P-Clouds Program Documentation

Wanjun Gu, A.P. Jason de Koning, Kathryn Hall, and David D. Pollock.

Beta version 0.9a. Updated December 19, 2011.

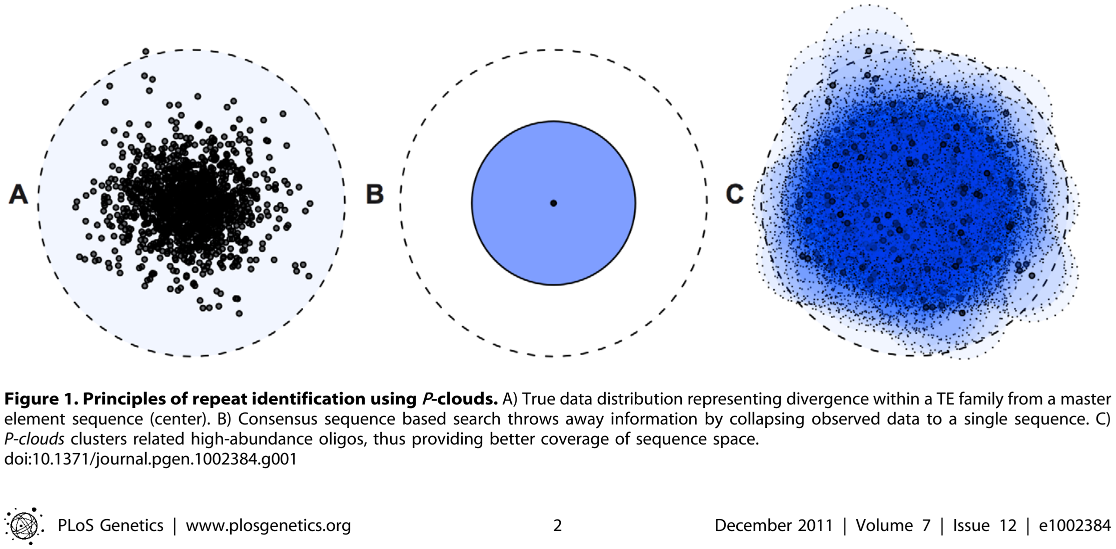

Figure 1. Principles of repeat identification using *P*-clouds. A) True data distribution representing divergence within a TE family from a master element sequence (center). B) Consensus sequence based search throws away information by collapsing observed data to a single sequence. C) *P-clouds* clusters related high-abundance oligos, thus providing better coverage of sequence space.
doi:10.1371/journal.pgen.1002384.g001

PLoS Genetics | www.plosgenetics.org 2 December 2011 | Volume 7 | Issue 12 | e1002384

For the most up-to-date version, please visit:

http://www.EvolutionaryGenomics.com/ProgramsData/PClouds/PClouds.html

Requests, questions, or bug reports should be directed to:

Wanjun Gu (wanjun.gu@gmail.com)
Kathryn Hall (midnightline@yahoo.com)
Jason de Koning (jason.de.koning@gmail.com)
David Pollock (David.Pollock@UCDenver.edu).

Introduction

The *P-Clouds* package is designed to rapidly detect repeated regions in whole eukaryotic genomes. There are three principle phases, which must be run in sequence. The first phase counts oligonucleotides in the entire genome (or for element-specific analyses, in the data set of known copies of a given repetitive element family). The second phase arranges oligonucleotides into related clusters, or repeat "probability clouds". The final "genome dissection" phase takes a set of *P*-clouds and annotates the genome by finding stretches of repeat regions having a high density of oligos that occur in *P*-clouds.

The main innovations are associated with the second and third phases (although some hard work was involved in creating a rapid and flexible counting program that works well with limited RAM), which achieve speed-ups relative to a self-self BLAST search by completely avoiding alignment and similarity searching. The method is described in detail in Gu *et al.* (2008), while the 'element-specific' *P-clouds* approach is described in de Koning *et al.* (2011).

