The CRISPR File Specification (v1.1)

Michael Imelfort and Connor Skennerton

18th November 2011

1 Introduction

The CRISPR file (.crispr) is an xml based document designed to describe all properties of the direct repeats and spacer sequences obtained from CRISPR loci from genomic and metagenomic datasets. It's key function is to enable a standard way of describing the complex arrangements of spacers that are seen in environmental samples. Each .crispr file can contain multiple <group> tags; each one corresponding to a different CRISPR locus. Each group then has three tags: <data>, <metadata> and <assembly>

1.1 Terminologies and Concepts

1.1.1 Clustered Regularly Interspersed Short Palindromic Repeats (CRISPRs)

CRISPRs are a class of repeatative elements found in many bacterial and archaeal genomes. They encode a complex microbial immune system capable of rapidly adapting to, and targeting invading DNA. CRISPR elements are composed of direct repeats (DRs), spacers and flanking sequences.

Direct Repeats Although the P in CRISPR is for palindromic, most DRs do not form perfect palindromes. There are many different types of repeats, some that do contain palindromic regions in them and others that do not. What is common is that within a CRISPR locus the sequence of the DR will remain identical. This is essential as the DR sequence is used to bound the spacer sequences that are used for the immune response.

Spacers Spacer sequences are small segments of invading DNA that is added to a CRISPR locus. These spacers are used as a targeting mechanism of degradative proteins.

The Leader Sequence The leader seuqence flanks one side of the the CRISPR locus and is thought to act as a promoter sequence.

2 Elements

2.1 Data

The data tag is designed list all of the spacers that a CRISPR contains as well as the direct repeats, any flanking sequences such as the leader sequence and the sources (reads/genomes) that this CRISPR came from.

```
<data>
        <sources>
                <source accession="read:x:y:z:1:HGITA" soid="SO1"/>
        </sources>
        <drs>
                <dr seq="GAGTCCCCGC" drid="DR1" confcnt="12" totcnt="20"/>
        </drs>
        <spacers>
                <spacer seq="TGAGCGGTCGC" spid="SP1"/>
        </spacers>
        <flankers>
                <flanker seq="GGAGTTCTAC" flid="FL1"/>
        </flankers>
</data>
```

2.1.1 Sources

The <sources> tag is used to define the reads/genome/contig that contained (wholly or part-of) the CRISPR. Each <source> tag has two requiredattributes: accession and soid. The accession attribute lists the original accession for the source and the soid provides a reference for that sourcefor use subsuquently in the file. The motivation for this is to reduce the number of characters in the file, since accessions/read names tend to be quite long.

2.1.2 Direct Repeats

Direct repeats are defined in the <drs> tag as a <dr>. Each <dr> tag has two requiredattributes: seq and drid; and two implied attributes: confcnt and totcnt

```
<drs>
        <dr seq="GAGTTCCCCGCGCCA" drid="DR1" confcnt="12" totcnt="20"/>
</drs>
```

The seq attribute is the sequence of the direct repeat in its lowest lexicographical form and should not contain any characters other than the IUPAC standard for DNA bases and should be in uppercase. The drid attribute is set by the programer, it must be an integer prefixed with "DR" however it does not need to be sequential – just unique for the <group> that it is contained in.

The two implied attributes confcnt and totcnt the number of spacers associated with this particular direct repeat variant.

2.1.3 Spacers

Spacers are defined inside the <spacers> tag as an empty element containing two required attributes seq and spid, as well as an implied attribute cov.

```
<spacers>
       <spacer seq="TGAGCGGTCGC" spid="SP1" cov="10"/>
</spacers>
```

The seq attribute is listed in the same directionality as the direct repeat. This is not necessarily is the lowest lexicographical form for the spacer but will be the form in which the direct repeat is in it's lowest lexicographical form. The spid attribute is set by the programer, it must be an integer prefixed with "SP" however it does not need to be sequential – just unique for the <group> that it is contained in. The cov attribute describes how many instances of this spacer were seen in the genome/metagenome/dataset, which may be used in assembly to resolve forks in the arrangement of spacers or to determine community abundance of particular spacers.

