harp

Haplotype-Assisted Read Parsing

what

This script performs allele-specific alignment of a set of reads to two reference genomes. It outputs sam files for each genome, which can be used in many downstream analyses.

how

The script in this package takes a set of single/paired-end fastq files, aligns them two reference genomes using bowtie2, and parses each read based on which genome they best align to. If a read aligns to both genomes, but not equally-well, the alignment to both genomes must be unique (have a high alignment score for both genomes) in order to avoid mismappings due to SNPs.

why

This was originally written to perform allele-specific alignment of repli-seq data in castaneus-musculus hybrid mouse cells. It has also been used for parsing Hi-C, RNA-seq, and ChIP-seq data. This R package contains the implementation of the read-parsing algorithm used in the manuscript:

Allele-specific control of replication timing and genome organization during development. Juan Carlos Rivera-Mulia, Andrew Dimond, Daniel Vera, Claudia Trevilla-Garcia, Takayo Sasaki, Jared Zimmerman, Catherine Dupont, Joost Gribnau, Peter Fraser and David M. Gilbert

who

This software was written by Daniel Vera (vera@genomics.fsu.edu)

software requirements

This software has only been tested on centos7 and ubuntu trusty, but is expected to work on most modern linux-based systems with the following software installed and in your $PATH: - bowtie2 - samtools >1.3 - gawk - GNU coreutils - R >3

And the following R packages should be installed: - devtools >1.13 (R package)

input requirements

- fastq files to parse
- bowtie2 indices for each haplotype.

installation

```
# in R:
devtools::install_github("dvera/harp")
```

usage

```
# make bowtie2 indices for each haplotype, assuming you have a fasta file for each
haplotype, where each differs only by SNPs:

mkdir bowtie2index && cd bowtie2index
bowtie2-build /path/to/genome1.fa genome1
bowtie2-build /path/to/genome2.fa genome2

# navigate to a directory with your fastq files
cd /path/to/fastqFiles

# open R
R
```

in R:

```
library(harp)
# define a vector of fastq files to parse
f <- files("*.fastq")

ref1 <- "/path/to/bowtie2index/genome1"
ref2 <- "/path/to/bowtie2index/genome2"

harp( f, index1prefix=ref1, index2prefix=ref2 )
```

output

The script will generate a series of files for each input fastq file: - *_unmapped.sam (did not map well to either genome) - *_parsed_genome1.sam (parsed to genome1) - *_parsed_genome2.sam (parsed to genome2) - *_ambiguous.sam (mapped well to at least one genome, but could not be confidently parsed)