

The Jackson Laboratory The Mouseion at the JAXlibrary

Faculty Research 2019

Faculty Research

1-14-2019

GOPHER: Generator Of Probes for capture Hi-C Experiments at high Resolution.

Peter Hansen

Salaheddine Ali

Hannah Blau

The Jackson Laboratory, hannah.blau@jax.org


Daniel Danis

The Jackson Laboratory, daniel.danis@jax.org

Jochen Hecht

See next page for additional authors

Follow this and additional works at: <https://mouseion.jax.org/stfb2019>

 Part of the [Life Sciences Commons](#), and the [Medicine and Health Sciences Commons](#)

Recommended Citation

Hansen, Peter; Ali, Salaheddine; Blau, Hannah; Danis, Daniel; Hecht, Jochen; Kornak, Uwe; Lupiáñez, Darío G; Mundlos, Stefan; Steinhaus, Robin; and Robinson, Peter N, "GOPHER: Generator Of Probes for capture Hi-C Experiments at high Resolution." (2019). *Faculty Research 2019*. 17.

<https://mouseion.jax.org/stfb2019/17>

This Article is brought to you for free and open access by the Faculty Research at The Mouseion at the JAXlibrary. It has been accepted for inclusion in Faculty Research 2019 by an authorized administrator of The Mouseion at the JAXlibrary. For more information, please contact ann.jordan@jax.org.

Authors


Peter Hansen, Salaheddine Ali, Hannah Blau, Daniel Danis, Jochen Hecht, Uwe Kornak, Darío G Lupiáñez, Stefan Mundlos, Robin Steinhaus, and Peter N Robinson

SOFTWARE

Open Access



GOPHER: Generator Of Probes for capture Hi-C Experiments at high Resolution

Peter Hansen¹, Salaheddine Ali², Hannah Blau³, Daniel Danis³, Jochen Hecht⁴, Uwe Kornak^{1,5}, Darío G. Lupiáñez⁶, Stefan Mundlos^{1,2,5}, Robin Steinhaus¹ and Peter N. Robinson^{3,7*} 

Abstract

Background: Target enrichment combined with chromosome conformation capturing methodologies such as capture Hi-C (CHC) can be used to investigate spatial layouts of genomic regions with high resolution and at scalable costs. A common application of CHC is the investigation of regulatory elements that are in contact with promoters, but CHC can be used for a range of other applications. Therefore, probe design for CHC needs to be adapted to experimental needs, but no flexible tool is currently available for this purpose.

Results: We present a Java desktop application called GOPHER (Generator Of Probes for capture Hi-C Experiments at high Resolution) that implements three strategies for CHC probe design. GOPHER's `simple` approach is similar to the probe design of previous approaches that employ CHC to investigate all promoters, with one probe being placed at each margin of a single digest that overlaps the transcription start site (TSS) of each promoter. GOPHER's `simple-patched` approach extends this methodology with a heuristic that improves coverage of viewpoints in which the TSS is located near to one of the boundaries of the digest. GOPHER's `extended` approach is intended mainly for focused investigations of smaller gene sets. GOPHER can also be used to design probes for regions other than TSS such as GWAS hits or large blocks of genomic sequence. GOPHER additionally provides a number of features that allow users to visualize and edit viewpoints, and outputs a range of files useful for documentation, ordering probes, and downstream analysis.

Conclusion: GOPHER is an easy-to-use and robust desktop application for CHC probe design. Source code and a precompiled executable can be downloaded from the GOPHER GitHub page at <https://github.com/TheJacksonLaboratory/Gopher>.

Keywords: Gene regulation, Nuclear organization, Promoter-enhancer interactions, Capture Hi-C, Java

Background

Functional elements that are widely separated in the linear sequence of the genome can be brought into contact with one another by the folding of the genome in three-dimensional space. A series of extensions of the original targeted chromosome conformation capture (3C) method that was introduced in 2002 [1] culminated in Hi-C, a global method for interrogating chromatin interactions that combines formaldehyde-mediated cross-linking of

chromatin with fragmentation, DNA ligation, and high-throughput sequencing to characterize interacting loci on a genome-wide scale [2]. Hi-C has been used to investigate the large scale organizational architecture of the genome, revealing the existence of megabase-sized local chromatin interaction domains termed topologically associating domains (TADs) [3]. Owing to the complexity of Hi-C libraries, it is not feasible to investigate interactions between specific gene promoters and their distal regulatory elements. For instance, roughly 100 million reads are required to obtain 40kb resolution [4]. Given that a linear increase of resolution requires a quadratic increase in total sequencing depth [5], obtaining the 5kb or better resolution that is desirable for investigating individual promoter-enhancer interactions would

*Correspondence: peter.robinson@jax.org

³The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, United States

⁷Institute for Systems Genomics, University of Connecticut, Farmington, CT 06032, United States

Full list of author information is available at the end of the article



be costly. Recently, capture Hi-C (CHC) and capture-C methodologies were developed as alternative approaches to overcome these difficulties. These techniques employ a hybridization technology similar to exome capture that enriches Hi-C libraries for viewpoint sequences representing loci of interest using biotinylated cRNA probes.

CHC has been used in a variety of experimental settings to provide more in-depth data for specific loci than would be feasible with Hi-C. For example, promoter CHC focuses on the enrichment of gene promoters in order to identify functional interactions with distal regulatory elements such as enhancers [6–10]. Other applications include the investigation of the potential regulatory effects of disease-associated single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS). Of note, the majority of these so-called GWAS hits are located in non-coding and likely regulatory sequences, whose effects are, in the lack of further evidence, commonly assigned to the nearest gene. CHC has suggested the inaccuracy of these assumptions by showing that some distal interactions are associated with stronger effects on expression than interactions with neighboring genes, thereby providing strong evidence that altered regulation of a distal gene underlies the mechanism of certain GWAS hits [11–18]. In particular, one study on 1999 SNPs associated with cardiovascular disease revealed that more than 90% of the SNP-target gene interactions did not involve the nearest neighbor, and 40% of the SNPs displayed interactions with two or more genes [19], demonstrating the value of CHC for understanding disease biology.

CHC has also been used to analyze gene regulation programs in differentiation and disease [20–22] by profiling interactions across large genomic regions and by characterizing the effects of structural variation on chromatin organization. For instance, one study investigated the effects of genomic duplications on the TAD architecture of the genome using CHC and 4C-seq methods, and showed that duplications can result in the formation of new chromatin domains (neo-TADs) with pathologic alterations of gene regulation [23].

CHC employs a set of biotinylated oligonucleotides that are designed to hybridize to and 'capture' target sequences; such oligonucleotides are usually referred to as baits or probes. Several technologies are commercially available for capture of exonic sequences in exome sequencing [24]. These methods can be adapted for CHC by means of a custom design for probes that hybridize to promoter sequences or other desired CHC target regions. Because of the diversity of CHC applications, users are faced with the challenge of designing probes for specific experimental settings.

To our knowledge, only two tools are available for capture Hi-C probe design. CapSequm [24] is a web application that can be readily used through a web browser, but the number of viewpoints is limited to 1000 viewpoints at a time. HiCapTools [25] overcomes this limitation, but is a command-line tool that needs to be compiled from source. Both CapSequm and HiCapTools implement an approach to probe design similar to what we call the 'simple approach' in this manuscript, and do not implement features that would be required to design probes according to the simple-patched and extended strategies that we introduce in this manuscript.

Here we present GOPHER (*Generator Of Probes for capture Hi-C Experiments at high Resolution*), an easy-to-use Java-based desktop application that provides a suite of methods and visualization tools for the automated design and subsequent manual curation of viewpoints. GOPHER enables all steps required for probe design to be performed in a unified framework that leads users from the download of the genome, alignability, and transcript files, through the choice of parameters such as target genes or regions, restriction enzymes, and desired thresholds for GC content, alignability, and digest length. Users can inspect the genomic context of each of the generated viewpoints, and can add or remove digests (restriction fragments) if desired. GOPHER implements three main approaches to probe design, including two that have not previously been available. GOPHER outputs a series of files including a probe file that can be used to order probes (baits) for the enrichment of the targeted regions in capture Hi-C experiments. Additionally, summary statistics are generated that can be used for documentation of the final design. Users can generate a digest file containing attributes of the selected and unselected digests relevant for downstream analysis.

