CaptureProbe: a java tool for designing probes for capture Hi-C applications

DEAR EDITOR,

Many functional elements associated with traits and diseases are located in non-coding regions and act on distant target genes via chromatin looping and folding, making it difficult for scientists to reveal the genetic regulatory mechanisms. Capture Hi-C is a newly developed chromosome conformation capture technology based on hybridization capture between probes and target genomic regions. It can identify interactions among target loci and all other loci in a genome with low cost and high resolution. Here, we developed CaptureProbe, a user-friendly, graphical Java tool for the design of capture probes across a range of target sites or regions. Numerous parameters helped to achieve and optimize the designed probes. Design testing of CaptureProbe showed high efficiency in the design success ratio of target loci and probe specificity. Hence, this program will help scientists conduct genome spatial interaction research. CaptureProbe and source code are available at https://sourceforge.net/projects/captureprobe/.

Genome level studies on traits and diseases in different organisms have revealed that the majority of associated genetic loci are located in non-coding regions and are enriched in different regulatory signals, thus suggesting their regulatory functions (Maurano et al., 2012; Welter et al., 2014; Zhang et al., 2014). Regulatory elements can act on multiple genes and distant target genes via chromatin looping (Maston et al., 2006). Therefore, elucidation of the regulatory mechanisms of these non-coding loci is not reliable when applying simple assignment to the nearest genes. Chromosome conformation capture with high-throughput sequencing (Hi-C) allows for the identification of physical chromatin interactions across an entire genome (Lieberman-Aiden et al., 2009). However, the enormous complexity of Hi-C libraries makes it costly to obtain sufficient spatial resolution to detect interactions among specific elements. To circumvent these issues, capture Hi-C technology with capture probes was developed to reduce the target regions for sequencing in order to identify interactions among target loci and all other loci in a genome at low cost (Mifsud et al., 2015; Sahlén et al., 2015; Schoenfelder et al., 2015). This technology has been used extensively in different studies to reveal the regulatory mechanisms of traits or disease-associated loci in non-coding regions (Baxter et al., 2018; Mishra & Hawkins, 2017). The design of capture probes is a necessary prerequisite for capture Hi-C experiments and can be complex work for researchers without programming experience.

Several software tools have been designed for capture Hi-C probes, including CapSequm (Davies et al., 2016), HiCapTools (Anil et al., 2018), and GOPHER (Hansen et al., 2019). These toolkits are important in capture Hi-C-related analysis but cannot meet all requirements of diverse experiments. For instance, CapSequm, which is a web application for designing capture probes, can only process 1 000 positions at a time and provides very limited design parameters (e.g., probe length, restriction enzyme). HiCapTools was designed to find probes for target sites, but not for target regions, which are very common candidates for genetic research. In addition, HiCapTools contains limited parameters, which reduces its flexibility when considering specific DNA sequencing contexts. Furthermore, it is a command-line program and requires a series of input files, and thus is not very user friendly. The recently developed program GOPHER can design capture probes for both target sites and regions and includes a user-friendly graphic user interface (GUI). However, its capture probe design capacity is currently limited to human and mouse.

In this study, we developed CaptureProbe, a Java tool with a graphical user-friendly interface that can design capture probes for both target genetic sites and regions without species limitation. CaptureProbe is easy to use, only requires

Received: 15 November 2019; Accepted: 26 November 2019; Online: 11 December 2019

Foundation items: This work was supported by the Ministry of Agriculture of China (2016ZX08009003-006) and Animal Branch of the Germplasm Bank of Wild Species, Chinese Academy of Sciences (Large Research Infrastructure Funding)

DOI: 10.24272/j.issn.2095-8137.2020.010

simple input files, and provides abundant parameters for probe design. Moreover, it can also give detailed statistical information about design results. Comparisons between CaptureProbe and other existing tools showed that it provides rich software functions and shows better or equivalent performance in designing capture probes.

