Crass: identification and assembly of CRISPR loci using unassembled metagenomic sequencing reads

Michael Imelfort, Connor T. Skennerton, Gene W. Tyson

Australian Centre for Ecogenomics, School of Chemisty and Molecular Biosciences & Advanced Water Management Centre, The University of Queensland, St. Lucia, Queensland, Australia

Background

- Clustered Regularly Interspersed Short Palindromic Repeats (CRISPR) are an adaptive bacterial and archaeal immune system.
- Analysis of CRISPRs in metagenomic datasets is hampered by their repetative nature
- Modern genome assemblers can't handle the diversity seen in metagenomic datasets - CRISPRs are difficult to assemble!
- Current CRISPR identification programs are designed for completed genomes
- We have developed a specific tool, Crass, for discovering CRISPRs in metagenomes

A New Way of Visualising CRISPRs

- Crass represents spacer arangements as a graph rather than an alignment

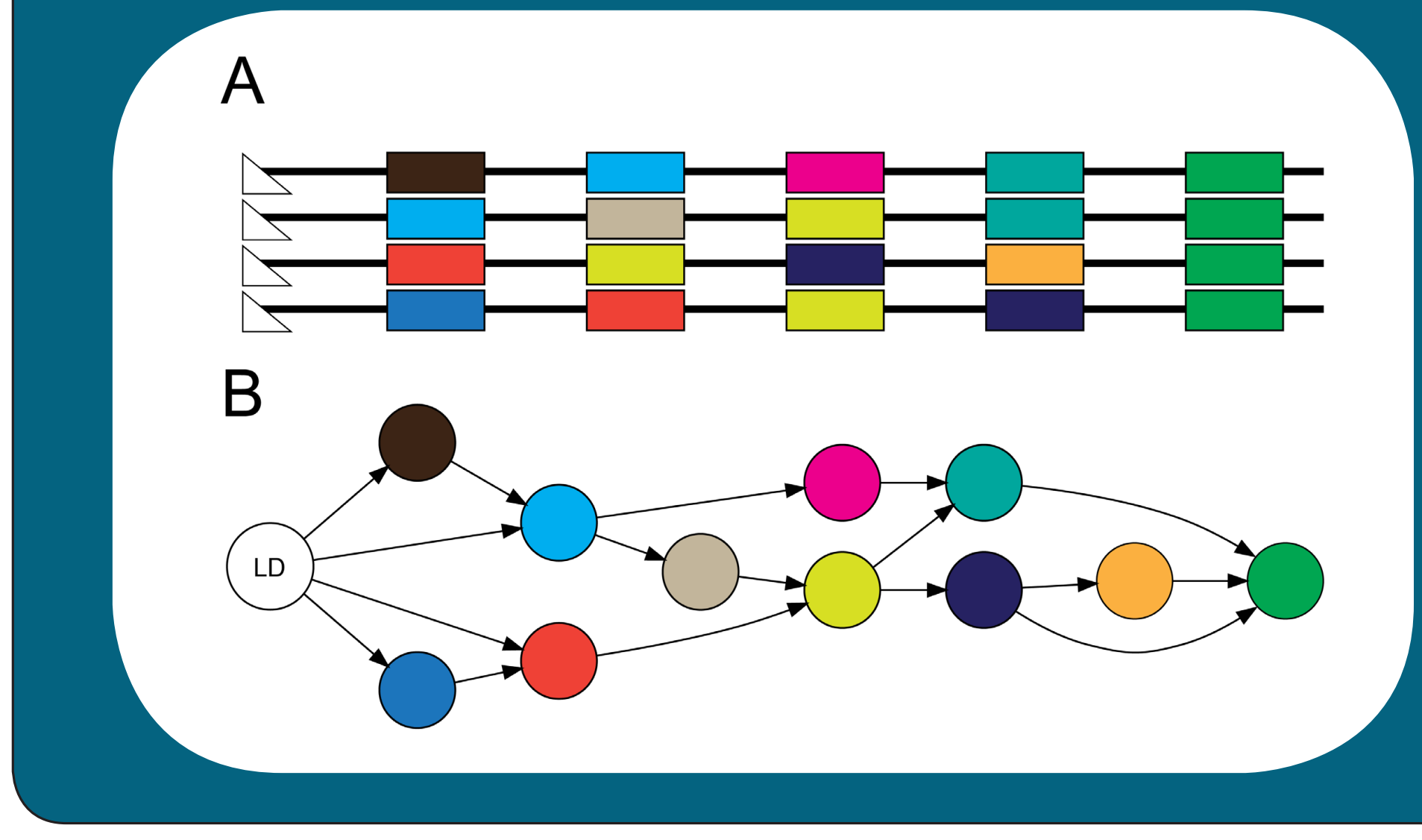
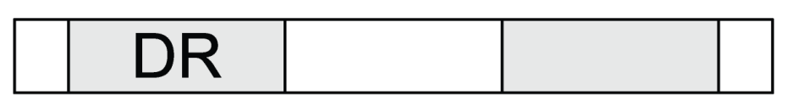