Results

We present an easy-to-use software application for the design of CHC probes that uses one of three approaches and allows users to set a wide range of parameters for different experimental situations. GOPHER implements three main strategies for probe design. The `simple` approach generates probes that are similar to those used for many previously published capture Hi-C studies: One digest is selected for each target region (often including a transcription start site of a gene), and two probes are placed at the outermost ends of the digest. The `simple-patched` approach "patches" viewpoints that are poorly covered by single digests. GOPHER additionally implements a new approach to probe design that we term `extended`, which is intended to provide greater resolution than the simple approach by performing restriction digestion with a 4-cutter instead of 6-cutter and selecting sets of multiple fragments per target region. In

general, the simple and simple-patched approaches are best suited for investigations of larger numbers of targets such as a promoterome in which all promoters of all coding genes are investigated [7, 8, 10], whereas the extended approach is more suited to investigate smaller numbers of genes (e.g., 500–1000) involved in a biological process of interest [6, 24, 26]. All approaches are also suitable for other categories of target regions such as GWAS hits or larger blocks of genomic sequence.

Data preparation and parameter settings

In order to design CHC probes, users need to download and preprocess a substantial amount of sequence and annotation data. GOPHER provides a graphical user interface (GUI) to streamline these tasks (Fig. 1a). Various genome builds for human and mouse can be selected from a drop-down menu, and downloading, unzipping and indexing of genome sequences can be performed with no software requirements other than a Java virtual machine

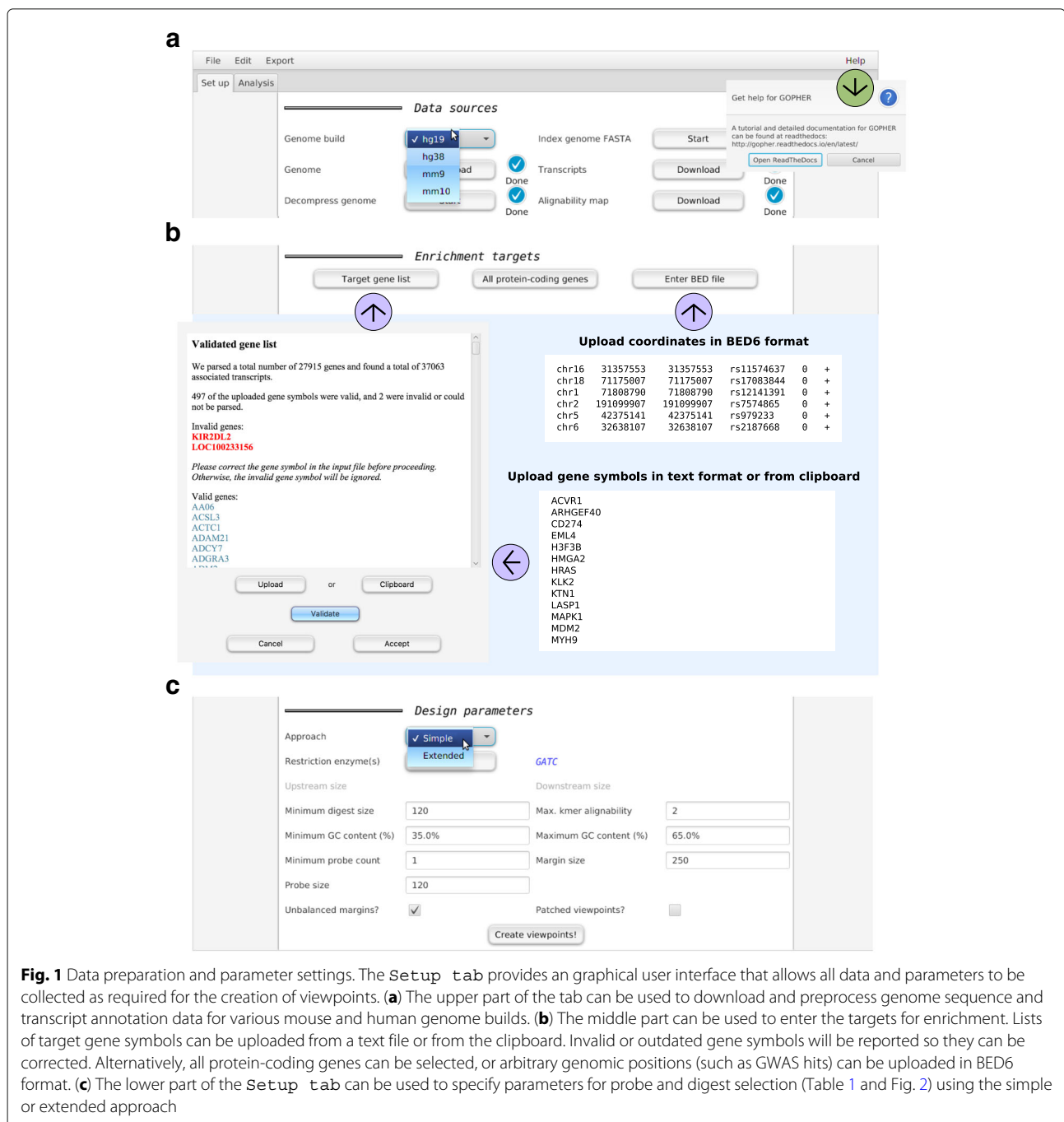


Fig. 1 Data preparation and parameter settings. The **Setup** tab provides a graphical user interface that allows all data and parameters to be collected as required for the creation of viewpoints. **(a)** The upper part of the tab can be used to download and preprocess genome sequence and transcript annotation data for various mouse and human genome builds. **(b)** The middle part can be used to enter the targets for enrichment. Lists of target gene symbols can be uploaded from a text file or from the clipboard. Invalid or outdated gene symbols will be reported so they can be corrected. Alternatively, all protein-coding genes can be selected, or arbitrary genomic positions (such as GWAS hits) can be uploaded in BED6 format. **(c)** The lower part of the **Setup** tab can be used to specify parameters for probe and digest selection (Table 1 and Fig. 2) using the simple or extended approach

(version 1.8). Furthermore, associated annotation data for transcription start sites and alignability are downloaded and parsed directly from the application. The progress of time-consuming steps such as indexing the genome file is indicated in the GUI. These steps have to be performed only once for a given genome build.

Following this, users specify the desired enrichment targets (Fig. 1b). For promoter CHC, gene symbols can be entered either from a text file or from the clipboard. GOPHER creates one viewpoint for all transcription start sites associated with the entered gene symbols. If gene symbols are used that do not occur in the downloaded annotation data, as can be the case if an invalid or outdated symbol is used (e.g., *P53* instead of the official gene symbol *TP53*), GOPHER will issue a warning and report a list of unmappable symbols that can be used to search for the current correct symbols. An alternative shortcut option allows promoters of all protein coding genes to be selected as targets. GOPHER also accepts a BED file with genomic positions. For instance, the coordinates of GWAS hits can be uploaded in BED6 format.

GOPHER allows the user to set a number of parameters that control the choice of viewpoints, digests, and probes (Table 1) using a graphical user interface (Fig. 1c). In the following sections, we describe how to choose parameters and how to visualize and edit viewpoints.

Selection of capture Hi-C probes and digests

Capture Hi-C probes must meet certain requirements that are substantially different from the those for standard use cases such as exome sequencing. Note that in this article, we refer to the DNA sequences produced by the sonication step of next-generation sequencing as *fragments*, and we refer to the DNA sequences produced by restriction digestion as *digests*. Within Hi-C libraries, interacting sequences are represented by hybrid molecules consisting of two pieces of digests from different genomic locations (Fig. 2a). The sonication step decreases the length of hybrid molecules, typically to around 300–500 bp. Therefore, valid interaction read pairs [26] map largely to the margins of digests adjacent to restriction enzyme cutting sites (Additional file 1: Figure S1). GOPHER takes this into account and places probes only within the margins of digests with a default size of 250 bp.

