**Supplementary Data**

**Supplementary Figure S1:** This diagram outlines relationships amongst different CRISPR detection methods. CRISPR detection can be performed either using specified direct repeats (reference-guided detection) or without prior knowledge of direct repeat sequences (*de novo* detection). *De novo* detection searches raw metagenomic reads for direct repeat sequences of the appropriate length and spacing (i.e., 25-60 bp long repeats with 25-60 bp spacers between them). *De novo* detection techniques either detect spacers in reads only (Minced) or assemble reads into arrays (Crass). Reference-guided CRISPR detection, on the other hand, searches reads for user-specified direct repeat sequences, and extracts spacers from between direct repeat sequences identified in reads containing direct repeats. While the query is user-specified, general strategies for generating a query include using direct repeats identified from assembled metagenomic contigs or direct repeats found in genomic CRISPR arrays that might be expected based on taxonomic profiles (e.g., using known genomic *Streptococcus pyogenes* direct repeats if *Streptococcus pyogenes* is found in the metagenome’s taxonomic profile).

**Supplementary Figure S2:** Evaluation of MetaCRAST, Crass, and Minced performance on simulated AMD and EBPR metagenomes. All data points represent the averages of six individual simulations and are presented with error bars representing two times the standard error above and two below the average. The true number of spacers expected in each simulated metagenome is marked with a black line (138 expected in the AMD metagenomes; 219 in the EBPR metagenomes).



**Supplementary Figure S3:** Total number of CRISPR spacers detected in real AMD and EBPR metagenomes using four different detection methods – Crass (*de novo*), MetaCRAST (using assembly-guided queries), MetaCRAST (using taxonomy-guided queries), and Minced (*de novo*). Taxonomy-guided and assembly-guided queries are provided as Supplementary Tables S4-S7.

****

**Supplementary Figure S4:** Comparison of spacers detected in the real AMD metagenome using Crass (*de novo*), MetaCRAST (using taxonomy-guided queries), MetaCRAST (using assembly-guided queries), and Minced (*de novo*). Comparison was performed using Venny 2.1 (<http://bioinfogp.cnb.csic.es/tools/venny/>).

****

**Supplementary Figure S5:** Comparison of spacers detected in the real EBPR metagenome using Crass (*de novo*), MetaCRAST (using taxonomy-guided queries), MetaCRAST (using assembly-guided queries), and Minced (*de novo*). Comparison was performed using Venny 2.1 (<http://bioinfogp.cnb.csic.es/tools/venny/>).

****

**Supplementary Figure S6:** Evaluation of MetaCRAST, Crass, and Minced performance on simulated metagenomes with varying proportions of *Ferroplasma acidarmanus* fer1 and *Leptospirillum* sp. Group II 'CF-1' genome sequences. Simulated metagenomes were generated with Grinder. The data points represent the average number of “true positive” spacers detected that matched spacers in corresponding *Ferroplasma* or *Leptospirillum* CRISPR arrays. All data points represent the averages of six individual simulations and are presented with error bars representing two times the standard error above and two below the average. The true number of spacers expected for each genome is marked with a black line (20 expected in the *Ferroplasma* genome; 118 in the *Leptospirillum* genome).



**Supplementary Table S1:** Details about the AMD and EBPR metagenomes used for evaluation of the tools. The AMD metagenome was downloaded from <http://data.imicrobe.us/project/view/1>, while the EBPR metagenome was downloaded from <http://data.imicrobe.us/project/view/22>.

|  |  |  |  |
| --- | --- | --- | --- |
| Study site | Sequencing technology | Total reads | Average read length (bp) |
| Acid mine drainage (AMD) | Sanger | 319,166 | 1021 |
| Enhanced biological phosphorus removal (EBPR) | Sanger | 224,516 | 985 |