Figure 1: Comparison between CRISPR representations. (A) Traditional alignment approach, each strain is represented as a row, beginning from the leader sequence. (B) The graph approach used in Crass, individual strains are merged into a single path where they have conserved spacers and diverge into separate paths where there is strain variation.

Algorithm Design

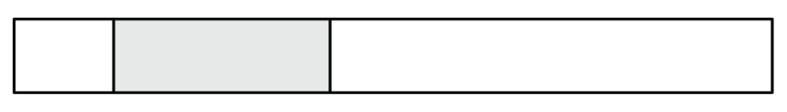
Stage 1: Direct Repeat Search

1. Primary search:
Identify sequences with repeated short sub-strings

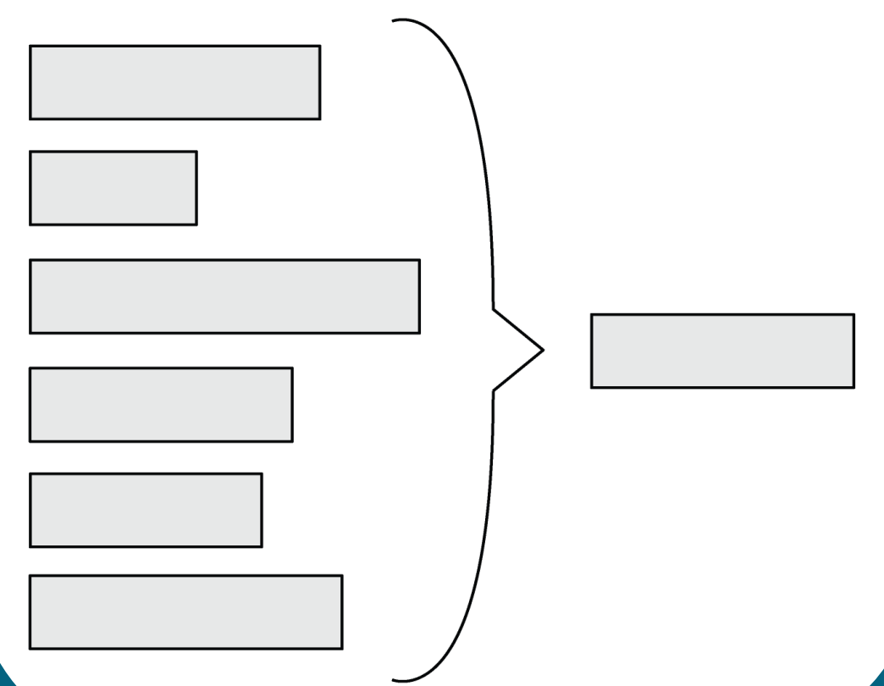


Create a database of potential DRs

2. Secondary search:
Identify sequences with one copy of any DR in the database

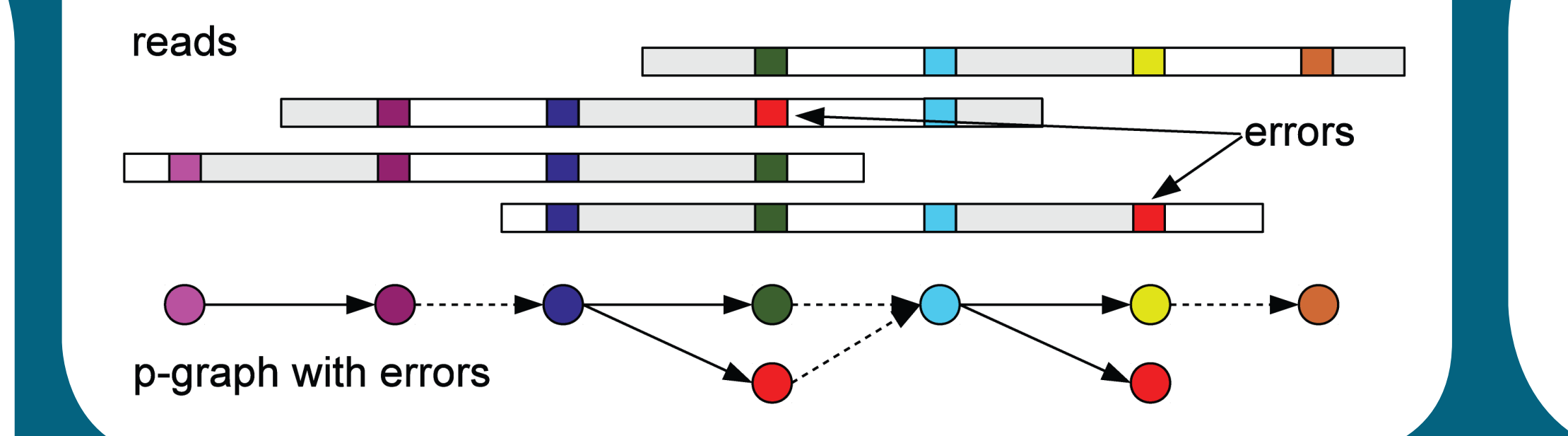