GOPHER considers alignability as well as GC content of probes (Fig. 2b). The mean k-mer alignability (Methods) of a probe reflects the average number of sequences in the target genome that are identical with k-subsequences of the probe. It is assumed that a higher k-mer alignability may increase the probability of unspecific cross hybridization of the probe to repetitive genomic sequences and thereby reduce the capture efficiency of the

Table 1 GOPHER parameters: The users may chose parameter settings that influence the design of probes and digests. In addition, approach-specific parameters can be chosen

Probe parameters	
Probe length	<i>Explanation:</i> Length of probes. <i>Default:</i> 120 bp
Minimum GC content	<i>Explanation:</i> The minimum proportion of G and C nucleotides. <i>Default:</i> 35%
Maximum GC content	<i>Explanation:</i> The maximum proportion of G and C nucleotides. <i>Default:</i> 65%
Alignability	<i>Explanation:</i> Maximum mean 50mer alignability. <i>Default:</i> 2
Digest parameters	
Margin size	<i>Explanation:</i> Width of the outermost ends of digests that will be tiled with probes. <i>Default:</i> 250 bp
Minimum digest size	<i>Explanation:</i> Smaller digests cannot be selected. <i>Default:</i> 120 bp
Minimum number of probes	<i>Explanation:</i> At least this number of probes have to be placed in each margin of a balanced digest. The total number of probes in both margins of unbalanced digest must be at least twice this value. <i>Default:</i> 1
Allow unbalanced margins	<i>Explanation:</i> Digest with unequal numbers of probes in each margin are selected during viewpoint creation. <i>Default:</i> <code>False</code>
Simple parameters	
Allow patching	<i>Explanation:</i> Digests that are not well centered at the TSS will be patched during viewpoint creation. <i>Default:</i> <code>False</code>
Extended parameters	
Maximum distance upstream	<i>Explanation:</i> Extension of the viewpoint in upstream direction <i>Default:</i> 5000 bp
Maximum distance downstream	<i>Explanation:</i> Extension of the viewpoint in downstream direction. <i>Default:</i> 1500 bp

probe. By default, GOPHER discards probes with mean k-mer alignabilities greater than 2; there is a tradeoff between the mean alignability threshold and the number

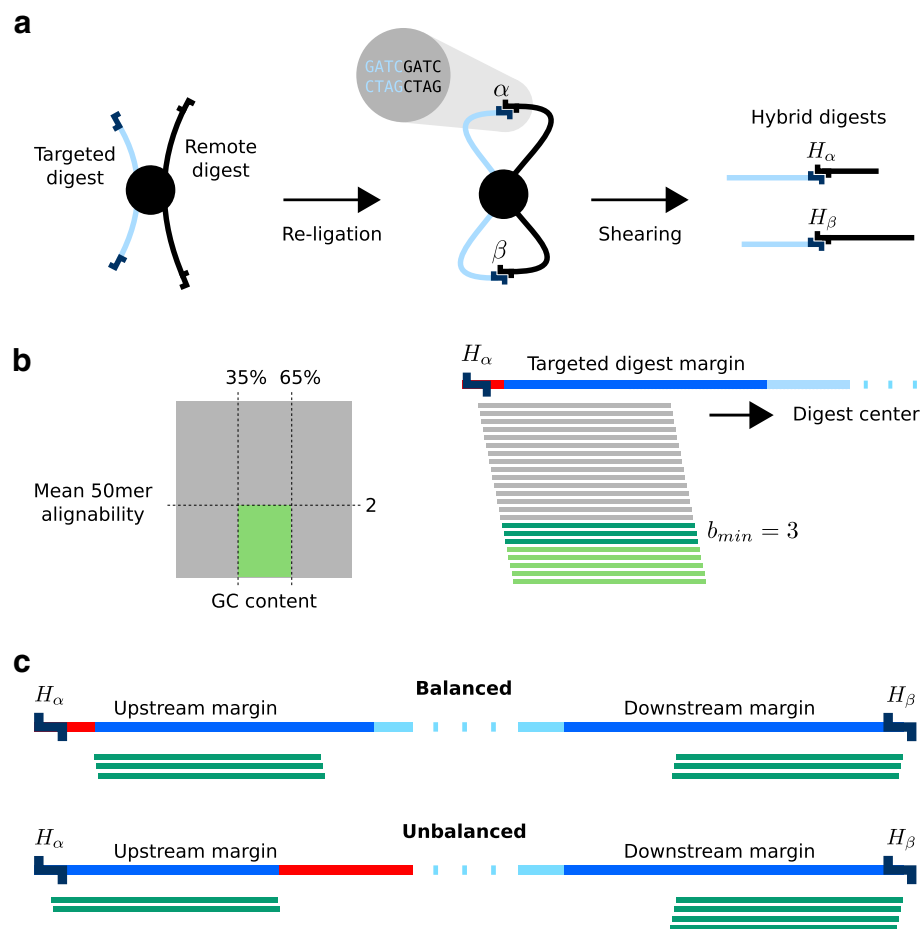


Fig. 2 Selection of probes and digests. **(a)** Idealized example of two cross-linked digests from a targeted region (light blue) and a remote interacting region (black). Re-ligation and shearing results in two hybrid digests H_α and H_β consisting of DNA from the targeted and a remote region. **(b)** We assume that the average length of the two parts corresponds to half of the average fragment length of sheared DNA in total. Therefore, only the margins of digests are defined to be target regions (blue). By default, GOPHER uses a margin size of 250 bp. For selection of usable probes only the uniqueness (alignability) of the probe sequence and GC content are taken into account. By default, usable probes are defined as those that have a mean 50mer alignability ≤ 2 and a GC content between 35 and 65% (light green area within square). GOPHER starts at the outermost end of targeted digests, moves towards the center and selects the first b_{min} usable probes (dark green). Regions for which no usable probes can be selected are depicted in red. **(c)** If b_{min} usable probes can be placed within each margin of a given digest, the digest is here referred to as balanced. Otherwise, if $2 \cdot b_{min}$ probes can be placed in both margins but with unequal numbers in the two margins, the digest is referred to as unbalanced. By default, GOPHER selects balanced digests only, and unbalanced digests can be manually selected after viewpoint creation, but if desired users can allow GOPHER to select unbalanced digests if no balanced digest can be found

of viewpoints for which probes can be designed, and the threshold can be adjusted by the user (Additional file 1: Figure S2). GOPHER restricts the GC content of selected probes between a lower threshold of 35% and an upper threshold of 65%, but these default thresholds can be adjusted by the user. For each margin of a given targeted digest, GOPHER starts at the outermost ends, moves towards the center and selects the first b_{min} usable probes. There is no restriction on the overlap between probes, because we reasoned that the sequences directly next to the cutting sites occur most likely within hybrid fragments (Additional file 1: Figure S1). Furthermore, complete tiling

of the margins is not an appropriate objective in this case. Therefore, if a margin contains more than one probe, it is often the case that the probes are only shifted by only 1 bp. The parameter b_{min} denotes the minimum number of probes (baits) necessary to select a digest for enrichment. By default, GOPHER demands that each of the two margins of a digest contain b_{min} probes; if this is the case, the digest is referred to as balanced. If the user allows unbalanced margins in the Setup tab of GOPHER, then any digest with at least $2 \cdot b_{min}$ valid probes will be selected. If the two margins do not have equal numbers of probes, then the digest is referred to as unbalanced

(Fig. 2c). GOPHER prefers balanced digests because they may be associated with a more even enrichment. However, if it is preferable for the experimental goals to have unbalanced digests rather than no digests at all for difficult sequences, then the user can select unbalanced margins or manually select individual digests after creation of viewpoints.

Viewpoint creation

Following data preparation and the choice of parameters, the user can click the **Create Viewpoints** button to cause GOPHER to read the genome sequence and alignability map in order to prepare an *in silico* digest and to evaluate each digest and candidate probe sequence with respect to *k*-mean alignability and GC content. A progress monitor tracks the creation of the viewpoints. Following this, the **Analysis** tab will be initialized to show a summary of the results and one row for each created viewpoint (Fig. 3). Users can click on individual viewpoints to show **Viewpoint editor** tabs that will be discussed below.