**Supplementary Table S2:** Profile used to generate simulated AMD metagenomes using Grinder. All simulated metagenomes contained 100,000 reads. 454 metagenomes were generated with this command: grinder -reference\_file AMDgenomes.fasta - abundance\_file AMDprofile.txt -total\_reads 100000 -read\_dist [one of 100, 150, 200, 250, 300, 400, or 600] normal 50 -homopolymer\_dist balzer. All 454 read length distributions were normal with a standard deviation of 50 bp. Illumina metagenomes were generated with this command: grinder -reference\_file AMDgenomes.fasta -abundance\_file AMDprofile.txt -total\_reads 100000 -read\_dist [one of 100, 150, 200, 250, 300, 400, or 600] -md poly4 3e-3 3.3e-8. All Illumina read length distributions were uniform with all reads having exactly the average read length.

|  |  |  |
| --- | --- | --- |
| Taxon | GenBank accession (chromosome) | Relative abundance (%) |
| *Leptospirillum* sp. Group II 'CF-1' | NZ\_CP012147.1 | 65 |
| *Ferroplasma acidarmanus* fer1 | NC\_021592.1 | 35 |

**Supplementary Table S3:** Profile used to generate simulated EBPR metagenomes using Grinder. All simulated metagenomes contained 100,000 reads. 454 metagenomes were generated with this command: grinder -reference\_file EBPRgenomes.fasta - abundance\_file EBPRprofile.txt -total\_reads 100000 -read\_dist [one of 100, 150, 200, 250, 300, 400, or 600] normal 50 -homopolymer\_dist balzer. All 454 read length distributions were normal with a standard deviation of 50 bp. Illumina metagenomes were generated with this command: grinder -reference\_file EBPRgenomes.fasta -abundance\_file EBPRprofile.txt -total\_reads 100000 -read\_dist [one of 100, 150, 200, 250, 300, 400, or 600] -md poly4 3e-3 3.3e-8. All Illumina read length distributions were uniform with all reads having exactly the average read length.

|  |  |  |
| --- | --- | --- |
| Taxon | GenBank accession (chromosome) | Relative abundance (%) |
| Candidatus *Accumulibacter phosphatis* clade IIA str. UW-1 | NC\_013194.1 | 100 |

**Supplementary Table S4:** Taxonomy-guided query used for real AMD metagenome. Direct repeats were obtained from CRISPRdb and corresponding genomic GenBank numbers associated with taxa using NCBI Efetch. The *Leptospirillum* sp. Group IV 'UBA BS' direct repeat was detected in the listed assembly using CRISPRFinder.

|  |  |  |
| --- | --- | --- |
| Taxon | Accession number  | Query sequence  |
| *Ferroplasma acidarmanus* fer1 | NC\_021592.1(GenBank) | ATTTCAATTCCTATATGGAATTATTTTAAC |
| *Ferroplasma acidarmanus* fer1 | NC\_021592.1(GenBank) | GTGTTTAGTCTATCTATAAGGGTTTGAAAT |
| *Leptospirillum* sp. Group II 'CF-1' | NZ\_CP012147.1(GenBank) | GTATTCCCCACGTTCGTGGGGATGAACCG |
| *Leptospirillum* sp. Group IV 'UBA BS' | GCA\_000496115.1 (Assembly) | GTTTTCCCCGCATGCGCGGGGGTGTTTCT |

**Supplementary Table S5:** Taxonomy-guided query used for the real EBPR metagenome. Direct repeats were obtained from CRISPRdb and corresponding genomic GenBank numbers associated with taxa using NCBI Efetch.

|  |  |  |
| --- | --- | --- |
| Taxon | Accession number  | Query sequence  |
| Candidatus *Accumulibacter phosphatis* clade IIA str. UW-1 | NC\_013194.1 (GenBank) | GTCTCAATCCCTTTGATTTCAGGGCTGGTTACTGAC  |
| Candidatus *Accumulibacter phosphatis* clade IIA str. UW-1 | NC\_013194.1 (GenBank) | GTTTCCCCCGCGTCAGCGGGGATAGGCCC |

**Supplementary Table S6:** Assembly-guided query used for the real AMD metagenome. CAP3 assembled AMD reads into contigs with default parameters. CRISPR DRs were detected in CAP3 contigs using PILER-CR. The DRs clustered with a similarity threshold of 0.9 were then used to search the real AMD metagenome.