3. DR clustering:
Group similar DR types together based on single-linkage clustering and kmer composition

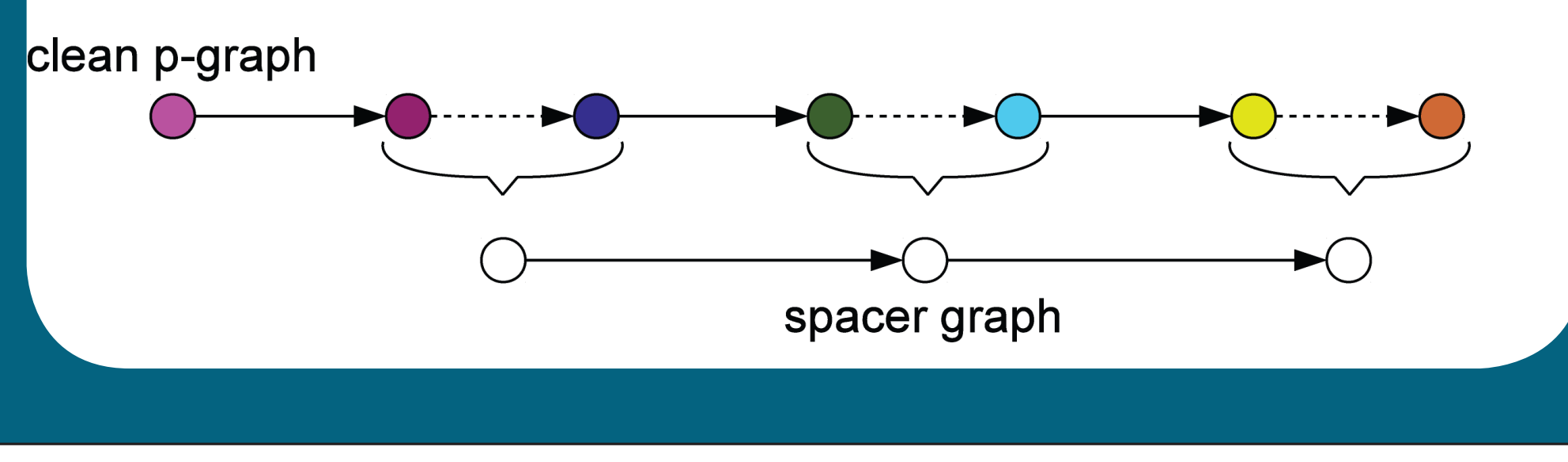


Stage 2: Graph Construction

1. Build P-Graph from spacer ends



2. Clean P-Graph and build spacer graph

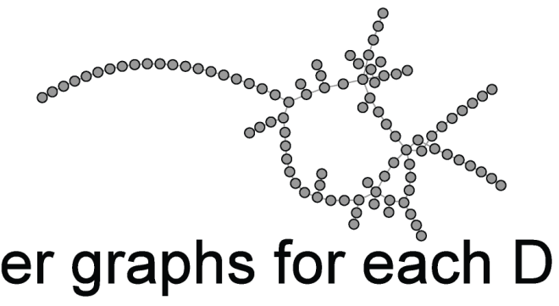


Stage 3: Output

1. XML (.crispr)

```
<?xml version="1.0" encoding="ISO8859-1" standalone="no" ?>
<crispr version="1.0">
  <group drseq="AGTTGGGATGTTTCCAATGTGACTAATATGAGAG" gid="G11">
    <data>
      <drs>
        <dr drid="DR1" seq="AGTTGGGATGTTTCCAATGTGACTAATATGAGAG"/>
      </drs>
      <spacers>
        <spacer cov="1" seq="GAATGTTTGGGCATAGTGAATTCAATCAAAATATTGGC" spid="SP12"/>
        <spacer cov="1" seq="AGATGTTTTATTTTAATGATAATTTTAATCAAGACCTAAAC" spid="SP15"/>
      </spacers>
    </data>
    <metadata>
      <notes>crass (0.2.12) run on 18_05_2012_111424 with command: crass -o crass_out_com...
      <file type="log" url="/home/user/crass.18_05_2012_111424.log"/>
      <file type="data" url="/home/user/Group_11_AGTTGGGATGTTTCCAATGTGACTAATATGAGAG_debug..
```

2. Images



Spacer graphs for each DR type are rendered using Graphviz

3. Reads

Fasta file is generated for each DR type identified. These files can be used for downstream assembly

Validating Crass

- Crass was validated against an Acid Mine Drainage (AMD) biofilm and the Global Ocean Survey (GOS) datasets