Creation of simple viewpoints

GOPHER's **simple** approach is intended for designs with a large number of target regions. In such cases the number of available probes may become a limiting factor. For instance, to capture the human promoters of protein-coding, noncoding, antisense, snRNA, miRNA and snoRNA transcripts about 22,000 *HindIII* restriction fragments (digests) were targeted with two probes each [7, 10]. Only one digest is targeted for each viewpoint; the digest that overlaps the transcription start site (TSS) is chosen if possible (Fig. 4). In many studies, the 6-cutter *HindIII* (~ 3700 bp) is employed for promoterome-wide investigations, but GOPHER allows a range of 6-cutters and 4-cutters such as *DpnII* (~ 430 bp) for different experimental goals. Depending on the cutting motif, some restriction enzymes may display a different distribution of digest sizes near to the transcription start sites. For instance, for *DpnII* the digests at TSS are on average 900 bp instead of 430 bp. Especially if 4-cutters are used (which tend to generate smaller digests than 6-cutters), we have observed that in some viewpoints, the digest

Viewpoint	Target	Genomic location	# Selected	Score	Active length	Total length	Fragment at TSS	Manually revised	Delete	Reset
Show	ABCA1	chr9:107,690,5...	1	100.00%	1122	1122	yes		Delete	Reset
Show	ABCC3	chr17:48,712,2...	1	100.00%	1740	1740	yes		Delete	Reset
Show	ACAP1	chr17:7,239,848	1	74.28%	1063	1063	yes		Delete	Reset
Show	ADAM12	chr10:128,077,...	1	86.35%	483	483	yes		Delete	Reset
Show	ADAM...	chr8:24,241,798	1	99.93%	815	815	yes		Delete	Reset
Show	ADAMTS7	chr15:79,103,8...	1	100.00%	3359	3359	yes		Delete	Reset
Show	ADAP2	chr17:29,248,7...	1	69.00%	175	175	yes		Delete	Reset
Show	ADCY9	chr16:4,166,186	1	100.00%	1843	1843	yes		Delete	Reset
Show	AGAP2	chr12:58,135,9...	1	100.00%	1034	1034	yes		Delete	Reset
Show	AGAP2	chr12:58,132,0...	0	0.0%	0	0	no		Delete	Reset
Show	AGPAT5	chr8:6,565,878	1	84.11%	960	960	yes		Delete	Reset
Show	AHR	chr7:17,338,276	0	0.0%	0	0	no		Delete	Reset
Show	AIM2	chr1:159,046,6...	1	98.45%	408	408	yes		Delete	Reset

Fig. 3 Simple viewpoint creation. Simple viewpoints can be created by clicking on **Create viewpoints!** after setting of appropriate parameters (Fig. 1c). Upon completion, the **Analysis** tab will be opened. At the top, summary statistics regarding the design are listed. In this case, GOPHER attempted to create simple viewpoints for 730 genes. GOPHER created at least one valid viewpoint (at least one selected digest) for 667 genes. Note that there are usually more viewpoints than genes, because one viewpoint for each TSS is created. For instance, two viewpoints were created for the gene *AGAP2*. If the simple approach is performed without patching, the mean size of viewpoints corresponds to the mean size of digests at TSS. Depending on the selected restriction enzyme, this size may be different from the mean size derived from all digests due to the different base composition in promoter regions. Overlapping viewpoints arising from multiple TSS on given digests lead to redundant digests and associated probes. GOPHER reports only the number of unique digests and does not export redundant digests. The unique digests are further classified as balanced and unbalanced. The number of probes and the capture size, i.e. the total region that is covered by probes, can be used for cost estimation. The table below the summary statistics contains information about individual viewpoints. Each viewpoint can be opened for visual inspection and editing. Manually adjusted adjusted viewpoints will be flagged and can be reset to their original state

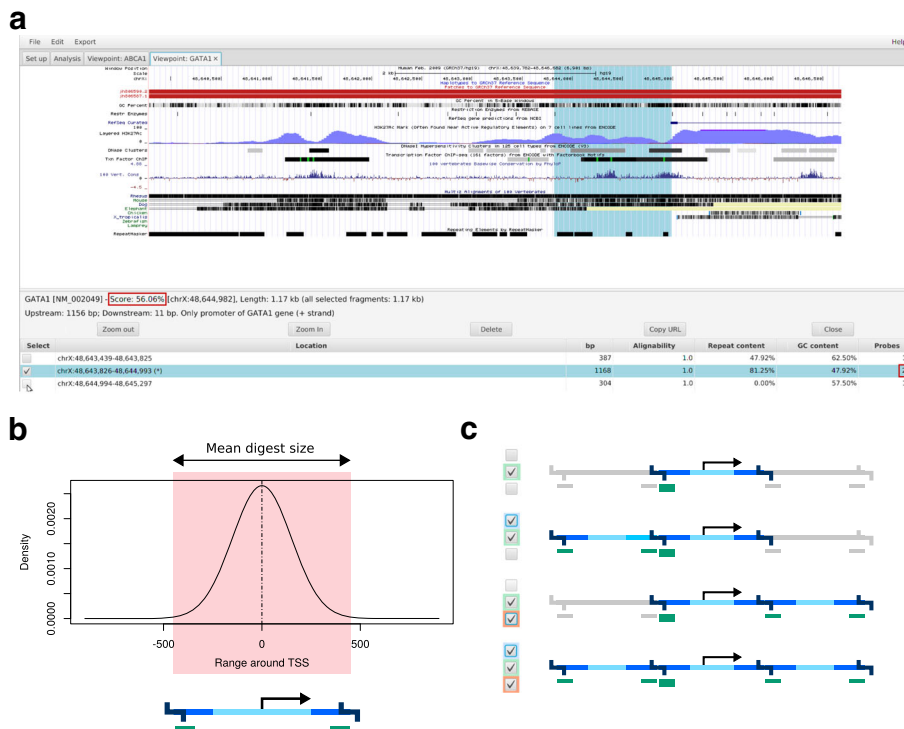


Fig. 4 Simple viewpoint creation. **(a)** From the **Analysis** tab (Fig. 3) each individual viewpoint can be opened in a separate tab for visual inspection. The upper part displays tracks from UCSC’s genome browser and can be used for evaluation and orientation during editing of viewpoints. In this case, the selected digest is not well centered at the TSS. Detailed information about the digest that contains the TSS (marked with an asterisk) and the two adjacent digests are shown below. The indicated information about alignability, GC and repeat content refers to selected probes. Note that in this case the digests containing the TSS is unbalanced due to high GC content at the downstream margin. **(b)** The score for simple viewpoints is close to 1 for digests that are not too short and well centered at the TSS, whereas it is close to 0.5 if the TSS occurs at the outermost ends of digests. Such viewpoints can be easily identified by sorting the viewpoint table in the **Analysis** tab by score. **(c)** The user can select and deselect each individual digest. For the *GATA1* viewpoint shown above, the adjacent downstream digest should be selected in order to center the viewpoint at the TSS

only barely overlaps the actual TSS, with a substantial amount of potentially important regulatory sequence (as judged by the presence of an H3K27Ac peak) being left out (Fig. 4a). GOPHER calculates a score for simple viewpoints that reflects how well the region around given TSS is covered by the associated digest (Fig. 4b). Viewpoints with poor coverage tend to have scores close of 0.5 or less and can be identified via sorting the table in the **Analysis** tab (Fig. 3). The **Viewpoint editor** tab allows the user to add additional adjacent digests by selecting the corresponding checkbox (Fig. 4c). With the **simple** approach, a total of three digests are shown, with the selected digest being in the middle. In some cases, the surrounding digests cannot be chosen because they are too short or no baits can be found which satisfy the chosen GC or alignability constraints. In this case, GOPHER shows “n/a” in red.