|  |  |
| --- | --- |
| PILER-CR DR information | DR sequence |
| Contig784[Array1;Pos=75] | CGGTTCATCCCCACGAACGTGGGGAATAC |
| Contig1409[Array2;Pos=26] | ATTTCAGAAAAACTAGTTAGTATGGAAG |
| Contig2393[Array5;Pos=904] | CTTTGAAACTTTCTAAATAAGATTCTAAC |
| Contig3459[Array7;Pos=1814] | GTTAGAATCTTATTTAGAAAGTTTCAAAGT |
| Contig3740[Array9;Pos=78] | CTTTCAATCCTATCAAGGTTCTATTTTTAC |
| Contig3754[Array10;Pos=255] | GTATTCCCCACGTTCGTGGGGATGAACCG |
| Contig3832[Array11;Pos=246] | GTTTAAAAAGCACTAGGTAGTATGGAAG |
| Contig3945[Array12;Pos=153] | GTTAAAATCGAACCTTAATAGGATTGAAAG |
| Contig3946[Array13;Pos=1080] | CTTTCAATCCTATTAAGGTTCGATTTTAAC |
| Contig4021[Array14;Pos=104] | GTAAAAATAGAACCTTGATAGGATTGAAAG |
| Contig4922[Array16;Pos=82] | ATTAGAAATATATCCTATAAGGAATTGATAC |
| Contig6764[Array22;Pos=263] | GTCTTCCCCACGCCCGTGGGGGTGTTTC |
| Contig7045[Array23;Pos=529] | GTAGTCCCCACGTATGTGGGGGTGAAGGG |
| Contig7402[Array24;Pos=53] | GTCTTAATCCCTTATTTATCAGGTCTTACCTTCGTTT |
| Contig8016[Array26;Pos=72] | GTGTTTAGTCTATCTATAAGGGTTTGAAAT |
| Contig8725[Array28;Pos=280] | GAAACACCCCCACGGGCGTGGGGAAGAC |
| Contig9087[Array29;Pos=204] | ATTTCCATAATAGAAATATTATGGCTCTATTGAAGC |
| Contig11151[Array31;Pos=153] | CTTCCATACTACCTAGTGCTTTTTAAAC |

**Supplementary Table S7:** Assembly-guided query used for the real EBPR metagenome. CAP3 assembled EBPR reads into contigs with default parameters. CRISPR DRs were detected in CAP3 contigs using PILER-CR. The DRs clustered with a similarity threshold of 0.9 were then used to search the real EBPR metagenome.

|  |  |
| --- | --- |
| PILER-CR DR information | DR sequence |
| Contig162[Array1;Pos=1038] | CTTTGAAGCTCGCCCCGATTTAGAGGGGATTAAGAC |
| Contig1390[Array2;Pos=112] | TTTCTAAGCCGCCATCACGGCGGCAAAC |
| Contig3577[Array4;Pos=288] | CGGTTCATCCCCACAGATACGGGGAACAC |
| Contig4649[Array8;Pos=1549] | GAGCGTGTCGTTGCCGGC |
| Contig4987[Array9;Pos=92] | CTGCCGTTATCCCTGATGCCGAAAGGCGTTGAGCAC |
| Contig5142[Array11;Pos=112] | GTTTGCCGCCGTGATGGCGGCTTAGAAA |
| Contig5455[Array12;Pos=1527] | CATTCTCCCAGCTAATTATGTTGGGAGTGGATTGAAACA |
| Contig6574[Array13;Pos=70] | GTTTCAATCCGCGCCCCTCGTTGCCGAGGGGCGATGC |
| Contig7071[Array14;Pos=54] | GTCTCAATCCCTTTGATTTCAGGGCTGGTTACTGAC |
| Contig7106[Array16;Pos=1039] | GGGCCTATCCCCGCTGACGCGGGGGAAAC |
| Contig8721[Array18;Pos=119] | GTCAGTAACCAGCCCTGAAATCAAAGGGATTGAGAC |
| Contig10087[Array20;Pos=540] | GTTCTCGCTCCCCGACTTCCTGAAGGGGATTAAGAC |
| Contig11041[Array22;Pos=71] | GTTGTGAATTGCTTTCAAATTCTTAAGTAACTTAGTTCTTGCACAAC |
| Contig11786[Array23;Pos=429] | GTTTCCCCCGCGTCAGCGGGGATAGGCC |
| Contig12588[Array27;Pos=221] | GTCACAAAGCCCTATTTACGGGCAGGGGTGACGGAC |
| Contig12822[Array28;Pos=605] | GTCGCCCGTCACTCCGGTGACGGGCGTGGATTGAAAC |