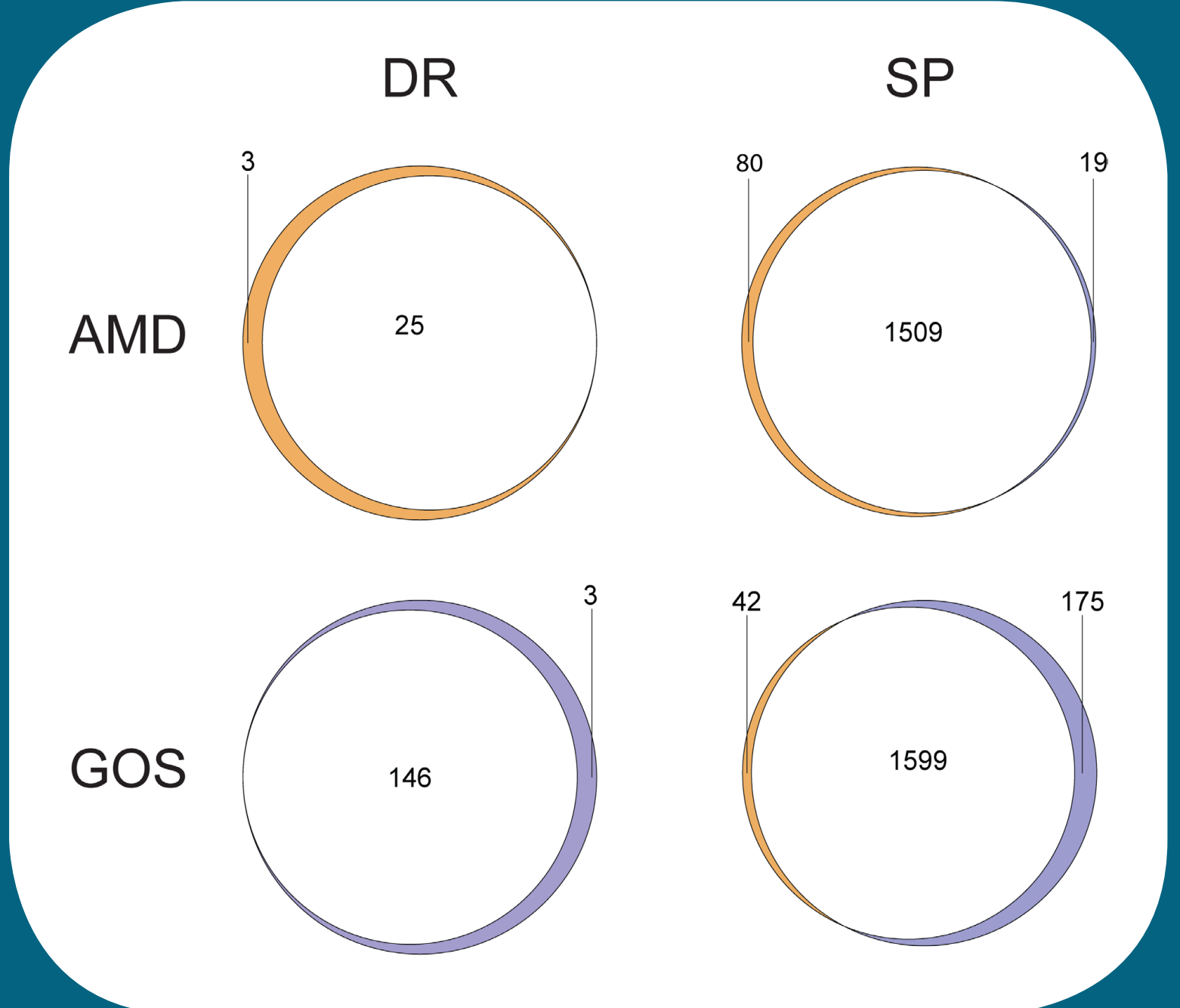


Figure 2: Unique direct repeats (DR) and spacers (SP) are coloured orange for Crass, or purple for the original analysis. The shared fraction is shown in white.

Discovery of Novel Phage-host Interactions

- Crass was used on a paired phage and microbial dataset constructed from an Enhanced Biological Phosphorus Removal (EBPR) reactor community
- We compared CRISPR loci discovered from a metagenomic assembly and from Crass
- Crass discovered 54 extra CRISPRs (Figure 3)
- CRISPRs found in the assembly were highly biased toward higher coverage
- Multiple strains could be seen in some CRISPRs (Figure 5)
- Only CRISPRs identified by Crass were found to contain targeting spacers in the phage metagenome (Figure 4)

