Supplementary material

0.1 Triplex-Inspector

Triplex-Inspector is a software pipeline that enables the screening of genomic sequences for sites that can be targeted by triplex-forming molecules. It generates dynamic HTML reports that allow efficient data visualisation and filtering based on Jquery Datatables[1] and JSON objects, which are readily supported by many modern web-browsers.

Triplex-Inspector has been incorporated into the Triplexator toolkit (Buske *et al.*, 2012). The workflow of Triplex-Inspector, outlined in Fig. 1, relies on third party software for several, in some cases optional, tasks.

- Unix shell - pipelining
- Python[2] v 2.7 & Biopython[3] — Cock *et al.* (2009) — sequence conversion and processing
- BEDtools[4] v2.x — Quinlan and Hall (2010) — data intersection
- Circos[5] v0.62 (optional) — Krzywinski *et al.* (2009) — circos graph generation
- SAMtools[6] v0.1.15+ & Pysam[7] (optional) — Li *et al.* (2009) — chromatin integration for bam formatted data
- bx-python[8] (optional) - chromatin integration for bigwig formatted data

The software comes with a suite of additional scripts for the generation of chromatin data from DNase I hypersensitivity deep sequencing data (bam files) that perform tag manipulation (duplicate removal, tag truncation to 5' nucleotide), replica pooling, and tag density enrichment calls, with the latter being based on F-Seq (Boyle *et al.*, 2008).

0.2 Comparison with existing tools

Triplex-Inspector makes it convenient to identify the *optimal* triplex-forming oligonucleotide target in a given gene. Unlike existing tools, e.g. TTS mapping (Jenjaroenpun and Kuznetsov, 2009) and TFO target sequence search (Gaddis *et al.*, 2006), which leave it to the user to investigate potential off-targets on a candidate-by-candidate basis, Triplex-Inspector automatically reports all potential off-targets for each candidate target in a unified framework, allowing the user to quickly find the best target site amongst all candidates simultaneously. In more detail, the three fundamental differences between Triplex-inspector and existing tools are as follows:

Triplex-Inspector can be applied to any genomic sequence. In contrast to the existing tools (Gaddis *et al.*, 2006; Jenjaroenpun and Kuznetsov, 2009), which are limited to human (hg18) and mouse (mm8) genomes, Triplex-Inspector can be used to identify optimal TFO targets for any species for which a genome sequence is available. Target sites can be put into their genomic context by leveraging gene annotations. Files in the format required by Triplex-Inspector are available for many genomes, for example from Ensembl at `http://ensembl.org/info/data/ftp/index.html`.

[1] `http://datatables.net/`
[2] `http://www.python.org/`
[3] `http://biopython.org/wiki/Biopython`
[4] `http://code.google.com/p/bedtools/`
[5] `http://circos.ca/`
[6] `http://samtools.sourceforge.net/`
[7] `http://code.google.com/p/pysam/`
[8] `https://bitbucket.org/james_taylor/bx-python/`

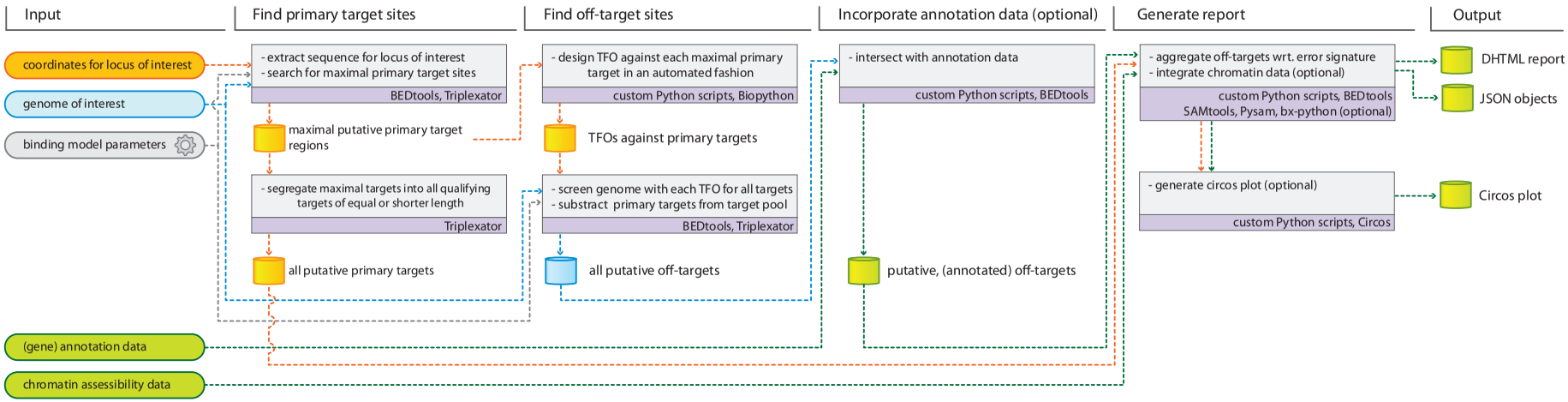

Figure 1: The Triplex-Inspector pipeline.