Simple patched viewpoints

The creation procedure of simple viewpoints may result in viewpoints that are not well centered at the TSS

and thus might miss relevant regulatory elements. In such cases adjacent digests can be additionally selected manually, which is time-consuming for larger numbers of viewpoints. Therefore, GOPHER provides the **simple patched** approach that automates the process of selecting the best digest (Fig. 5). First, simple viewpoints are generated as described above. For viewpoints whose score is less than 0.6, GOPHER tries to add one of the two adjacent digests. GOPHER selects the digest that is closer to the TSS if it satisfies length, alignability, and GC content criteria. After patching, the simple viewpoint score is recalculated, and poor-quality viewpoints can be identified by sorting as for the simple approach.

Extended viewpoints

Some published CHC studies target all promoters of the genome by placing single probes at the the outermost ends of TSS-containing *HindIII* restriction fragments [7, 8, 10, 27]. The tools CapSequm [6, 28] and HiCapTools [25] can be used generate probes for this class of experiment, and GOPHER’s simple and simple-patched approaches

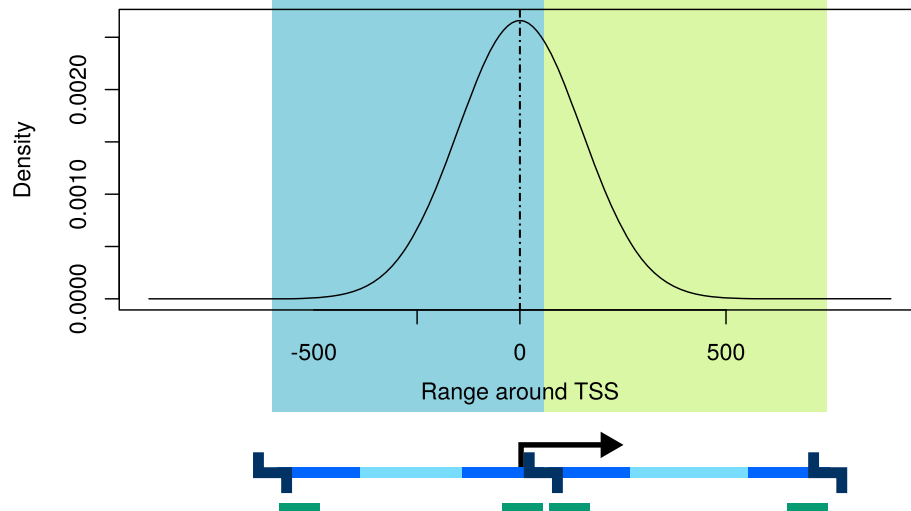


Fig. 5 Simple patched viewpoints. In the first run of simple viewpoint creation, only digests that contain the TSS are selected. This may result in viewpoints that are not well centered at the TSS. Such viewpoints can be manually revised by selecting adjacent digests in upstream or downstream direction (Fig. 4c). GOPHER provides an option referred to as “patching” that automates this process. If the score of a simple viewpoint is below 0.6, GOPHER will add the appropriate adjacent digest if possible

are mainly intended for this setting. On the other hand, some CHC studies targeted only a few hundred promoters for enrichment using the 4-cutter *DpnII* [6, 29]. These studies provide higher resolution and deeper coverage at the viewpoints.

We developed the extended approach to enable probe design for the latter class of experiment. GOPHER will select all digests that are located within or overlap with the window specified by upstream size and downstream size and that display a valid size, mean k-mer-alignability, and GC content. GOPHER calculates an empirical viewpoint score for extended viewpoints that can be used in the same way as the score for the simple and simple-patched approaches to rank the viewpoints and if desired prioritize viewpoints with low scores for manual inspection. Similar to the score for simple viewpoints, viewpoints that are well covered receive a good score, and selected digests near to the TSS contribute more to the score than others (Methods; Fig. 6).

Non-TSS applications

GOPHER can be used to target genomic regions in which GWAS hits are embedded. Users should generate a BED6 file [30] with the genomic coordinates of the GWAS hits and adjust the parameters such as restriction enzyme and probe criteria according to experimental goals. GOPHER can also be used to generate probes to “tile” larger genomic regions by choosing the extended approach with a BED file that contains a position at the center of each desired genomic region. The size of the genomic region is

controlled by adjusting the Upstream and Downstream size in GOPHER’s Setup tab.

GOPHER features

GOPHER shares a record of the project file in the user’s home directory, so that users can return to a probe design project at multiple occasions after saving the current results via the File menu. If desired, the project file can also be exported and shared with other GOPHER users. GOPHER outputs a number of files, including a documentation of the current project and results, a BED file with the chromosomal location of the selected digests and a probe file that can be used for ordering probes from commercial suppliers. We note that if two viewpoints share the same selected digest, the digest margins and probes that are output as input for probe design are made unique. The same is true for exported probes. GOPHER also can export a digest file with data on each of the in silico digests including information on GC content, alignability, length, and whether a probe was chosen or not. This file can be used to provide this data to downstream analysis programs for normalization or interaction calling. Table 2 provides an overview of GOPHER’s output files.

The allTracks.bed output file can be loaded as a custom track into UCSC’s genome browser [30]. This file contains five tracks: Genomic positions (e.g., the TSS), Viewpoints, Restriction fragments, Target regions and Probes (Additional file 1: Figure S3). The viewpoints are displayed in different grayscales that reflect the scores of viewpoints (black for

