Data analysis procedure for SMURF-seq reads

Contents

1 Introduction

SMURF-seq is a protocol to efficiently sequence short DNA molecules on a long-read sequencer by randomly ligating them to form long molecules. The SMURF-seq protocol involves cleaving the genomic DNA into short fragments. These fragmented molecules are then randomly ligated back together to form artificial, long DNA molecules. The long re-ligated molecules are sequenced following the standard MinION library preparation protocol. After (or possibly concurrent with) sequencing, the SMURF-seq reads are mapped to the reference genome in a way that simultaneously splits them into their constituent fragments, each aligning to a distinct location in the genome (for most fragments).

This manual explains how to map SMURF-seq reads, generate copy-number profiles from the mapped fragments, and perform additional optional analysis of the sequenced read or the mapped fragments.

2 Mapping SMURF-seq reads

At this point, we assume that the reads generated from a SMURF-seq experiment are base-called and are in a fastq or fasta file.

The reads sequenced using SMURF-seq protocol needs to be mapped to the reference genome to identify the fragment locations. The reads can be aligned leveraging long-read mapping tools that are designed for split-read alignment.

We present the option of aligning SMURF-seq reads using BWA-MEM [1], Minimap2 [2], and LAST [3], and we present several parameter recommendations for each tool.

2.1 Prerequisites

Reference genome: The CNV analysis procedure described in section 3 makes use of the human reference genome build `hg19`, and thus this build has to used to generate CNV profiles using the procedure described here. However, other reference genomes can be utilized if the user does not used the procedure in 3

Software required:

1. Mapping tool: One of BWA, Minimap2, or LAST.
2. samtools [4] (version: 1.9)

Environment variables:

2.2 Mapping SMURF-seq reads to the reference genome

A user has an option of using either of the tools listed above following the procedure in section 2.2.1, 2.2.2, or 2.2.3 respectively. We recommend using BWA-MEM as this produced higher fragment counts.

2.2.1 Mapping SMURF-seq reads with BWA-MEM

Reference genome index creation

Aligning SMURF-seq reads

[1] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[2] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 1:7, 2018.

[3] Szymon M Kielbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, pages gr–113985, 2011.

[4] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.