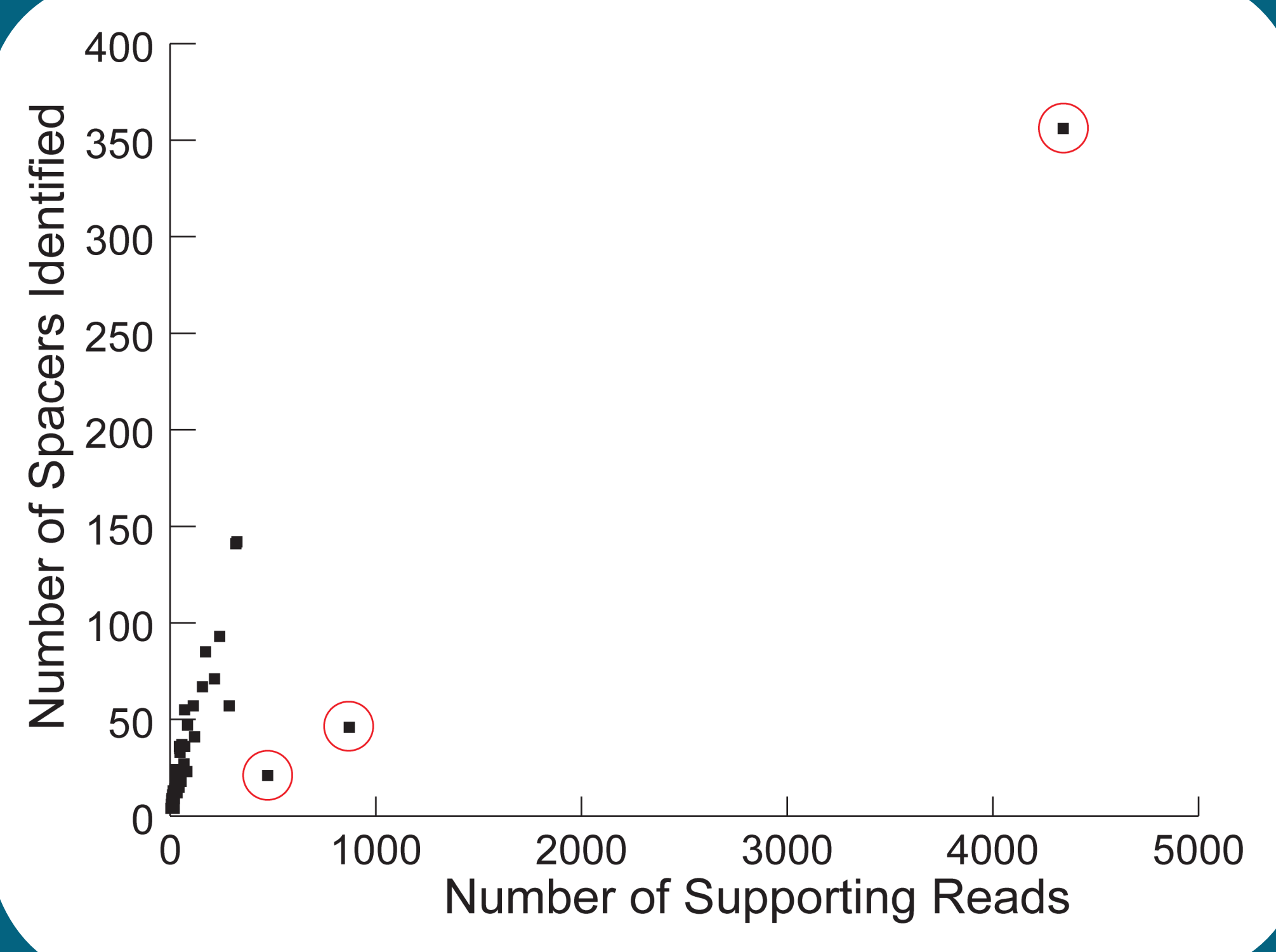


Figure 3: CRISPR loci discovered in the EBPR microbial dataset. The three loci that are circled in red were identified in the assembly. All other CRISPRs were identified only by Crass.