To achieve good performance in capturing informative ligation fragments, CaptureProbe designs probes based on the structural features of the Hi-C library. The Hi-C library consists of ligated restriction fragments originally in close spatial proximity in the nucleus (Lieberman-Aiden et al., 2009). Usually, these ligation fragments are sheared to a specific size range to ensure suitability for high-throughput sequencing. Therefore, CaptureProbe designs probes to capture both ends of the target restriction fragment (overlapping target sites or regions) and selects probes nearest to the end of the target restriction fragment. The program initially starts probe design from both ends of the target restriction fragment and moves inward by 1 bp for each cycle. To improve capture efficiency and specificity of probes, CaptureProbe calculates the GC content and missing and repeated bases (missing bases: n/N, repeated bases: marked in lowercase) in the probe sequence and chooses the first probe that meets all parameter limitations provided by the user. CaptureProbe can avoid redundancy probe sequences caused by target site overlap in the same restriction fragment.

Running CaptureProbe is very simple, requiring only the coordinate file of the target sites/regions and the sequence file (fasta format). CaptureProbe can limit any repeated sequences marked in the sequence files when designing the probes. Users can employ windows to directly specify the path of the required files and to set parameters. All real-time configuration information can be printed for users to check progress. After running, CaptureProbe can print detailed information on the results of the capture probe design for users to evaluate the results. CaptureProbe will generate a series of result files for users to customize probes and to check the design state of each target site or region.

We systematically compared software function and probe design performance between CaptureProbe and other existing tools (Tables 1–2). As CapSequm function is limited, comparison analysis was not included. Both CaptureProbe and GOPHER showed rich functions and user-friendly GUI (Table 1).

Table 1 Functional comparisons among CaptureProbe and other tools

Tool	CaptureProbe	HiCapTools	GOPHER
Supported species	All	All	Human, Mouse
Type of target loci	Site/Region	Site	Site/Region
GUI	√	×	√
Detailed probe information	√	×	√
Probe GC content limitation	√	×	√
Probe missing base limitation	√	×	×
Repeated sequence limitation	√	√	×
Design margin limitation	√	×	√
Fragment size limitation	√	×	√
Fragment missing base limitation	√	×	×
Mapping score limitation	×	√	√

Table 2 Comparisons of capture probe design among CaptureProbe and other tools

Tool	CaptureProbe	HiCapTools	GOPHER
Testing site number (*n*)	20 000	20 000	20 000
Both ends with probes (%)	42.40	21.26	85.76
Only upstream with probe (%)	18.77	23.14	2.67
Only downstream with probe (%)	19.28	23.77	3.02
Total sites with probes (%)	80.45	68.17	91.45
Total sites with no probes (%)	19.57	31.83	8.56
Probe GC content <25% (%)	0.00	3.03	0.00
Probe GC content >65% (%)	0.00	0.34	0.00
Probe with extreme GC content (%)	0.00	3.37	0.00
Probe with unique alignment (%)	92.84	77.44	83.34
Probe with multiple aligments (%)	7.10	22.31	16.41
Probe with no alignment (%)	0.06	0.25	0.25

In this study, we only evaluated design performance for target sites as the mechanism is the same for target sites and regions. Twenty thousand random target sites (not from gap regions) in the human genome (hg38) were generated for testing. The same parameters were set for all tools: i.e., probe length, 120 bp; repeat sequence length, 6 bp; restriction enzyme, Hand III; minimal fragment length, 300 bp; design margin size, 500 bp; probe GC content, 25%–65%; with all other parameters set using default values. Firstly, we compared the design success ratio among the three programs. GOPHER showed the highest design success ratio (91.45%), followed by CaptureProbe (80.45%), and finally

HiCapTools (68.17%). We next accessed the specificity of the probes, using BLASTN (Altschul et al., 1990) to align all probes to the genome sequence. CaptureProbe demonstrated the highest ratio of unique alignment (92.84%) among the programs (GOPHER: 83.34%, HiCapTools: 77.44%). As HiCapTools could not filter GC content in the probe sequences, partial probes of HiCapTools (3.37%) showed extremely high GC content (<25% or >65%), which did not match the efficient capture range (Agilent Technologies). Furthermore, we also found that small probes from GOPHER contained ambiguous characters (N).