2.1.4 Flankers

The <flankers> element is not a required part of <data> however if present decribes any flanking sequences found on either side of the spacer array, such as the leader sequence. Each <flanker> contains two required attributes: seq and flid.

```
<flankers>
       <flanker seq="GGAGTTCTAC" flid="FL1"/>
</flankers>
```

The requirements for these attributes are the same as for spacers, namely that the sequence be in the form to have the associated direct repeat be in its lowest lexicographical form; and the flid being an integer, unique to a <group>.

2.2 Metadata

The <metadata> tag can hold a range of information that links with a particular group. The metadata can take two forms, as either a <notes> tag that holds freeform infromation inputed by the user or outputed by a program; or as a URL to a local or remote file.

```
<metadata>
     <program>
              <name>crass</name>
              <version>0.2.13</version>
              <command>crass -K 9 raw/combined.fa </command>
```

```
      </program>
        <notes>
                E. coli dataset of ca. 20Million 100bp reads
        </notes>
        <file type="image" url="./crispr1.jpg"/>
        <file type="sequence" url="./crispr1.fa"/>
        <file type="data" url="./crispr1_mimmarks.gcdml"/>
</metadata>
```

2.2.1 Notes

The <notes> tag should not be a dump for large amounts of internal or external data (such as a program log) but instead be a place for the user to store information about program parameters or basic dataset information. More extensive data and information should be referred to using the <file> tag (see below).

2.2.2 External Files

The <file> tag can be used to refer to a file that contains extra data about this CRISPR pointed to by the url attribute. There are currently three acceptable types of files defined by the type attribute: image, sequence and data. Image files could show a graphical representation of the spacer arrangment produced by programs such as Graphviz. Sequence files refer to fasta or fastq files containing reads that belong to this CRISPR. The final option is generic and may reference any file that holds extra information about the dataset, such as a MIGS/MIMS standards compliant metadata file.

2.3 Assembly

The <assembly> tag defines how all of the spacers, flankers and direct repeats defined in the <data> section are arranged together. An <assembly> tag contains a one or more <contig> tags that contain references to individual spacers.

```
<assembly>
        <contig cid="C2">
                <cspacer spid="SP6">
                        <bspacers>
                                <bs spid="SP36" drid="DR1" drconf="0" />
                        </bspacers>
                        <fspacers>
                                <fs spid="SP27" drid="DR1" drconf="0" />
                        </fspacers>
                </cspacer>
                <concensus>
                        TCAGCTTTATAAATCCGGAGATACGGAAACTA
                </concensus>
        </contig>
</assembly>
```

2.3.1 Contigs

Contigs are linear sections of a graph where there are no branches between spacers; that is to say that every spacer in a contig contains a single forward and backward link. Therefore spacers which have multiple forward or backward spacers links are placed into a contig on their own. This is to make it easy to assembly and join multiple different branching paths through a CRISPR locus as spacers that join multiple divergent arrangments would be in both pathways.

Each contig is defined with a <contig> tag, which requires a cid attribute that must be unique to the <group>. A cid must be an integer prefixed with "C". Each <contig> must contain one or more <cspacer> tags and may also contain a <concensus> tag.

<cspacer> contains a single required attribute spid that refers to one of the spacers defined in the <data> tag. It should also contain references to any linked <spacer> or <flanker> sequences using the <bspacers>, <fspacers>, <bflankers> and <fflankers> tags, although if a spacer does not connect to anything, then none of these tags will be present.

<bspacers> & <fspacers> These two tags contain lists of the forward linking and backward linking spacers for the current <cspacer>. There does not ned to be both types represented for a <cspacer> – there may be neither if the spacer joins only to flankers. The <fs> and <bs> tags denote the forward and backward spacers, respectively. They have a single required attribute spid as well as two implied attributes drid and drconf. If no drid attribute is provided then the default direct repeat for the group will be used.

<bflankers> & <fflankers> These two tags work in the same way as <bspacers> and <fspacers> and contain information about any flanking sequences that join to this spacer. Each individual flanker is marked with either <bf> or <ff> tags.

