4C-Seq primer database

Tanay, deLaat Groups

Version 0.7, July 2012

1. Contents

The 4cseq_primer_db consists of 12 tables (6 for the mouse mm9 genome assembly, 6 for the human hg18 genome assembly) detailing all the recommended primer pairs for use in 4C-Seq experiments. Each table represents a different primary and secondary digestion combination, including *Csp6I-BfaI*, *DpnII-BfaI*, *DpnII-Csp6I*, *NlaIII-BfaI*, *NlaIII-Csp6I*, and *NlaIII-DpnII*.

2. Primer design

4C primer pairs carry additional 5’ overhangs composed of the adapter sequences (obtained from Illumina technical support) necessary for Illumina single read sequencing (GA-II and Hi-seq 2000). The strategy therefore produces sequencing reads (36-mers) composed of the 4C primer sequence (20 nucleotides, specific for a given viewpoint), followed by 16 nucleotides that identify a captured sequence. The reading primer always hybridizes to, and ends at the 3’ side of, the entire first restriction recognition site. This design ensures analysis of primary ligation events only and provides sufficient sequence information to unambiguously identify most captured sequences (mapability of 16-mers directly adjacent to a given four basepair site is 68%, using *Nla*III and *Dpn*II as restriction enzyme combination). The non-reading primers, with a size between 18 to 27 bp, were designed at a distance of ≤120 bp from the secondary restriction site.

PCR primers were designed taking into account the following additional rules. The viewpoint fragment preferably had a size of at least 500 bp to allow efficient cross-linking to other DNA fragments. The fragment end (the nucleotide sequences of the viewpoint fragment between the primary and secondary restriction site to which both 4C primers hybridize) was at least 300 bp and preferably more than 350 bp, to allow efficient circularization during the second ligation step. Primer3 (Rozen, Methods Mol Biol 2000) was used to find the optimal primer pair for a given viewpoint fragment, with the following adaptations to the default settings: optimal temperature of 55°C, minimum of 45°C and maximum of 65°C, GC content between 35% and 65%. Primers were checked against the mouse genome with megablast (Zhang, J Comput Biol 2000) (settings -p 88.88 -W 12 -e 1 -F T) requiring primers on the reading side to be matched uniquely in the genome, and primers in the non-reading side to have a maximum of three perfectly matching Blast high-scoring segment pairs (HSP). Both primers were also required to have fewer than 30 HSPs with an identity of at least 88.88%(16/18bp).

3. Table structure

Each primer database table contains the following fields:

chr_alias – Chromosome number
loc_XXXX_site – Starting genomic coordinate of primary restriction site (digested by enzyme XXXX)
loc_YYYY_site – Starting genomic coordinate of secondary restriction site (digested by enzyme YYYY)
XXXX_frag_size – Length of fragment between two primary restriction sites
XXXX_YYYY_frag_size – Length of fragment end between adjacent primary and secondary restriction sites
self_lig_size_add – Describes the size of PCR product generated by amplification of the 4C template resulting from a primary restriction fragment that underwent self-ligation. *XXXX_YYYY_frag_size* + *self_lig_size_add* = final molecule size
not_cut_XXXX_size_add - Describes the size of PCR product generated by amplification of the 4C template resulting from non-digestion of the primary restriction site. *XXXX_YYYY_frag_size* + *not_cut_XXXX_size_add* = final molecule size

re_name_XXXX – Name of primary restriction enzyme XXXX
pr_seq_XXXX – Forward primer sequence (lies directly on primary ligation junction)
hit_100_XXXX – Number of sites in the genome that perfectly match forward primer sequence
hit_89_XXXX – Number of sites in the genome that match forward primer sequence by >89%
re_ name_YYYY – Name of secondary restriction enzyme YYYY
pr_seq_YYYY – Reverse primer sequence (lies upstream of the secondary restriction site, within the viewpoint fragment end)
hit_100_YYYY – Number of sites in the genome that perfectly match reverse primer sequence
hit_89_YYYY – Number of sites in the genome that match reverse primer sequence by >89%
pp_id – Primer pair ID number
pr_id_XXXX – Forward primer ID number
pr_id_YYYY – Reverse primer ID number
latin_name – Latin name of the organism being studied
ucsc – the genome assembly to which primers were designed

4. Illumina adapter sequence

In order to generate 4C-Seq products that can directly by sequenced by Illumina machines, the following sequences must prefix the sequences found in the primer database tables:

Single-End:
AATGATACGGCGACCACCGAACACTCTTTCCCTACACGACGCTCTTCCGATCT | *Reading adapter*
CAAGCAGAAGACGGCATACGA | *Non-reading adapter*

Paired-End:
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT | *Reading adapter PE1*
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT | *Reading adapter PE2*