BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 5 2009, pages 666–667
doi:10.1093/bioinformatics/btp001

Sequence analysis

Disperse—a software system for design of selector probes for exon resequencing applications

J. Stenberg[1,2], M. Zhang[1] and H. Ji[1,3,*]

[1]Department of Medicine, Division of Oncology, Stanford University School of Medicine, Clark Center W300, 318 Campus Drive, Stanford, CA 94305-5440, USA, [2]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, 751 85 Uppsala, Sweden and [3]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304, USA

Received on April 27, 2008; revised on December 20, 2008; accepted on December 30, 2008

Advance Access publication January 21, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: Selector probes enable the amplification of many selected regions of the genome in multiplex. Disperse is a software pipeline that automates the procedure of designing selector probes for exon resequencing applications.

Availability: Software and documentation is available at http://bioinformatics.org/disperse

Contact: genomics_ji@stanford.edu

The advent of next-generation sequencing systems has brought with it interesting new possibilities for genomic research (Shendure *et al*., 2004). One application is resequencing selected regions of the human genome. To accomplish this, methods for sample preparation are required that allow the selective enrichment of many arbitrary regions of the genome.

Several methods have been proposed to this end, such as hybridization of fragmented genomic DNA to oligonucleotide arrays, followed by washing and amplification of the retained material (Albert *et al*., 2007; Okou *et al*., 2007), and various probe-based approaches using PCR to achieve multiplex amplification (Dahl *et al*., 2007; Fredriksson *et al*., 2007; Porreca *et al*., 2007).

One of these approaches, the selector technique, allows amplification of arbitrarily selected restriction fragments in a single-primer pair PCR (Dahl *et al*., 2005). Selectors are oligonucleotide constructs that consist of a general sequence motif flanked by target-specific end sequences. Each selector targets a single restriction fragment and guides the circularization of this fragment by ligation, incorporating the general sequence motif into all fragments. Either the complete fragment is circularized, or a portion of the fragment is cleaved off using the endonucleolytic cleavage activity of a polymerase, and the 3′ part of the fragment is circularized. All targeted fragments are then amplified in multiplex. In previous work, we have demonstrated how this can be used for amplification of a large set of exons for subsequent resequencing using massively parallel pyrosequencing (Dahl *et al*., 2007).

Designing selector probes for exon resequencing of a set of genes has previously involved querying databases through web interfaces and running a series of command line utilities and interactive programs. This procedure has been time consuming and not readily reproducible, which prompted us to develop an integrated software system, Disperse, to facilitate this task.

Given a list of genes as HUGO gene symbols, and a set of design parameters, Disperse will generate a set of selector probe sequences, designed to select the largest possible portion of the targeted sequence. The design work is performed in a pipelined fashion, where a number of steps are executed sequentially, each step utilizes the results of previous steps. Each step is implemented as a Perl script or a command line Java program, and all steps can be executed in the order using a pipeline script. The design parameters and the input set of gene names are defined in text files, and the intermediate and final results are also stored as text files. The steps are as follows:

(1) Determine coordinates for all coding regions of target genes by lookup in a local copy of CCDS data (http://www.ncbi.nlm.nih.gov/projects/CCDS/). For genes not found, use NCBI eUtils (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) to acquire GenBank (Benson *et al*., 2008) files and extract coding sequence coordinates. If any gene is not found in this step, halt the execution.

(2) Define regions of interest (ROIs) by extending all coding regions by a number of bases on each flank, and merging overlapping ROIs.

(3) Use the fastacmd program to retrieve reference sequence for ROIs and flanking regions from a local Blast database file (http://www.ncbi.nlm.nih.gov/blast).

(4) Extract SNP data for target regions from local SNP data file containing an extract of dbSNP (Sherry *et al*., 2001).

(5) Collate sequence and SNP data by creating a copy of the target file with SNP data included using degeneracy symbols.

(6) Select a reaction combination and generate a list of all accepted fragments for the combination using the PieceMaker API (application programming interface) (Stenberg *et al*., 2005a).

(7) Select a subset of the generated fragments, retaining optimal coverage of the targeted regions using the PieceMaker API.

*To whom correspondence should be addressed.

(8) Generate an amplicon output file listing the circularized parts of the fragments of the selected subset, including their genomic coordinates.