Spacer Targets in the EBPR Virome

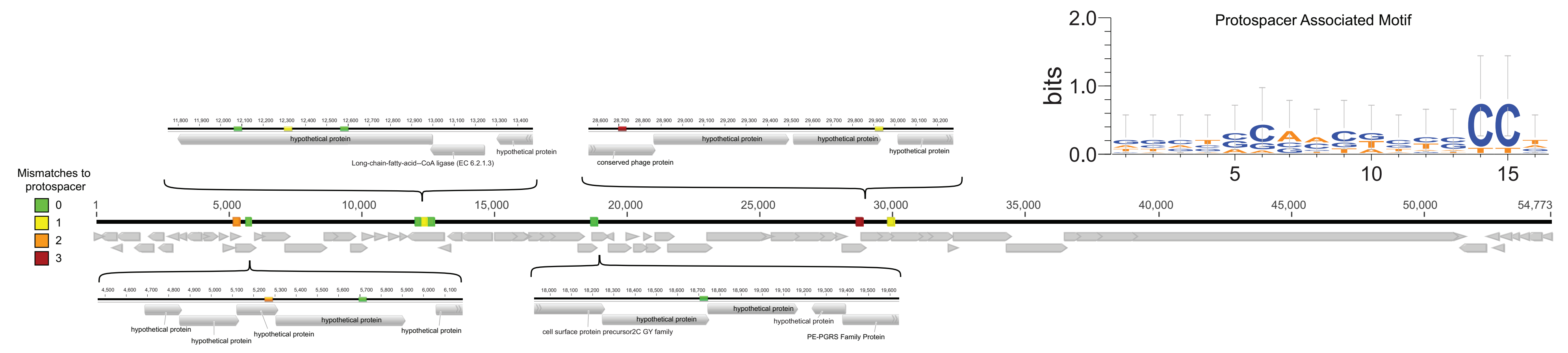


Figure 4: Genome representation of the most dominant phage in the EBPR dataset. Protospacers are indicated as colored bars on the genome, with the colour indicating the number of mismatches between the spacer and the genome. The protospacer associated motif (PAM) for the protospacers is shown in the top right corner.