Concensus

3 Document Type Definition (DTD)

```
<!DOCTYPE crispr [
<!ELEMENT crispr (group+)>
<!ATTLIST crispr version CDATA "1.1">
<!ELEMENT group (data+,metadata*,assembly+)>
<!ATTLIST group gid CDATA #REQUIRED>
<!ATTLIST group drseq CDATA #REQUIRED>
```

```
<!ELEMENT data (sources*,drs+,spacers+,flankers*)>
<!ELEMENT sources (source+)>
<!ELEMENT source EMPTY>
<!ATTLIST source soid CDATA #REQUIRED>
<!ATTLIST source accession CDATA #REQUIRED>
<!ELEMENT drs (dr+)>
<!ELEMENT dr EMPTY>
<!ATTLIST dr seq CDATA #REQUIRED>
<!ATTLIST dr drid CDATA #REQUIRED>
<!ATTLIST dr confcnt CDATA #IMPLIED>
<!ATTLIST dr totcnt CDATA #IMPLIED>
<!ELEMENT spacers (spacer+)>
<!ELEMENT spacer (source*)>
<!ATTLIST spacer seq CDATA #REQUIRED>
<!ATTLIST spacer spid CDATA #REQUIRED>
<!ATTLIST spacer cov CDATA #IMPLIED>
<!ELEMENT source (spos?, epos?)>
<!ATTLIST source soid CDATA #REQUIRED>
<!ELEMENT spos CDATA #REQUIRED>
<!ELEMENT epos CDATA #REQUIRED>
<!ELEMENT flankers (flanker*)>
<!ELEMENT flanker EMPTY>
<!ATTLIST flanker seq CDATA #REQUIRED>
<!ATTLIST flanker flid CDATA #REQUIRED>
<!ELEMENT metadata (program?,notes?,file*)>
<!ELEMENT program (name, version, command?)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT version (#PCDATA)>
<!ELEMENT command (#PCDATA)>
<!ELEMENT notes (#PCDATA)>
<!ELEMENT file EMPTY>
<!ATTLIST file type (image|sequence|log|data) #REQUIRED>
<!ATTLIST file url CDATA #REQUIRED>
<!ELEMENT assembly (contig+)>
<!ELEMENT contig (sources*,consensus*,cspacer+)>
<!ATTLIST contig cid CDATA #REQUIRED>
<!ELEMENT sources (source+)>
<!ELEMENT source EMPTY>
<!ATTLIST source soid CDATA #REQUIRED>
<!ELEMENT consensus (#PCDATA)>
<!ELEMENT cspacer (bspacers?,fspacers?,bflankers?,fflankers?)>
<!ATTLIST cspacer spid CDATA #REQUIRED>
<!ELEMENT bspacers (bs+)>
```

```
<!ELEMENT bs EMPTY>
<!ATTLIST bs spid CDATA #REQUIRED>
<!ATTLIST bs drid CDATA #IMPLIED>
<!ATTLIST bs drconf (0|1) #IMPLIED>
<!ELEMENT fspacers (fs+)>
<!ELEMENT fs EMPTY>
<!ATTLIST fs spid CDATA #REQUIRED>
<!ATTLIST fs drid CDATA #IMPLIED>
<!ATTLIST fs drconf (0|1) "0">
<!ELEMENT bflankers (bf)>
<!ELEMENT bf EMPTY>
<!ATTLIST bf flid CDATA #REQUIRED>
<!ATTLIST bf drid CDATA #IMPLIED>
<!ATTLIST bf drconf (0|1) "0">
<!ATTLIST bf directjoin (0|1) "0">
<!ELEMENT fflankers (bf)>
<!ELEMENT ff EMPTY>
<!ATTLIST ff flid CDATA #REQUIRED>
<!ATTLIST ff drid CDATA #IMPLIED>
<!ATTLIST ff drconf (0|1) "0">
<!ATTLIST ff directjoin (0|1) "0">
]>
```

4 Example