes in the input set.

4 DISCUSSION

We have shown that Roary can construct the pan genomes of large collections of bacterial genomes using a desktop computer, where it was not previously computationally possible with other methods. Further speedups in running time are possible by providing more processors to Roary. On simulated data Roary is the only application to correctly identify all clusters. This increased accuracy comes from using the context provided by conserved gene neighbourhood information. Roary scales well on large real datasets, identifying large numbers of core genes, even in the presence of a varied open pan genome.

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust (grant number WT 098051).

REFERENCES


Roary: Rapid large-scale prokaryote pan genome analysis