**Supplementary Table S8:** Taxonomic profile of the real AMD metagenome determined using MetaPhyler.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Taxon | % Abundance | Depth of coverage | Number of reads | Similarity with reference |
| Ferroplasma | 67.68 | 16.69 | 960 | 88.54 |
| Leptospirillum | 23.48 | 5.79 | 318 | 95.43 |
| Euryarchaeota{phylum} | 2.35 | 0.58 | 105 | 83.49 |
| Thermoplasmatales{order} | 2.03 | 0.5 | 59 | 85.97 |
| Actinobacteria{phylum} | 1.06 | 0.26 | 34 | 85.92 |
| Chlorobiaceae{family} | 0.76 | 0.18 | 28 | 88.91 |
| Fusobacteria{phylum} | 0.6 | 0.14 | 30 | 88.18 |
| Actinobacteria (class){class} | 0.38 | 0.09 | 19 | 86.93 |
| Proteobacteria{phylum} | 0.31 | 0.07 | 23 | 85.18 |
| Aquificae{phylum} | 0.24 | 0.06 | 10 | 88.09 |
| Deinococcus-Thermus{phylum} | 0.21 | 0.05 | 10 | 83.41 |
| Chlorobi{phylum} | 0.14 | 0.03 | 15 | 86.03 |
| Firmicutes{phylum} | 0.13 | 0.03 | 4 | 84.3 |
| Actinomycetales{order} | 0.13 | 0.03 | 8 | 88.6 |
| Bacteroidetes{phylum} | 0.09 | 0.02 | 8 | 84.46 |
| Delftia | 0.09 | 0.02 | 4 | 99.56 |
| Caulobacteraceae{family} | 0.08 | 0.02 | 2 | 94.3 |
| Chromatiales{order} | 0.03 | 0 | 4 | 85.45 |
| Deltaproteobacteria{class} | 0.02 | 0 | 1 | 88.89 |
| Nocardioidaceae{family} | 0.02 | 0 | 1 | 96.59 |
| Micrococcaceae{family} | 0.01 | 0 | 1 | 87.25 |
| Frankia | 0.01 | 0 | 1 | 91.46 |
| Crenarchaeota{phylum} | 0.01 | 0 | 2 | 84 |
| Thermoplasma | 0 | 0 | 1 | 89.74 |
| Thermotogae{phylum} | 0 | 0 | 1 | 87.32 |
| Veillonellaceae{family} | 0 | 0 | 1 | 91.55 |