Triplex-Inspector detects and evaluates *all* off-target sites. Triplex-Inspector uses Triplexator, an exhaustive sequence search engine that is designed to detect all off-target occurrences. By contrast, existing web services (Gaddis *et al.*, 2006; Jenjaroenpun and Kuznetsov, 2009) rely on a heuristic method, BLAST (Altschul *et al.*, 1997), to detect off-target sites. Heuristic methods cannot guarantee to detect all off-targets and may miss eligible sites, especially for short sequences such as targets of PNAs.

In addition, by contrast to web-services, Triplex-Inspector is provided as a compilable package for download, which enables Triplex-Inspector to be linked into computational pipelines without being dependent on the availability of maintenance from third party service providers.

0.3 DNase I hypersensitivity data

In the SupFG case study, we utilised DNase I hypersensitivity data generated from mouse liver cells in the Stamatoyannopoulous lab at the University of Washington as part of the ENCODE project (John *et al.*, 2011).

Putative triplex target sites in *c-Myc*, *Bcl-2* and *CCR5*

We employ Triplex-Inspector on three concrete biologically relevant use-cases that illustrate the power of computer-guided target site evaluation. In the first case, we find a target site in the *c-Myc* locus that is less likely to suffer from G4-quadruplex competition compared to targets used in previous studies. In the second case we focus only on the exons of the 200kb long locus of the *bcl-2* gene and find a target sites without interrupting pyrimidines, which can be expected to form stronger triplexes compared to sites used in previous studies. Finally, we apply Triplex-Inspector to the *CCR5* locus and report targets that have two orders of magnitude fewer sequence copies (equivalent off-targets) compared to previously used target sites.

0.4 *c-Myc*

The gene *c-Myc* encodes a transcription factor that has been implicated as a human oncogene. *c-Myc* regulates many genes with functions in cell-proliferation and has been found to be constitutively expressed in many cancers [see Dang (2012) for a review].

Several sites within the *c-Myc* locus have been targeted with a variety of TFOs of different length (Cooney *et al.*, 1988; Carbone *et al.*, 2004; Belotserkovskii *et al.*, 2007). Several approaches target the promoter P2 located in the first exon of *c-Myc* and report modulated transcription of the downstream gene.

After running Triplex-Inspector on the *c-Myc* locus (parameter setting: minimum target size of 17 nt, minimum guanine rate of 45%, maximal 10% pyrimidine-interruptions in the target) we find the previously used target site (Belotserkovskii *et al.* (2007), hg19, chr8:128748411-128748437, `GGGAAAAAGAAcGGAGGGAGGGA`) among the top-ranking sites. This site has no other copies in the human genome but contains a pyrimidine interruption in the centre, which could weaken triplex formation. It also contains three `GpGpG` sites, which can complicate triplex-mediated targeting utilising purine TFOs, as competition with G4-quadruplexes are likely to occur.

We find an alternative 22 nt long target site located in the first intron (hg19, chr8:128749058-128749097, `GAGAGGAGAAGGcAGAGGGAA`) with only one `GpGpG` triplet, thus reducing the expected effect of G4-quadruplex com-

petition. Similar to the above target site, this target site has no additional sequence copies across the human genome (see Supplementary Fig. 2), and also contains a pyrimidine interruption, which is however slightly off-centre and may therefore allow more stable triplex formation. In either case it is expected that the remaining nucleotides, which can engage in hydrogen bonding to stabilise triplex formation, are able to compensate for this interruption.

0.5 B-cell lymphoma 2 (*Bcl-2*)

Bcl-2 is a proto-oncogene with an important role in programmed cell death (apoptosis). Disruption or damage to the *bcl-2* gene has been shown to cause a number of cancers [see Davids and Letai (2012) for review]. In human, the *bcl-2* locus covers almost 200 kb on chromosome 18, most of which is comprised by the second intron (RefSeq ID: NM_000633). Previous studies focused on the 3' UTR (hg19, chr18:60794609-60794627, target site `AGAGGGAAGGAAcAGAGG`), and the promoter (hg19, chr18:60986996-60987015, target site `GAGGGGtGGGGAGAAGGAGG`), when targeting *bcl-2* by means of TFOs (Shen *et al.*, 2001, 2003). We similarly limit Triplex-Inspector to investigate only exons with 100 bp of flanking sequence up and downstream. Our search reveals a superior 24 nt long putative target site just upstream of the last exon (hg19, chr18:60796000-6079603, `GGAGAGGAGGGAAGAAAAAGAAAG`) that has *no copies* across the human genome. Moreover, this site contains *no pyrimidine interruption* and only one G-triplet (`GpGpG`), which makes this site particularly suitable for triplex-mediated targeting due to minimal competition with G4-quadruplex formation of the TFO. Importantly, any off-target that overlaps an annotated gene or exon (Gencode v14) is either at least 5 nt and 7 nt shorter, respectively, or it contains a mismatch. Thus the primary target site we identify using Triplex-Inspector can be expected to provide superior affinity over any off-target.

