GEAN manual

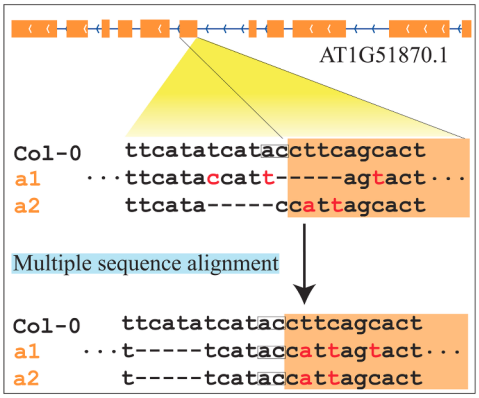
Baoxing Song

June 12, 2019

Contents

1 GEAN

Here we provide a solution for INDEL inconsistent alignment problem which could lead to false positive splice sites disturb or ORF-shift predication. And whole genome MSA is all developped basing on the genetic features.



By soving this problem, GEAN could also use to transform the well annotated genetics feature of model species to the genome of other natural variation individuls or phylogenetically nearby species with whole genome avaliable.

The inconsistent alignment problem could affect the function impact annotation of INDELs (non-coding INDEL V.S. ORF-shift INDEL), SNP (non-coding region SNP, coding region SNP), by re-alignment, those variants could be moved to non-coding regions.

1.1 Install

1.1.1 Dependencies

CPU support avx2 GNU GCC >=6.0 Cmake >= 3.0 Due to the high computational density of weighted sequence alignment algorithm, GEAN only fullly works on hadware platform with CPU support AVX2 constructions. Some functions could work on most of hardware plarform. As long as you are not using a very old machine, AVX2 should be valiable.

1.1.2 Installation

```
git clone https://github.com/baoxingsong/GEAN.git
cd GEAN
cmake CMakeLists.txt
make
```

1.2 Usage

1.2.1 Examples

example on different purpose and using genome with different genome complexity could be found on our github: https://github.com/baoxingsong/GEAN/tree/master/example

```
/home/bs674/software/bin/gean
```

```
## Program gean
## Usage: gean <command> [options]
## Commands:
##  -- variant calling:
##     pseudogeno  create pseudo genome sequence
##     lift        transform coordinate to another accession
```

```
##     revlift     transform coordinate of another accession to reference
##     liftgff     transform all the GFF/GTF coordinates
##     revliftgff  transform all the GFF/GTF coordinates back to reference
##     reanva      update variants records for functional annotation
##     gff2seq     get the protein/CDS/gene sequence of GFF/GTF file
##     annowgr     annotate re-sequenced genome
##     randomVar   assign a random position for each variant
##
##  -- whole genome wide MSA:
##     premsa      cut the whole genome sequence into fragments
##     msatosdi    generate sdi files from MSA results
##
##  -- de novo assembly genome:
##     transgff    trans reference gff/gtf to de novo assembly genome
##     spltogff    trans reference gff/gtf to de novo assembly genome using sam file
##     purifygff   purify the result from transgff
##     sinsyn      keep syntenic genes priorly and single copy genes (for inner species)
##     sinsyn2     keep syntenic genes priorly and single copy genes (for inter species)
##     quotasyn    quota syntenic blocks
##     orf         keep only ORF conserved genes
##     varcall     variant calling for de novo genome sequence
```

1.2.2 variant calling

Those functions are designed for whole-genome resequencing variant calling data. It works very for sdi file, which is a very simple file format. For VCF format, you should make sure there is no heterozygous variant calling result. If your are working heterozygous line, you could do phasing and seperate your variant calling result into two or more VCF files. please make sure you have only those variants records pass quanlity control in the input file. Since the VCF file format really diverse from different variant calling software, we could not make our software work with all the VCF files, it is highly recommanded to reform yoyr vcf file into sdi file.

1.2.2.1 get pseudo genome sequence using reference genome sequence and variant calling result

```
/home/bs674/software/bin/gean pseudogeno
```

```
## Usage: gean pseudogeno -r reference -v variants -o output
## Options
##  -h         produce help message
##  -r FILE    reference genome in fasta format
##  -v FILE    variant calling result in vcf/sdi format
##  -p STRING  prefix for vcf records
##  -o FILE    output pseudo genome in fasta format
```

** -prefix is the prefix of chromosome name for vcf/sdi variant records. Like the chromosome in TAIR10 reference genome is Chr1, Chr2, Chr3, Chr4 and Chr5. While the chromosomes in vcf files from the 1001 genomes project were indicated with 1, 2, 3, 4 and 5. So `-prefix Chr` should be set to make the software work properly. If this parameter is not set correctly, the software would act as no variant records in the input vcf/sdi file.

