Primer Scheme Specifications

Chris Kent[a*], Bede Constantinides[a]

[a] Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK.

1. Abstract

Amplicon Sequencing has become a dominant method for genomic surveillance. However, the lack of defined file format has lead to incompatibility issues with downstream analysis, and constantly evolving formats. Here we describe a universal specification for the primer.bed and the corresponding reference.fasta files, which will aid compatibility.

Key words: Data standards; Primer Schemes; Amplicon Sequencing.

Contents

2. primer.bed file

A primer.bed file describes an amplicon sequencing primer scheme and is generated by tooling. Its purpose is to encapsulate all the information needed to i) reproduce a primer scheme and ii) facilitate correct bioinformatic analysis of resulting sequencing data. It therefore incorporates both

[*] Corresponding author. Address: Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK.. Email: chrisgkent@protonmail.com

Table 1. The column structure and description of a BedLine

Column	Field Name	Type	Brief description	Restrictions
1	chrom	String	Chromosome name	[A-Za-z0-9_]
2	primerStart	Int	Primer start position	u64
3	primerEnd	Int	Primer end position	u64
4	primerName	String	Primer name	[a-zA-Z0-9\-]+_[0-9]+_(LEFT|RIGHT)_[0-9]+
5	pool	Int	Primer pool	u64
6	strand	String	Primer strand	[-+]
7	primerSeq	Optional(float)	Primer weight for PCR reactions	f64 > 0
8	attributes	String	list of key=value pairs separated by `;`	f64 > 0

wet lab and analytical elements. These include primer sequences, their associated pools, and relative concentrations, as well as their coordinates with respect to one or more reference genome sequences.

2.1. Format overview

primer.bed files are tab-delimited text files. Lines prefixed with Each line can either be a comment line (prefixed with #) or a BedLine, which represents a single unique primer (Oligonucleotide) that forms part of an associated amplicon. A compliant primer.bed file contains one or more amplicons.

The format of primer.bed is based on the Browser Extensible Data (BED) specification, with seven required columns followed by one optional column.

2.2. Comment Line

The comment line is minimally parsed, but has the option to contain key=value pairs. If the line contains a single = it will be split, with the left|right sides being key|value respectively.

2.3. BedLine field descriptions

2.3.1. chrom

The name of the corresponding reference sequence chromosome for the primer. This must match a valid sequence ID inside an accompanying reference sequence FASTA file, by convention named reference.fasta.

2.3.2. primerStart

The start position of the primer on the chrom.

2.3.3. primerEnd

The non-inclusive end position of the primer on the chrom. Must be greater than primerStart.

2.3.4. primerName

The name of the primer in the form "{prefix}_{ampliconNumber}_{direction}_{primerNumber}".

- prefix: Must match regex [a-zA-Z0-9\-]. See best practices
- ampliconNumber: The number of the amplicon for its relevant chrom. Must be a positive integer incrementing from 1.
- direction: The direction of the primers. Must be either LEFT or RIGHT.

- `primerNumber`: The number of the primer. Must be a positive integer incrementing from 1.

2.3.5. `pool`

The PCR pool the primer belongs to. Must be a positive integer incrementing from 1.

2.3.6. `strand`

The strand of the primer. Must be either `+` or `-`. Required to match the `primerName:direction` (`LEFT==+`, `RIGHT==-`)

2.3.7. `primerSeq`

The sequence of the primer in the 5′ to 3′ direction. Unrestricted to contain any character[1], and parsed by only removing white space.

2.3.8. `primerAttributes`

An **optional** list of a `key=value` pair to denote additional primer attributes, in the form of `pw=1.0;ps=10.0`. This is intentionally flexible to allow the storage of additional information.

Some key are reserved and undergo validation;

- `pw|primerWeight`: To ensure all amplicons perform similarly, the concentration of individual primers can be altered. Primer Concentration in the PCR should be scaled by `primerWeight * [typical PCR conc]`. This is restricted to positive numerics (`f64 > 0`).