Strain Variation in an EBPR CRISPR

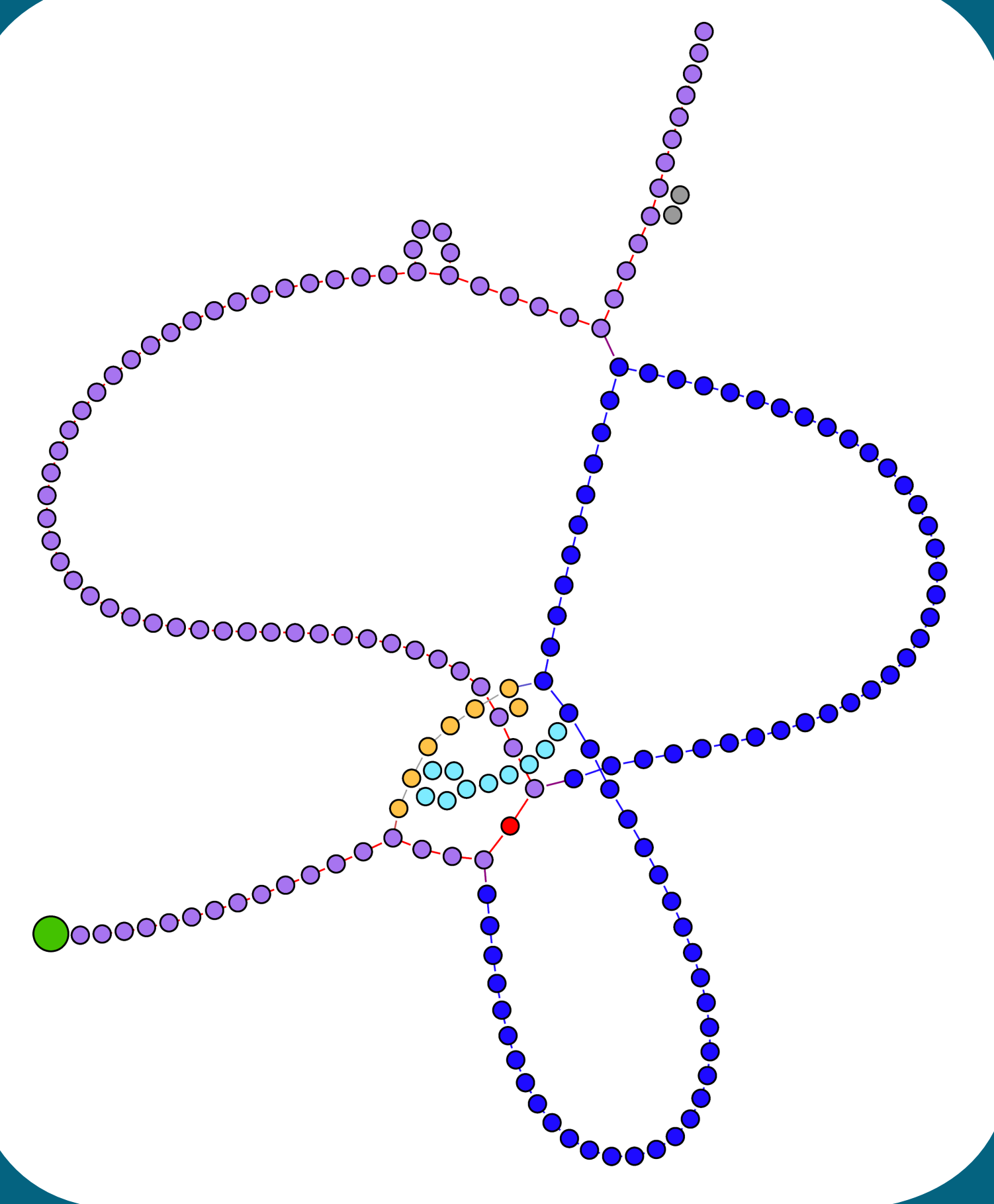


Figure 5: Spacer arrangement of the most abundant CRISPR in the EBPR dataset. This loci contains four strains; from the leader sequence (green) all strains begin on a unified purple segment and then diverge through either the yellow, red, cyan or blue paths. All of the strains finish in a conserved tail section, distal from the leader sequence.

Discussion

- Identifying CRISPRs from metagenomic assemblies is highly biased toward loci that have a large number of supporing reads
- Crass uses a novel graph based approach to overcome the repetitiveness of CRISPR loci and provides a graphical output that describes the strain variation found in the community
- The extended spacer complement identified by Crass was crucial to discovering phage-host links

1. Crass is freely available from: https://github.com/ctSkennerton/crass
2. Edgar, R. C. PILER-CR: Fast and accurate identification of CRISPR repeats. BMC Bioinformatics 8, 18, doi:10.1186/1471-2105-8-18 (2007).
3. Bland, C. et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8, 209, doi:10.1186/1471-2105-8-209 (2007).
4. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research 35, W52-W57 (2007).
5. Andersson, A. F. & Banfield, J. F. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. Science 320, 1047-1050, doi:10.1126/science.1157358 (2008).
6. Sorokin, V. A., Gelfand, M. S. & Artamonova, I. I. Evolutionary Dynamics of Clustered Irregularly Interspaced Short Palindromic Repeat Systems in the Ocean Metagenome. Applied and Environmental Microbiology 76, 2136-2144, doi:10.1128/aem.01985-09 (2010).
7. We thank the Australian Research Council (ARC) for funding this project, Grant DP1093175, and for funding Connor Skennerton's Australian Postgraduate Award (APA)