1.2.2.2 liftover reference coordinate to pseudo-genome-sequence

Project/liftover a certain reference genome-sequence coordinate to re-sequencing accession/line pseudo-genome-sequence.

```
/home/bs674/software/bin/gean lift
```

```
## Usage: gean lift -v variants -c chromosome -p position
## Options
##  -h         produce help message
##  -r FILE    reference genome in fasta format
##  -v FILE    variant calling result in vcf/sdi format
##  -f STRING  prefix for vcf records
##  -c STRING  chromosome
##  -p INT     the position/coordinate in reference genome
```

1.2.2.3 liftover pseudo-genome-sequence coordinate to reference genome sequence

Project/liftover a certain coordinate of re-sequencing accession/line pseudo-genome-sequence to reference genome-sequence.

```
/home/bs674/software/bin/gean revlift
```

```
## Usage: gean revlift -v variants -c chromosome -p position
## Options
##  -h         produce help message
##  -r FILE    reference genome in fasta format
##  -v FILE    variant calling result in vcf/sdi format
##  -f STRING  prefix for vcf records
##  -c STRING  chromosome, should be consistent with the chromosome information in sdi file (The coordi
##  -p INT     the position/coordinate in re-sequenced genome
```

1.2.2.4 liftover reference gff/gtf/gff3 annotation to pseudo-genome-sequence

Inference the gene structure (gtf/gff file) annotation of re-sequencing accession/line by purely coordinate liftover.

```
/home/bs674/software/bin/gean liftgff
```

```
## Usage: gean liftgff -v variants -i inputGffFile -o outputGffFile
## Options
##  -h         produce help message
##  -r FILE    reference genome in fasta format
##  -v FILE    variant calling result in vcf/sdi format
##  -i FILE    the input GFF/GTF file of reference line/accession
##  -f STRING  prefix for vcf records
##  -o         the output GFF/GTF file of target line/accession
```

1.2.2.5 liftover pseudo-genome-sequence gff/gtf/gff3 annotation to reference genome sequence

Project/liftover the gene structure (gtf/gff file) annotation of re-sequencing accession/line to reference genome-sequence by purely coordinate liftover.

```
/home/bs674/software/bin/gean revliftgff
```

```
## Usage: gean revliftgff -v variants -i inputGffFile -o outputGffFile
## Options
##  -h         produce help message
##  -r FILE    reference genome in fasta format
```

```
##  -v FILE    variant calling result in vcf/sdi format
##  -i FILE    the input GFF/GTF file of non-reference line/accession
##  -f STRING  prefix for vcf records
##  -o         the output GFF/GTF file of reference line/accession
```

1.2.2.6 re calling variant calling by align the genic region sequencing and keep the completeness of ORF

Realign the sequence using ZDP algorithm to solve the inconsistent INDEL alignment problem and recall all variants which could cause false positive ORF-state shit predication.

```
/home/bs674/software/bin/gean reanva
```

```
## Usage: gean reanva -i inputGffFile -r inputGenome -a similar segments -s new genome sequence -o outpu
## Options
##  -h        produce help message
##  -i FILE   GFF/GTF file
##  -r FILE   reference genome sequence
##  -v FILE   variant calling result in vcf/sdi format
##  -f STRING prefix for vcf records
##  -o FILE   output file
##  -m INT    minimum intron size
```

- By ORF-states, this software has following criteria:
  1) Splicing sites is one of motif in "SpliceSites", which is included in the release
  2) The minimum length of intron is larger than a certain value
  3) CDS sequence length is larger than a certain value
  4) The length of CDS sequence is divisible by 3
  5) No premature stop codon
  6) End with end codon
  7) Start with start codon The IUPAC Codes of DNA sequence could be well dealt with. The result of ORF-states are included in the CDS sequence

1.2.2.7 extract sequece using genome sequence and annotation file

Extract CDS sequence, C-DNA sequence and protein sequence for each protein-coding transcript. And predict the protein coding potential (termed as ORF-state)

```
/home/bs674/software/bin/gean gff2seq
```

```
## Usage: gean gff2seq -i inputGffFile -r inputGenome -p outputProteinSequences -c outputCdsSequences -g
## Options
##  -h        produce help message
##  -i FILE   reference genome in GFF/GTF format
##  -r FILE   genome sequence in fasta format
##  -m INT    minimum intron size for ORF stats checking
##  -p FILE   output file of protein sequence in fasta format
##  -c FILE   output file of CDS (without intron) in fasta format
##  -g FILE   output file of CDS (with intron) in fasta frormat
```

1.2.2.8 annotate the pseudo-genome-sequence

Transform the reference gene structure annotation to re-sequencing accession/lines with several complementary methods.

```
/home/bs674/software/bin/gean annowgr
```

```
## Usage: gean annowgr -i inputGffFile -r referenceGenomeSequence -v variants -o outputGffFile
## Options
##  -h        produce help message
##  -i FILE   reference GFF/GTF file
##  -n FILE   the de novo annotation GFF of the target accession
##  -r FILE   reference genome in fasta format
##  -v FILE   variant calling result in vcf/sdi format
##  -o FILE   the output GFF/GTF file
##  -m INT    minimum intron size
##  -d        remove reference ORF shift transcripts (default false)
##  -f STRING prefix for vcf records
##  -t INT    number of threads, default: 4
##  -l INT    longest transcript to align. default(50000)
```