2.4. Examples

2.4.1. Simple example

A seven column `primer.bed` file, with no `primerAttributes` or `comment lines`.

```
MN908947.3  100 131 example_1_LEFT_1  1 + CTCTTGTAGATCTGTTCTCTAAACGAACTTT
MN908947.3  419 447 example_1_RIGHT_1 1 - AAAACGCCTTTTTCAACTTCTACTAAGC
MN908947.3  344 366 example_2_LEFT_1  2 + TCGTACGTGGCTTTGGAGACTC
MN908947.3  707 732 example_2_RIGHT_1 2 - TCTTCATAAGGATCAGTGCCAAGCT
```

2.4.2. Complex example

An eight column `primer.bed` file. With `primerAttributes` defined, and `comment lines` providing a `chrom` alias and explaining the `gc` `primerAttributes`.

```
# example scheme
# gc=fraction gc
# MN908947.3=sars-cov-2
MN908947.3  100 131 example_1_LEFT_1  1 + CTCTTGTAGATCTGTTCTCTAAACGAACTTT pw=1.4;gc=0.35
MN908947.3  419 447 example_1_RIGHT_1 1 - AAAACGCCTTTTTCAACTTCTACTAAGC  pw=1.4;gc=0.36
MN908947.3  344 366 example_2_LEFT_1  2 + TCGTACGTGGCTTTGGAGACTC  pw=1;gc=0.55
MN908947.3  707 732 example_2_RIGHT_1 2 - TCTTCATAAGGATCAGTGCCAAGCT pw=1;gc=0.44
```

2.5. Best Practices

`primer.bed` contain information about how to replicate the primer pools used in multiplexed PCR. They do not contain information about the PCR protocol, input material, or sequencing method and analysis. Therefore, additional information is needed for true reproducibility.

[1] "This is unrestricted (rather than IUPAC-only) to allow Primer Modification. Such as `/56-FAM/{primerSeq}` to represent 5′ 6-FAM fluorescent dye labeled primer"

2.5.1. Other metadata standards

To explicitly differentiate different versions of `primer.bed`, this spec is designed to fit into larger metadata standards, such as primal-page with PrimalScheme Labs or primaschema with pha4ge primer-schemes

2.5.2. Other tooling

`primalbedtools` is a python package that carries out schema validation and conversion, and common operations on `primer.bed` files.

2.5.3. `primerName:prefix`

The `primerName:prefix` should be as unique as possible (ideally a short uuid. For example `359ba5`) and different for `chrom` and the scheme generation run.

- Using `primerName:prefix` like `scheme` or `sars-cov-2` might seen easier, however, will result in a freezer / LIMS full of simular names leading to pooling mistakes.

2.5.4. comment line

The `comment line`'s `key=value` pattern is non-validated and should be non-critical to the bed file function. Although it is recommended that is it used to explained non-standard `primerAttributes`.

Another use is providing aliases for different `chrom`.

3. reference.fasta file

A `reference.fasta` file contains the DNA sequences of all the primary-reference genomes, used in primerscheme generation. Its purpose is to provide a reference genome, and coordinate system to be used for referenced-based assembly and consensus generation.

3.1. Format overview

`reference.fasta` files are typical `.fasta` format files, with text representing the nucleotide sequence of the reference. Each genome starts with a header line (starting with `>`) that denotes the id of the genome, followed by lines of nucleotide data.

All `chrom` fields of the BedLines must have a corresponding id in the `reference.fasta`.

3.2. Examples

3.2.1. Single fasta

```
>MN908947.3
ATTAAAGGTTTATACCTTCCCA...
```

> The corresponding `primer.bed` file should contain the `chrom` `MN908947.3`

3.2.2. Multi fasta

```
>MN908947.3
ATTAAAGGTTTATACCTTCCCA...
>NC_006432.1
CGGACACACAAAAAGAAAGAAA...
```

> The corresponding `primer.bed` file should contain the `chrom` `MN908947.3` and `NC_006432.1`

3.3. Best practices

As the `reference.fasta` is often used for referenced-based assembly, using high quality genome with minimal `Ns` or ambiguous bases is advisable. Using RNA sequences in the `reference.fasta` is not advice, as DNA is expected.