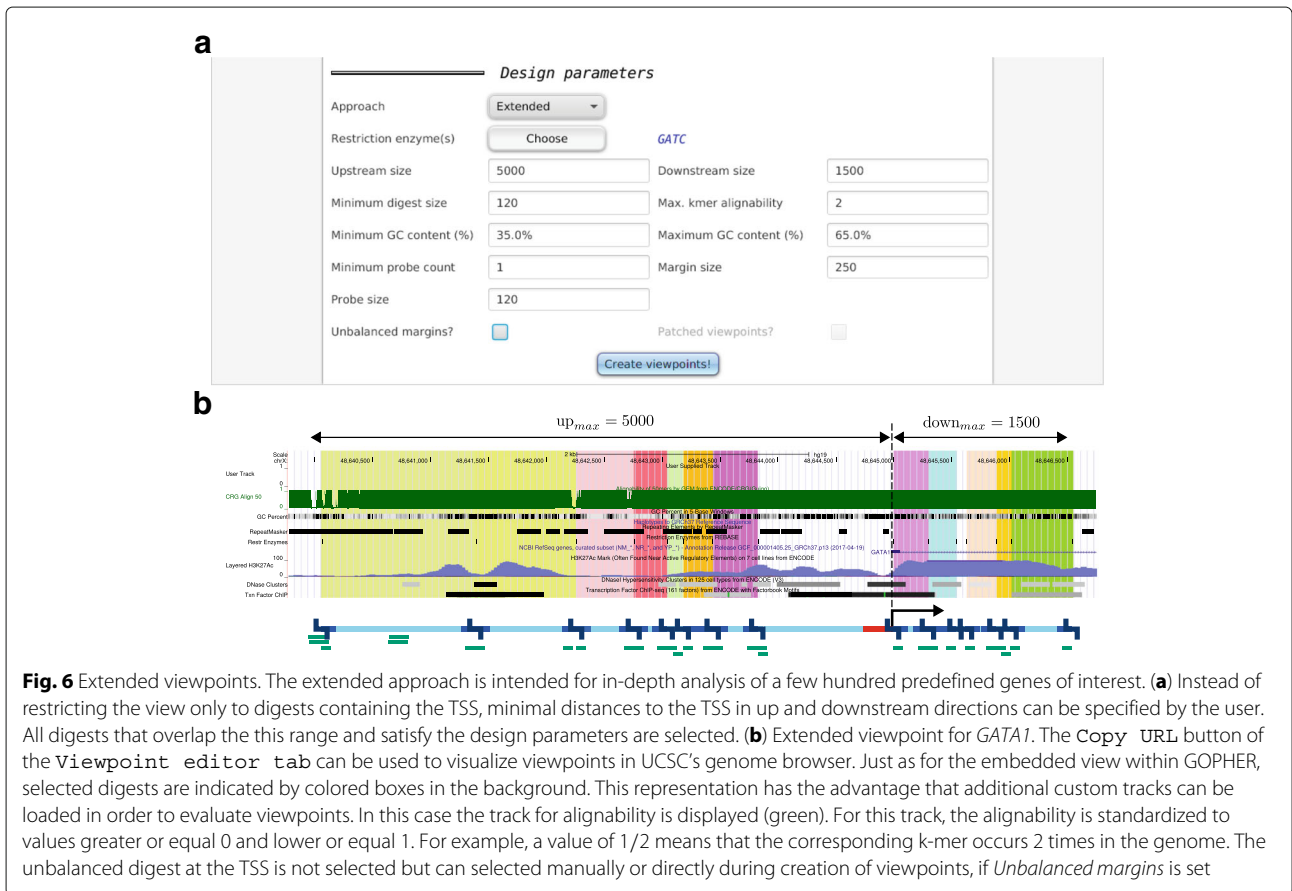


Fig. 6 Extended viewpoints. The extended approach is intended for in-depth analysis of a few hundred predefined genes of interest. **(a)** Instead of restricting the view only to digests containing the TSS, minimal distances to the TSS in up and downstream directions can be specified by the user. All digests that overlap the this range and satisfy the design parameters are selected. **(b)** Extended viewpoint for *GATA1*. The Copy URL button of the Viewpoint editor tab can be used to visualize viewpoints in UCSC's genome browser. Just as for the embedded view within GOPHER, selected digests are indicated by colored boxes in the background. This representation has the advantage that additional custom tracks can be loaded in order to evaluate viewpoints. In this case the track for alignability is displayed (green). For this track, the alignability is standardized to values greater or equal 0 and lower or equal 1. For example, a value of 1/2 means that the corresponding k-mer occurs 2 times in the genome. The unbalanced digest at the TSS is not selected but can be selected manually or directly during creation of viewpoints, if *Unbalanced margins* is set

Table 2 GOPHER output files: Seven files can be exported from GOPHER via the **Export** menu. These files can be used for downstream analyses, documentation, visualization, and ordering of probes. The file prefix <PREFIX> corresponds to the name of the project. The prefix of the digest file additionally contains the tag for the genome build

Filename	Description
<PREFIX>_DigestedGenome.txt	<i>Content of lines:</i> Coordinates, length, GC and repeat content, indication of selection, number of probes for each digest. <i>Purpose:</i> Can be used for downstream analyses.
<PREFIX>_viewPoints.tsv	<i>Content of lines:</i> Gene symbol, coordinates, UCSC custom URL, number of selected digests, score, length, indication whether digest with TSS is selected. <i>Purpose:</i> Documentation and sharing of results.
<PREFIX>_uniqueTargetDigests.bed	<i>Content of lines:</i> Coordinates and names of targeted digests. Duplicates arising from overlapping viewpoints are removed. <i>Purpose:</i> Probe design with external tools.
<PREFIX>_uniqueTargetDigestMargins.txt	<i>Content of lines:</i> Coordinates and names of margins of targeted digests. <i>Purpose:</i> Probe design with external tools.
<PREFIX>_allTracks.bed	<i>Content of lines:</i> Contains multiple sections for viewpoints, digests, margins, and probes. <i>Purpose:</i> Can be uploaded as multiple custom tracks to UCSC.
<PREFIX>_ProbeFile.bed	<i>Content of lines:</i> Coordinates and names of probes. <i>Purpose:</i> Ordering of probes.
<PREFIX>_agilentProbeFile.txt.zip	<i>Content of lines:</i> Coordinates, names, and sequences of probes. <i>Purpose:</i> Ordering of probes.

viewpoints with scores of 100%). Restriction fragments and target regions (margins of restriction fragments) are depicted in blue. Probes are displayed in different greyscale that reflect the mean k-mer-alignabilities (MKA) of probes (black for a MKA of 1 and gray for higher values). Currently, alignability maps for mm9 and hg19 are available within UCSC but not for mm10 and hg38.

Discussion

In CHC projects published to date, probes have been designed by the online tool CapSequm, by various (undocumented) scripts, and some have been manually designed or revised [6, 10–19, 23, 29]. In our experience, CHC probe design is time consuming and hard to document, and different studies have used different viewpoint and probe definitions without the motivation behind the differences being explicitly stated. For example, for the promoter of *Prrx1* two different viewpoints (Additional file 1: Figure S4) were used in two different studies [6, 29].

The current work presents a formalization of the criteria used to evaluate CHC probes, and provides an easy to use Java desktop application that implements three strategies for CHC probe design. GOPHER's *simple* approach is similar to the probe design strategy taken in previous publications that employ CHC to investigate all promoters, with one probe being placed at each margin of single digests that overlap TSS. This approach is suitable for large numbers of target genes, e.g., the promoterome, and is often used with a six-cutter to obtain relatively large enriched digests. If four-cutters are used, in general many target genes will be discarded because digests are too short or not well centered at the TSS. GOPHER's *simple-patched* approach is an extension of the *simple* approach that intends to improve coverage of viewpoints, which is accomplished by adding adjacent digests in upstream or downstream direction. GOPHER's *extended* approach is intended mainly for focused investigations of smaller gene sets; here, we recommend the use of a four-cutter to obtain higher resolution, which together with GOPHER's default viewpoint size (5000 bp upstream and 1500 bp downstream) will tend to generate viewpoints with 5–15 enriched digests.

GOPHER provides a series of features that allow users to visualize and edit viewpoints, and also outputs a series of files that are useful for a number purposes. For instance, an appropriately formatted probe file can readily be used for ordering, or a file containing the size, GC-content, alignability, and number of probes for individual digests can be used for normalization of interaction counts [31, 32]. Finally, it is also possible to export entire projects to files that can be shared between GOPHER users.

Conclusion

GOPHER allows probes for capture Hi-C viewpoints to be created within a few hours according to one of three different design approaches using clear and consistent rules. The graphical user interface allows post hoc inspection and editing of individual viewpoints. GOPHER will enable a wider range of researchers to employ CHC by providing an easy-to-use Java desktop application for CHC probe design. Source code, precompiled executables and detailed documentation are available on the GOPHER GitHub page at <https://github.com/TheJacksonLaboratory/Gopher>.

Methods

Implementation

GOPHER is a desktop Java application written using the JavaFX library for designing graphical user interfaces. GOPHER requires Java 1.8. The source code of GOPHER can be downloaded from the GOPHER GitHub page:

```
https://github.com/TheJacksonLaboratory/Gopher
```

From here, users can also download a precompiled application. Detailed documentation is available on a *readthedocs* site. GOPHER is freely available for academic use.

Data and data preparation

GOPHER allows the user to download all necessary data directly from the GOPHER application prior to generating probes. GOPHER downloads the genome sequence, transcript annotations, and alignability map for the human (hg19 or hg38) and mouse (mm9 or mm10) genomes. Default values are provided for all relevant parameters, and the user can change parameters directly in the application window (Table 1).

In silico digestion of the target genome

GOPHER downloads genome files from the UCSC Genome Browser Database [30]. It extracts this file (which is downloaded as a g'zipped file) and combines the individual chromosome files into a single file, which it writes to disk (for instance, the file for the human hg38 genome would be named `hg38.fa`). It then uses the HTSJDK library, a Java API for high-throughput sequencing data (HTS) formats that is part of the SAMtools suite [33], to create a FASTA index file. During the creation of viewpoints, GOPHER utilizes functions of HTSJDK together with these files to identify all digests on the basis of the enzyme or enzymes chosen by the user.