How to Use P-Clouds

There is a single control file "Controlfile", which specifies parameters that can be varied. Variables include the oligonucleotide length, thresholds for *P*-cloud creation (see Gu *et al.,* 2008), cutoffs for defining repeat regions, input and output file names, and switches that determine which of the three phases will be run. The program is written in C and includes a "Makefile" to automate compilation. At this time, *P-Clouds* has been tested on Linux and Mac OS X. **Once the program is compiled by running 'make', the *P-Clouds* binary ('Pclouds.out') should be placed in a folder with the control file and the necessary inputs.** There should be sufficient hard disk space for the output: for a large eukaryotic genome, this will be around 20 GB.

Input Files

The main input file is specified in the controlfile with '**CountGenome**' (*e.g.*, CountGenome=chr20.up), which should refer to a specially formatted file containing nucleotide sequences to be analysed for repetitive content. When multiple sequences or contigs are being used, they must be combined using the pre-processor program to a single file (see below).

Three Phase Analysis

Selection of which of the three phases of the *P-Clouds* analysis are to be run can be specified using the **CALCOUNTS**, **GETPCLOUDS**, and **DISSECTION** parameters, respectively for Phases 1-3. For a straightforward analysis, these can all be set to 1, which will result in a complete analysis that sequentially executes each of the three phases.

The output of the count enumeration phase (Phase 1) is called "out_lumped.txt", which is then used as the input for creating the *P*-clouds (Phase 2). Each line contains an oligonucleotide sequence observed more than once, followed by its count in the input sequence(s). After construction of the *P*-clouds in Phase 2, the file given by '**GenomeInput'** (*e.g.*, GenomeInput=chr20.up) is then annotated for repeat regions in Phase 3. This is often the same as the original input genome, but it could be another genome sequence to be annotated if desired.

Primary Output

The main outputs of the genome analysis are stored in the file indicated by '**RepeatRegion'** (*e.g.*, RepeatRegion=chr20_cleaned**.region**), which contains the start and stop points for each annotated repetitive region of the genome that fits the window size and percent cutoff parameters set in the control file. These coordinates refer to the combined input file, if multiple contigs were provided. To convert these coordinates to refer to individual contigs that were combined with the pre-processor, the post-processor program must be used (see below).

Auxilliary Output

Other outputs include files specified by '**CloudAnnotation**' (*e.g.*, CloudAnnotation=chr20_cleaned.cloud'), which contains the space separated number of associated *P*-clouds for every site in the genome. Information on the *P*-clouds themselves are stored in: "**mainclouds.info**", which contains the *P*-cloud number, the 'seed' repeat, and the number of oligonucleotides in the *P*-cloud; "**mainclouds.assign**", which contains each core oligonucleotide (*i.e.,* each oligo with count greater than "endthreshold") followed by the *P*-cloud number it is assigned to; and "**accclouds.assign**", which contains all oligonucleotides in the outer layer of the cloud, (*i.e.,* with repeat numbers less than endthreshold but greater than "copythreshold"; see Gu *et al.,* 2008 for detailed descriptions of these parameters).

Pre-processing input files

If multiple contigs are to be analysed simultaneously, they need to be combined into a single file using the preprocessor. To compile the pre-processor:

```
    cd preprocessor; make
```

The pre-processor can then be run on any FASTA formatted file to create an input file suitable for *P-Clouds* using:

```
    preprocessor <fastafile> <output filename>
```

Post-processing output annotation files

When the pre-processor has been used, the .region output files refer to coordinates on the combined files. To convert these coordinates back to coordinates on the original sequences, the post-processor must be used. To compile the post-processor:

```
    cd postprocessor; make
```

The post-processor can then be run on the original FASTA formatted file and the region file output by *P-Clouds* Phase 3 using:

```
    postprocessor <fastafile> <region file> <annotation file (output)>
```

References

Gu W, Castoe TA, Hedges DJ, Batzer MA, Pollock DD (2008). Identification of repeat structure in large genomes using repeat probability clouds. *Anal Biochem.* 380(1):77-83.

de Koning APJ, Gu W, Castoe TA, Batzer MA, and DD Pollock (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics* 7(12): e1002384.