Here, we present a very simple and user-friendly Java tool (CaptureProbe) that facilitates rapid capture probe design for target chromosome capture applications with no species limitation. CaptureProbe provides rich software functions and shows good probe design performance. Comparisons with existing software demonstrated that CaptureProbe has a good design success ratio and better probe specificity. CaptureProbe will be useful for a wide range of scientists studying genome spatial interactions.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Y.F.M., Y.P.Z., and H.B.X. designed the research. Y.F.M. implemented the Java code and analyzed the data. Y. F. M., Y. P. Z., and H. B. X. wrote the paper. A. C. A. and Y. B. S. revised and edited the manuscript. All authors read and approved the final version of the manuscript.

Yun-Fei Ma[1,2,3], Adeniyi C. Adeola[1], Yan-Bo Sun[1], Hai-Bing Xie[1,*], Ya-Ping Zhang[1,4,*]

[1] *State Key Laboratory of Genetic Resources and Evolution*, *and Yunnan Laboratory of Molecular Biology of Domestic Animals*, *Kunming Institute of Zoology*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650223, *China*

[2] *Kunming College of Life Science*, *University of Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650204, *China*

[3] *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*

[4] *State Key Laboratory for Conservation and Utilization of Bio-resource*, *and Key Laboratory for Animal Genetic Diversity and Evolution of High Education in Yunnan Province*, *Yunnan University*, *Kunming*, *Yunnan* 650091, *China*

*Corresponding authors, E-mail: xiehb@mail.kiz.ac.cn; zhangyp@mail.kiz.ac.cn

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3): 403−410.

Anil A, Spalinskas R, Åkerborg Ö, Sahlén P. 2018. HiCapTools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications. *Bioinformatics*, **34**(4): 675−677.

Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, Simigdala N, Martin LA, Andrews S, Wingett SW, Assiotis I, Fenwick K, Chauhan R, Rust AG, Orr N, Dudbridge F, Haider S, Fletcher O. 2018. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nature Communications*, **9**(1): 1028.

Davies JO, Telenius JM, Mcgowan SJ, Roberts NA, Taylor S, Higgs DR, Hughes JR. 2016. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods*, **13**(1): 74−80.

Hansen P, Ali S, Blau H, Danis D, Hecht J, Kornak U, Lupiáñez DG, Mundlos S, Steinhaus R, Robinson PN. 2019. GOPHER: Generator of probes for capture Hi-C experiments at high resolution. *BMC Genomics*, **20**(1): 40.

Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950): 289−293.

Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, **7**: 29−59.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**(6099): 1190−1195.

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, Leproust E, Follows GA, Fraser P, Luscombe NM, Osborne CS. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, **47**(6): 598−606.

Mishra A, Hawkins RD. 2017. Three-dimensional genome architecture and emerging technologies: looping in disease. *Genome Medicine*, **9**(1): 87.

Sahlén P, Abdullayev I, Ramsköld D, Matskova L, Rilakovic N, Lötstedt B, Albert TJ, Lundeberg J, Sandberg R. 2015. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, **16**(1): 156.

Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, Nagano T, Katsman Y, Sakthidevi M, Wingett SW, Dimitrova E, Dimond A, Edelman LB, Elderkin S, Tabbada K, Darbo E, Andrews S, Herman B, Higgs A, Leproust E, Osborne CS, Mitchell JA, Luscombe NM, Fraser P. 2015. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, **25**(4): 582−597.

Welter D, Macarthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, **42**(D1): D1001−1006.

Zhang X, Bailey SD, Lupien M. 2014. Laying a solid foundation for Manhattan--'setting the functional basis for the post-GWAS era'. *Trends in Genetics*, **30**(4): 140−149.