Mean k-mer alignability of probes

For mouse mm9 and human hg19 genome builds, we used the CRG Alignability tracks for a

k-mer size of 50 that are available as default tracks in UCSC's genome browser (wgEncodeEM002940 and wgEncodeEH000320). The alignability is standardized to values in the range [0,1]. For instance, 0.25 means that the corresponding k-mer occurs 4 times within the entire genome. These tracks were created in the course of the ENCODE project [34] using the mappability program of the GEM suite (GEM Multitool) [35]. For the mm10 and hg38 genome builds, no alignability tracks were available. Therefore, we generated new tracks using GEM version 1.7.1. In accordance with the existing maps, we allowed at most 2 mismatches (-m 2). The index was built on all chromosomes including non-canonical chromosomes. The output of GEM was converted to bedGraph format using utility programs of the kentUtils tool suite [36]. Lines for non-canonical chromosomes were removed and compressed files were uploaded to ftp://ftp.jax.org/robinp/GOPHER/alignability_maps. GOPHER automatically downloads the file as required for the genome build being analyzed.

Within GOPHER, the alignability map $p \mapsto a_k(p)$ indicates how often the k-mer starting at position p occurs in the entire genome. For instance, $a_{50}(42) = 2$ means that the 50mer starting at position 42 occurs two times in total. GOPHER uses binary search for efficient retrieval of all k-mer alignabilities within a given target region which can be used for fast calculation of mean k-mer alignabilities of digests. For a probe of length l starting at position p , the mean k-mer alignability is calculated as follows:

$$a_k(p) = \frac{1}{l-k+1} \cdot \sum_{i=0}^{l-k} a_k(p+i). \quad (1)$$

By default, probes with a mean k-mer alignability above 2 are rejected, but users can adjust this threshold value.

Mean G/C content of probes

GOPHER uses the HTSJDK library to retrieve the sequences of candidate digests and probes and counts the number of G, C, A, and T bases in order to calculate the GC content. By default, probes with a GC content of between 35 and 65% are accepted, but users can adjust these threshold values.

Score for simple viewpoints

Ideally, the selected digest of a simple viewpoint is centered at the TSS and not too short. In order to capture this intuition, we estimate the average size $\bar{\mu}_D$ from all digests in the genome and use a normal distribution with a mean $\mu = 0$ that corresponds to the TSS and a standard deviation $\sigma = 1/6 \cdot \bar{\mu}_D$. The score is then calculated as the area under the curve of the probability density function within

the range that is covered by the digest containing the TSS. The score can take on values between 0 and 1, and digests that are not well centered at the TSS will be assigned a score close to 0.5.

Score for extended viewpoints

The score for extended viewpoints is very similar to the score for simple viewpoints but it allows asymmetric length ratios for the targeted regions upstream and downstream of the TSS. Given the user-specified maximum allowed upstream and downstream distances from the TSS (Table 1), an extended viewpoint can be seen as a set of selected digests that overlap the specified range $[TSS - up_{max}, TSS + down_{max}]$. For a good viewpoint, the selected digests should cover as many positions as possible within the specified range, especially near the TSS.

To calculate the empirical score, we model the coverage of a viewpoint and its associated upstream and downstream sizes (5000 bp and 1500 bp by default) as based on normal distributions with mean at the transcription start site (or other central genomic position) and standard deviations calculated as 1/6 times the upstream or downstream size. Since 6 standard deviations cover $\sim 99.999\%$ of the total probability, if chosen digests cover all of the upstream space they will add 0.5 to the total probability and similarly with the downstream digests. In general, a digest will contribute more to the total probability the closer it is to the TSS. The probability is multiplied by 100 and the score is reported as a percentage.

Availability and requirements

- **Project name:** GOPHER: Generator Of Probes for capture Hi-C Experiments at high Resolution
- **Project home page:** <https://github.com/TheJacksonLaboratory/Gopher>
- **Operating system:** Platform independent
- **Programming language:** Java
- **Other requirements:** Java 1.8
- **License:** JAX
- **Any restrictions to use by non-academics:** License needed

Additional file

Additional file 1: Supplementary figures for the distribution of mapped Hi-C interaction read pairs within digests, an exploration of alignability thresholds, heterogeneous definitions of viewpoints and explanation of GOPHER's allTracks.bed file. (PDF 1430 kb)

Abbreviations

3C: Chromosome conformation capture; CHC: Capture Hi-C; HTS: High-throughput sequencing; GEM: Genome multitool; GOPHER: Generator of probes for capture hi-c experiments at high resolution; GUI: Graphical user interface; GWAS: Genome-wide association study; MKA: Mean

k-mer-alignability; SNP: Single nucleotide polymorphism; TAD: Topologically associating domain; TSS: Transcription start site

Acknowledgements

The authors thank Nomi Harris for critical comments on the manuscript.

Funding

This work was supported by an Agilent Thought Leader Award to P.N.R. Additional support was provided by a grant from the European Union Seventh Framework Programme (FP7/2007-2013) ("SYBIL" grant No. 602300). P.N.R gratefully acknowledges additional support from the Donald A. Roux family fund.

Availability of data and materials

Source code and executable can be downloaded from the GOPHER GitHub page at <https://github.com/TheJacksonLaboratory/Gopher>. Documentation is available on a readthedocs page at <https://gopher.readthedocs.io/en/latest/>.

Authors' contributions

PH and PNR developed the algorithms and strategies implemented in GOPHER. PH and PNR implemented the Java code with support from RS, HB and DD. PH and PNR wrote the manuscript with support all of the coauthors. JH, DL, SA, SM, and UK provided insights about CHC that were critical in designing the three approaches of GOPHER to probe design. All authors read and approved the final manuscript.

Authors' information

Peter Hansen: peter.hansen@charite.de; Salaheddine Ali: sali@molgen.mpg.de; Hannah Blau: hannah.blau@jax.org; Daniel Danis: daniel.danis@jax.org; Jochen Hecht: jochen.hecht@crg.eu; Uwe Kornak: uwe.kornak@charite.de; Darío G. Lupiáñez: dario.lupianez@mdc-berlin.de; Stefan Mundlos: stefan.mundlos@charite.de; Robin Steinhaus: robin.steinhaus@charite.de; Peter N. Robinson: peter.robinson@jax.org.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Institute of Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany. ²Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany. ³The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, United States. ⁴Genomics Unit, Centre for Genomic Regulation, Carrer del Dr. Aiguader 88, Barcelona 08003, Spain. ⁵Berlin Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, Berlin 13353, Germany. ⁶Epigenetics and Sex Development Group, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin-Buch 13125, Germany. ⁷Institute for Systems Genomics, University of Connecticut, Farmington, CT 06032, United States.