**Supplementary Table S9:** Taxonomic profile of the real EBPR metagenome determined using MetaPhyler.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Taxon | % Abundance | Depth of coverage | Number of reads | Similarity with reference |
| Candidatus Accumulibacter | 61.54 | 10.42 | 463 | 97.82 |
| Proteobacteria{phylum} | 8.78 | 1.48 | 149 | 83.93 |
| Xanthomonadaceae{family} | 6.22 | 1.05 | 60 | 90.8 |
| Bacteroidetes{phylum} | 4.61 | 0.78 | 56 | 84.12 |
| Comamonadaceae{family} | 1.49 | 0.25 | 11 | 92.36 |
| Gammaproteobacteria{class} | 1.35 | 0.22 | 29 | 88.69 |
| Betaproteobacteria{class} | 1.31 | 0.22 | 19 | 87.49 |
| Flavobacteriaceae{family} | 1.21 | 0.2 | 8 | 90.22 |
| Alphaproteobacteria{class} | 1.14 | 0.19 | 14 | 89.3 |
| Actinomycetales{order} | 0.99 | 0.16 | 8 | 88.53 |
| Burkholderiales{order} | 0.94 | 0.15 | 14 | 88.32 |
| Rhodobacteraceae{family} | 0.92 | 0.15 | 6 | 90.84 |
| Verrucomicrobia{phylum} | 0.71 | 0.12 | 7 | 84.87 |
| Fibrobacteres{phylum} | 0.67 | 0.11 | 40 | 86.58 |
| Actinobacteria{phylum} | 0.6 | 0.1 | 13 | 85.71 |
| Pseudoxanthomonas | 0.59 | 0.1 | 3 | 98.9 |
| Flavobacterium | 0.57 | 0.09 | 4 | 93.04 |
| Rhizobiales{order} | 0.56 | 0.09 | 4 | 87.59 |
| Rhodobacterales{order} | 0.51 | 0.08 | 3 | 88.28 |
| Flavobacteriales{order} | 0.49 | 0.08 | 5 | 87.78 |
| Chloroflexi{phylum} | 0.48 | 0.08 | 3 | 86.54 |
| Acidovorax | 0.45 | 0.07 | 3 | 93.63 |
| Sphingomonadaceae{family} | 0.41 | 0.06 | 3 | 92.46 |
| Shigella | 0.4 | 0.06 | 2 | 99.06 |
| Rhodocyclaceae{family} | 0.39 | 0.06 | 17 | 85.19 |
| Phyllobacteriaceae{family} | 0.35 | 0.05 | 3 | 90.43 |
| Flavobacteria{class} | 0.29 | 0.04 | 2 | 88.47 |
| Firmicutes{phylum} | 0.21 | 0.03 | 7 | 86.69 |
| Ectothiorhodospiraceae{family} | 0.18 | 0.03 | 2 | 85.99 |
| Beijerinckiaceae{family} | 0.18 | 0.03 | 2 | 90.27 |
| Alteromonadales{order} | 0.17 | 0.02 | 5 | 89.19 |
| Mesorhizobium | 0.15 | 0.02 | 2 | 89.46 |
| Magnetospirillum | 0.11 | 0.01 | 1 | 93.24 |
| Aeromonas | 0.09 | 0.01 | 2 | 99.37 |
| Cytophagaceae{family} | 0.09 | 0.01 | 2 | 90.65 |
| Acetobacteraceae{family} | 0.09 | 0.01 | 1 | 91.27 |
| Neisseriaceae{family} | 0.08 | 0.01 | 2 | 89.1 |
| Chlorobi{phylum} | 0.05 | 0 | 2 | 86.33 |
| Xanthomonas | 0.04 | 0 | 2 | 87.1 |
| Synergistetes{phylum} | 0.03 | 0 | 1 | 84.88 |
| Methylophilales{order} | 0.03 | 0 | 1 | 87.6 |
| Clostridiaceae{family} | 0.03 | 0 | 1 | 89.71 |
| Moraxellaceae{family} | 0.03 | 0 | 1 | 86.67 |
| Acinetobacter | 0.03 | 0 | 3 | 93.75 |
| Bacteroidales{order} | 0.03 | 0 | 2 | 88.43 |
| Caulobacter | 0.03 | 0 | 1 | 92.96 |
| Clostridia{class} | 0.03 | 0 | 1 | 90.48 |
| Cyanobacteria{phylum} | 0.02 | 0 | 2 | 85.71 |
| Photobacterium | 0.02 | 0 | 1 | 89.16 |
| Burkholderiaceae{family} | 0.02 | 0 | 1 | 89.23 |
| Campylobacter | 0.02 | 0 | 3 | 90.48 |
| Bacillales{order} | 0.02 | 0 | 2 | 87.4 |
| Roseovarius | 0.01 | 0 | 1 | 91.41 |
| Aurantimonadaceae{family} | 0.01 | 0 | 1 | 93.55 |