1.2.2.9 simulate random variants

Assign a random position for each variant in a variant calling result file, which could be used to compare the different between observed variant calling and random variants.

```
/home/bs674/software/bin/gean generateRandomSdi
```

```
## Program gean
## Usage: gean <command> [options]
## Commands:
##  -- variant calling:
##     pseudogeno  create pseudo genome sequence
##     lift        transform coordinate to another accession
##     revlift     transform coordinate of another accession to reference
##     liftgff     transform all the GFF/GTF coordinates
##     revliftgff  transform all the GFF/GTF coordinates back to reference
##     reanva      update variants records for functional annotation
##     gff2seq     get the protein/CDS/gene sequence of GFF/GTF file
##     annowgr     annotate re-sequenced genome
##     randomVar   assign a random position for each variant
##
##  -- whole genome wide MSA:
##     premsa      cut the whole genome sequence into fragments
##     msatosdi    generate sdi files from MSA results
##
##  -- de novo assembly genome:
##     transgff    trans reference gff/gtf to de novo assembly genome
##     spltogff    trans reference gff/gtf to de novo assembly genome using sam file
##     purifygff   purify the result from transgff
##     sinsyn      keep syntenic genes priorly and single copy genes (for inner species)
##     sinsyn2     keep syntenic genes priorly and single copy genes (for inter species)
##     quotasyn    quota syntenic blocks
##     orf         keep only ORF conserved genes
##     varcall     variant calling for de novo genome sequence
```

1.2.3 Whole genome wide multiple sequence alignment pipeline

1.2.3.1 cut the (pseudo-)genome sequence of a population of individuals into fragments to perform multiple sequence alignment for each fragment.

```
/home/bs674/software/bin/gean premsa
```

```
## Usage: gean premsa -i inputGffFile -r referenceGenomeSequence -v variants
## Options
##  -h         produce help message
##  -i FILE    the input GFF/GTF file of reference line/accession
##  -r FILE    reference genome
##  -v FILE    list of variant calling results files
##  -f STRING  prefix for vcf records
##  -m INT     minimum intron size
##  -t INT     number of threads, default: 4
##  -w INT     window size, default: 10000
##  -s INT     window overlap size, default: 500
##  -p INT     output catch size (default 100)
##  -l INT     longest transcript to align. default(50000)
```

1.2.3.2 variant calling using multiple sequence alignment

perform variant calling from the multiple sequence alignment of sequence fragments of a population of genome sequences

```
/home/bs674/software/bin/gean msatosdi
```

```
## Usage: gean msatosdi -a accessionList -c chromosomeLi -m MSAresultFolder -o outputFolder  -r referen
## Options
##  -h         produce help message
##  -c FILE    chromosome list
##  -m FOLDER  folder of MSA result
##  -o FOLDER  output folder
##  -r FILE    reference genome in fasta format
##  -t INT     number of threads, default: 4
##  -v FILE    list of variant calling results files
##  -f STRING  prefix for vcf records
```

1.2.4 project reference annotation to de novo assemly genome sequence

pipeline to project the reference gene structure annotation to a de novo assembly genome sequence highly similar with the reference genome sequence #### transgff liftover reference gene structure annotation to a de novo assembly genome sequence using whole genome sequence alignment. The result file contains duplication gene annotations records, which might do not compile with other software and could be purified with the following function.

```
/home/bs674/software/bin/gean transgff
```

```
## Usage: gean transgff -i inputGffFile -r inputGenome -a similar segments -s new genome sequence -o ou
## Options
##  -h        produce help message
##  -i FILE   reference GFF/GTF file
##  -r FILE   reference genome sequence
##  -a FILE   similar segments
```

```
##  -s FILE   target genome sequence
##  -o FILE   output GFF/GTF file
##  -w INT    sequence alignment window width (default: 60)
##  -sl       run in slow model (default false)
##  -l INT    longest transcript to align. default(50000)
##  -m INT    minimum intron size
```

1.2.4.1 purifygff

remove those duplication gene structure annotations generated from the transff function

```
/home/bs674/software/bin/gean purifygff
```

```
## Usage: gean purifygff -i inputGffFile -s inputGenome -o output GFF/GTF file
## Options
##  -h        produce help message
##  -i FILE   GFF/GTF file
##  -s FILE   target genome sequence
##  -o FILE   output GFF/GTF file
##  -x INT    minimum gene length
##  -m INT    minimum intron size
```

1.2.5 Acknowledgements

The GEAB team would like to thank all our enthusiastic users who have contacted us with suggestions to improve the codebase, request new functions, point out bugs, and beta-test the initial versions of GEAN. Many thanks also to everyone who has used GEAN and cited the publication – we are glad it has proven useful in your research, and a good citation record will help us to obtain future funding to keep developing GEAN.

1.2.6 Citation

If you use GEAN, please cite: Baoxing Song, Qing Sang, Hai Wang, Huimin Pei, Fen Wang and Xiangchao Gan. (2019) A weighted sequence alignment strategy for gene structure annotation lift over from reference genome to a newly sequenced individual. bioRxiv. doi:10.1101/615476