Received: 24 September 2018 Accepted: 16 December 2018

Published online: 14 January 2019

References

- Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science (New York, N.Y.)* 2002;295:1306–11. <https://doi.org/10.1126/science.1067799>.
- Denker A, de Laat W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* 2016;30:1357–82. <https://www.ncbi.nlm.nih.gov/pubmed/27340173>.
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012;485:376–80. <https://doi.org/10.1038/nature11082>. 1206.5533.
- Lajoie BR, Dekker J, Kaplan N. The hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods (San Diego, Calif.)* 2015;72:65–75. <https://doi.org/10.1016/j.ymeth.2014.10.031>.
- Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol.* 2016;17:743–55. <https://doi.org/10.1038/nrm.2016.104>.
- Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, De Gobbi M, Taylor S, Gibbons R, Higgs DR. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet.* 2014;46(2):205–12. <https://doi.org/10.1038/ng.2871>.
- Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett S, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, Osborne CS. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet.* 2015;47(6):598–606. <https://doi.org/10.1038/ng.3286>.
- Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, Nagano T, Katsman Y, Sakthidevi M, Wingett S, 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. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* 2015;25(4):582–97. <https://doi.org/10.1101/gr.185272.114>. 1311.1716.
- Schoenfelder S, Javierre B-M, Furlan-Magaril M, Wingett S, Fraser P. Promoter capture hi-c: High-resolution, genome-wide profiling of promoter interactions. *J Visualized Exp JoVE.* 2018. <https://doi.org/10.3791/57320>.
- Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett S, Várnai C, Thiecke MJ, Burden F, Farrow S, Cutler AJ, Rehnström K, Downes K, Grassi L, Kostadima M, Freire-Pritchett P, Wang F, Consortium B, Stunnenberg HG, Todd JA, Zerbino DR, Stegle O, Ouwehand WH, Frontini M, Wallace C, Spivakov M, Fraser P. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell.* 2016;167(5):1369–1384. <https://doi.org/10.1016/j.cell.2016.09.037>.
- McGovern A, Schoenfelder S, Martin P, Massey J, Duffus K, Plant D, Yarwood A, Pratt AG, Anderson AE, Isaacs JD, Diboll J, Thalayasingam N, Ospelt C, Barton A, Worthington J, Fraser P, Eyre S, Orozco G. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* 2016;17:212. <https://doi.org/10.1186/s13059-016-1078-x>.
- Dryden N, Broome L, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa I, Assiotis I, Fenwick K, Maguire S, Campbell J, Natrajan R, Lambros M, Perrakis E, Ashworth A, Fraser P, Fletcher O. Unbiased analysis of potential targets of breast cancer susceptibility loci by capture Hi-C. *Genome Res.* 2014;24:1854–68. <https://doi.org/10.1101/gr.175034.114>.
- Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, Whiffin N, Carnicer MJ, Broome L, Dryden N, Nagano T, Schoenfelder S, Enge M, Yuan Y, Taipale J, Fraser P, Fletcher O, Houlston RS. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun.* 2015;6:6178. <https://doi.org/10.1038/ncomms7178>.
- Du M, Tillmans L, Gao J, Gao P, Yuan T, Dittmar RL, Song W, Yang Y, Sahr N, Wang T, Wei G-H, Thibodeau SN, Wang L. Chromatin interactions and candidate genes at ten prostate cancer risk loci. *Sci Rep.* 2016;6:23202. <https://doi.org/10.1038/srep23202>.
- Martin P, McGovern A, Massey J, Schoenfelder S, Duffus K, Yarwood A, Barton A, Worthington J, Fraser P, Eyre S, Orozco G. Identifying causal genes at the multiple sclerosis associated region 6q23 using capture Hi-C. *PLoS ONE.* 2016;11:0166923. <https://doi.org/10.1371/journal.pone.0166923>.
- Pan DZ, Garske KM, Alvarez M, Bhagat YV, Boock J, Nikkola E, Miao Z, Raulerson CK, Cantor RM, Civelek M, Glastonbury CA, Small KS,

- Boehnke M, Lusis AJ, Sinsheimer JS, Mohlke KL, Laakso M, Pajukanta P, Ko A. Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS. *Nat Commun*. 2018;9:1512. <https://doi.org/10.1038/s41467-018-03554-9>.
17. Choy M-K, Javierre BM, Williams SG, Baross SL, Liu Y, Wingett S, Akbarov A, Wallace C, Freire-Pritchett P, Rugg-Gunn PJ, Spivakov M, Fraser P, Keavney BD. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. *Nat Commun*. 2018;9:2526. <https://doi.org/10.1038/s41467-018-04931-0>.
 18. Baxter JS, Leavy OC, Dryden N, Maguire S, Johnson N, Fedele V, Simigdala N, Martin L-A, Andrews S, Wingett S, Assiotis I, Fenwick K, Chauhan R, Rust AG, Orr N, Dudbridge F, Haider S, Fletcher O. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat Commun*. 2018;9:1028. <https://doi.org/10.1038/s41467-018-03411-9>.
 19. Montefiori LE, Sobreira DR, Sakabe NJ, Aneas I, Joslin AC, Hansen GT, Bozek G, Moskowitz IP, McNally EM, Nóbrega MA. A promoter interaction map for cardiovascular disease genetics. *eLife*. 2018;7. <https://doi.org/10.7554/eLife.35788>.
 20. Siersbæk R, Madsen JGS, Javierre BM, Nielsen R, Bagge EK, Cairns J, Wingett S, Traynor S, Spivakov M, Fraser P, Mandrup S. Dynamic rewiring of promoter-anchored chromatin loops during adipocyte differentiation. *Mol Cell*. 2017;66:420–4355. <https://doi.org/10.1016/j.molcel.2017.04.010>.
 21. Novo CL, Javierre B-M, Cairns J, Segonds-Pichon A, Wingett S, Freire-Pritchett P, Furlan-Magaril M, Schoenfelder S, Fraser P, Rugg-Gunn PJ. Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition. *Cell Rep*. 2018;22:2615–27. <https://doi.org/10.1016/j.celrep.2018.02.040>.
 22. Burren OS, Rubio García A, Javierre B-M, Rainbow DB, Cairns J, Cooper NJ, Lambourne JJ, Schofield E, Castro Dopico X, Ferreira RC, Coulson R, Burden F, Rowston SP, Downes K, Wingett S, Frontini M, Ouwehand WH, Fraser P, Spivakov M, Todd JA, Wicker LS, Cutler AJ, Wallace C. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol*. 2017;18:165. <https://doi.org/10.1186/s13059-017-1285-0>.
 23. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K, Kempfer R, Jerković I, Chan W-L, Spielmann M, Timmermann B, Wittler L, Kurth I, Cambiaso P, Zuffardi O, Houge G, Lambie L, Brancati F, Pombo A, Vingron M, Spitz F, Mundlos S. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*. 2016;538:265–9. <https://doi.org/10.1038/nature19800>.
 24. Davies JOJ, Telenius JM, McGowan SJ, Roberts NA, Taylor S, Higgs DR, Hughes JR. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods*. 2015;13(1):74–80. <https://doi.org/10.1038/nmeth.3664>.
 25. Anil A, Spalinskas R, Åkerborg O, Sahlén P, Berger B. Hicaptools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications. *Bioinformatics* (Oxford, England). 2018;34:675–7. <https://doi.org/10.1093/bioinformatics/btx625>.
 26. Andrey G, Schöpflin R, Jerković I, Heinrich V, Ibrahim DM, Paliou C, Hochradel M, Timmermann B, Haas S, Vingron M, Mundlos S. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Res*. 2016;213066–116. <https://doi.org/10.1101/gr.213066.116>.
 27. Schoenfelder S, Sugar R, Dimond A, Javierre B-M, Armstrong H, Mifsud B, Dimitrova E, Matheson L, Tavares-Cadete F, Furlan-Magaril M, Segonds-Pichon A, Jurkowski W, Wingett S, Tabbada K, Andrews S, Herman B, LeProust E, Osborne CS, Koseki H, Fraser P, Luscombe NM, Elderkin S. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet*. 2015;47(10):1179–86. <https://doi.org/10.1038/ng.3393>.
 28. McGowan SJ, Hughes JR, Han ZP, Taylor S. MIG: Multi-Image Genome viewer. *Bioinformatics*. 2013;29(19):2477–8. <https://doi.org/10.1093/bioinformatics/btt406>.
 29. Andrey G, Schöpflin R, Jerković I, Heinrich V, Ibrahim DM, Paliou C, Hochradel M, Timmermann B, Haas S, Vingron M, Mundlos S. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Res*. 2017;27(2):223–33. <https://doi.org/10.1101/gr.213066.116>.
 30. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, Hinrichs AS, Haussler M, Guruvadoo L, Navarro Gonzalez J, Gibson D, Fiddes IT, Eisenhart C, Diekhans M, Clawson H, Barber G, Armstrong J, Haussler D, Kuhn RM, Kent WJ. The UCSC genome browser database: 2018 update. *Nucleic Acids Res*. 2018;46:762–9. <https://doi.org/10.1093/nar/gkx1020>.
 31. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* (Oxford, England). 2012;28:3131–3. <https://doi.org/10.1093/bioinformatics/bts570>.
 32. Geeven G, Teunissen H, de Laat W, de Wit E. peakC: a flexible, non-parametric peak calling package for 4C and capture-C data. *Nucleic Acids Res*. 2018. <https://doi.org/10.1093/nar/gky443>.
 33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPD. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (Oxford, England). 2009;25:2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
 34. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74. <https://doi.org/10.1038/nature11247>.
 35. Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, Ribeca P. Fast computation and applications of genome mappability. *PLoS ONE*. 2012;7(1). <https://doi.org/10.1371/journal.pone.0030377>.
 36. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* (Oxford, England). 2010;26(17):2204–7. <https://doi.org/10.1093/bioinformatics/btq351>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