(9) Assemble selector probe sequences using the ProbeMaker (Stenberg *et al*., 2005b) command line interface.

(10) Generate consolidated output files to provide overview of the design.

By default, all these steps are carried out when the pipeline script is executed. If manual intervention is required at some stage, e.g. to add a region of interest outside coding regions, the script can be set to run only a subset of the steps at a time.

The results of a design job are affected by the design parameters and the pool of restriction reactions to choose from. The design parameters determine, among other things, acceptable fragment lengths and the length of the ends of the selector probes. Reactions, which may contain one or more restriction enzymes, are picked from a pool of reactions defined by the user. The size and contents of this pool, and the maximum number of reactions allowed, will affect the design results as measured by the portion of the ROIs that is covered by selected fragments. The likelihood of finding acceptable fragments that cover all ROIs will increase with the size of the pool, and with the number of reactions picked. Also, the time required to generate and evaluate fragments will increase linearly with the number of reactions, while the time required to select a reaction combination will increase with the number of possible reaction combinations, which depends on both the number of reactions in the pool, and the maximum number of reactions allowed.

To demonstrate the performance of the software, we designed a set of selector probe sequences for the coding regions of the 206 genes on chromosome 18 present in the CCDS data. We used similar design settings as in our previous work (Dahl *et al*., 2005), allowing selected fragments of length 100–200 bases, with a max flap length of 500. We allowed a combination of up to three reactions to be picked from a pool of 120 reactions. This design took 42 min on a 2.4 GHz Intel Quad-core processor computer running 64-bit openSUSE 10.3 with a 32-bit Java virtual machine, and resulted in 90.7% coverage of the targeted region using 5519 selector probes. The results are provided on the project web site.

To investigate the effects of using a smaller reaction pool, we performed the design using a randomly selected subset of 24 reactions from the pool of 120 reactions. This was completed in 8 min, yielding a slightly smaller coverage of 89.2%. Another run, picking up to six reactions from the smaller pool, resulted in 98.2% coverage and took 34 min. We estimate that a design allowing six reactions to be picked from the larger pool would yield close to full coverage, but would take several months to complete on a typical workstation. The conclusion is that for each design, it is reasonable to do some form of prescreening of enzymes and reactions, for example by analyzing the length distribution of generated fragments, before running the design pipeline.

Disperse depends on external data sources, programs and libraries. To our knowledge, all required data sources are freely available, as are the external dependencies, with the exception of the PieceMaker program (version 1.3.2) which is freely available for non-commercial users only. Instructions for acquiring PieceMaker are available on the Disperse web site. Disperse is written in Java and Perl, and has been tested on machines running openSUSE Linux, Gentoo Linux, Windows XP and MacOS X, using Perl 5 and Java 1.6.

ACKNOWLEDGEMENTS

We wish to thank Magnus Isaksson, Department of Genetics and Pathology, Uppsala University, for assistance with software testing.

Funding: National Institue of Health (5K08CA96879–6, 2P01HG000205, 5R21CA128485 to H.P.J.); Reddere Foundation (to H.P.J.); Liu Bie Ju Cha and Family Fellowship in Cancer (toH.P.J.); Olle Engkvist Byggmästare Foundation (to J.S.).

Conflict of Interest: none declared.

REFERENCES

Albert,T.J. *et al*. (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–905.

Benson,D.A. *et al*. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.

Dahl,F. *et al*. (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.*, **33**, e71–e77.

Dahl,F. *et al*. (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl Acad. Sci. USA*, **104**, 9387–9392.

Fredriksson,S. *et al*. (2007) Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.*, **35**, e47–e53.

Okou,D.T. *et al*. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, **4**, 907–909.

Porreca,G.J. *et al*. (2007) Multiplex amplification of large sets of human exons. *Nat. Methods*, **4**, 931–936.

Shendure,J. *et al*. (2004) Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.*, **5**, 335–344.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Stenberg,J. *et al*. (2005a) PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. *Nucleic Acids Res.*, **33**, e72–e77.

Stenberg,J. *et al*. (2005b) ProbeMaker: an extensible framework for design of sets of oligonucleotide probes. *BMC Bioinformatics*, **6**. 229–235.