0.6 C-C chemokine receptor type 5 (*CCR5*)

CCR5 encodes a G-protein-coupled transmembrane receptor on the surface of white blood cells that is required for HIV-1 entry and therefore constitutes an attractive drug target [see Lai (2012) for a review].

In the past, a 12 nt site in the *CCR5* locus (`GAGAAGAAGAGG`) and its 2 nt shorter subsequence (`GAGAAGAAGA`) has been targeted with TFOs (Belousov *et al.*, 1998) and PNAs (Schleifman *et al.*, 2011), respectively. These sites have 2,068 and 23,354 identical copies in the human genome (hg19), several hundred of which overlap exons. These sites are likely to be bound by the TFO or PNA with the same efficiency as the intended target site.

Triplex-Inspector identifies a potential 18 nt long target site that is located in intron 1 of *CCR5* and has only 20 indistinguishable off-targets (genomic copies), of which only 7 are located in genes and only 2 overlap an exon. While this target site contains a mismatch, it is otherwise comparable to the above-mentioned sites with regard to guanine content and number of GpGpG triads. Alternatively, a 17 nt target site (hg19, chr3:46417280-46417297, `AAAGGGAGAGAGAGAGG`) downstream of the previously chosen site could be used if no mismatches are a prerequisite. This target has 61 genomic copies, of which 23 are in genes, and none of which overlap any annotated exon (Gencode v14).

References

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.

Belotserkovskii, B. P., Silva, E. D., Tornaletti, S., Wang, G., Vasquez, K. M., and Hanawalt, P. C. (2007). A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J Biol Chem*, **282**(44), 32433–32441.

Belousov, E. S., Afonina, I. A., Kutyavin, I. V., Gall, A. A., Reed, M. W., Gamper, H. B., Wydro, R. M., and Meyer, R. B. (1998). Triplex targeting of a native gene in permeabilized intact cells: covalent modification of the gene for the chemokine receptor ccr5. *Nucleic Acids Res*, **26**(5), 1324–1328.

Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S. (2008). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**(21), 2537–2538.

Buske, F. A., Bauer, D. C., Mattick, J. S., and Bailey, T. L. (2012). Triplexator: Detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res*, **22**(7), 1372–1381.

Carbone, G. M., McGuffie, E., Napoli, S., Flanagan, C. E., Dembech, C., Negri, U., Arcamone, F., Capobianco, M. L., and Catapano, C. V. (2004). DNA binding and antigene activity of a daunomycin-conjugated triplex-forming oligonucleotide targeting the P2 promoter of the human c-myc gene. *Nucleic Acids Res*, **32**(8), 2396–2410.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Cooney, M., Czernuszewicz, G., Postel, E. H., Flint, S. J., and Hogan, M. E. (1988). Site-specific oligonucleotide binding represses transcription of the human c-myc gene in vitro. *Science*, **241**(4864), 456–459.

Dang, C. V. (2012). Myc on the path to cancer. *Cell*, **149**(1), 22–35.

Davids, M. S. and Letai, A. (2012). Targeting the b-cell lymphoma/leukemia 2 family in cancer. *J Clin Oncol*, **30**(25), 3127–3135.

Gaddis, S. S., Wu, Q., Thames, H. D., DiGiovanni, J., Walborg, E. F., MacLeod, M. C., and Vasquez, K. M. (2006). A web-based search engine for triplex-forming oligonucleotide target sequences. *Oligonucleotides*, **16**(2), 196–201.

Jenjaroenpun, P. and Kuznetsov, V. (2009). TTS Mapping: integrative WEB tool for analysis of triplex formation target DNA Sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC Genomics*, **10**(Suppl 3), S9.

John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., Hager, G. L., and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*, **43**(3), 264–268.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res*, **19**(9), 1639–1645.

Lai, Y. (2012). Ccr5-targeted hematopoietic stem cell gene approaches for hiv disease: current progress and future prospects. *Curr Stem Cell Res Ther*, **7**(4), 310–317.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and , . G. P. D. P. S. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16), 2078–2079.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

Schleifman, E. B., Bindra, R., Leif, J., del Campo, J., Rogers, F. A., Uchil, P., Kutsch, O., Shultz, L. D., Kumar, P., Greiner, D. L., and Glazer, P. M. (2011). Targeted disruption of the ccr5 gene in human hematopoietic stem cells stimulated by peptide nucleic acids. *Chem Biol*, **18**(9), 1189–1198.

Shen, C., Buck, A., Mehrke, G., Polat, B., Gross, H., Bachem, M., and Reske, S. (2001). Triplex forming oligonucleotide targeted to 3'utr downregulates the expression of the bcl-2 proto-oncogene in hela cells. *Nucleic Acids Res*, **29**(3), 622–628.

Shen, C., Rattat, D., Buck, A., Mehrke, G., Polat, B., Ribbert, H., Schirrmeister, H., Mahren, B., Matuschek, C., and Reske, S. N. (2003). Targeting bcl-2 by triplex-forming oligonucleotide–a promising carrier for gene-radiotherapy. *Cancer Biother Radiopharm*, **18**(1), 17–26.

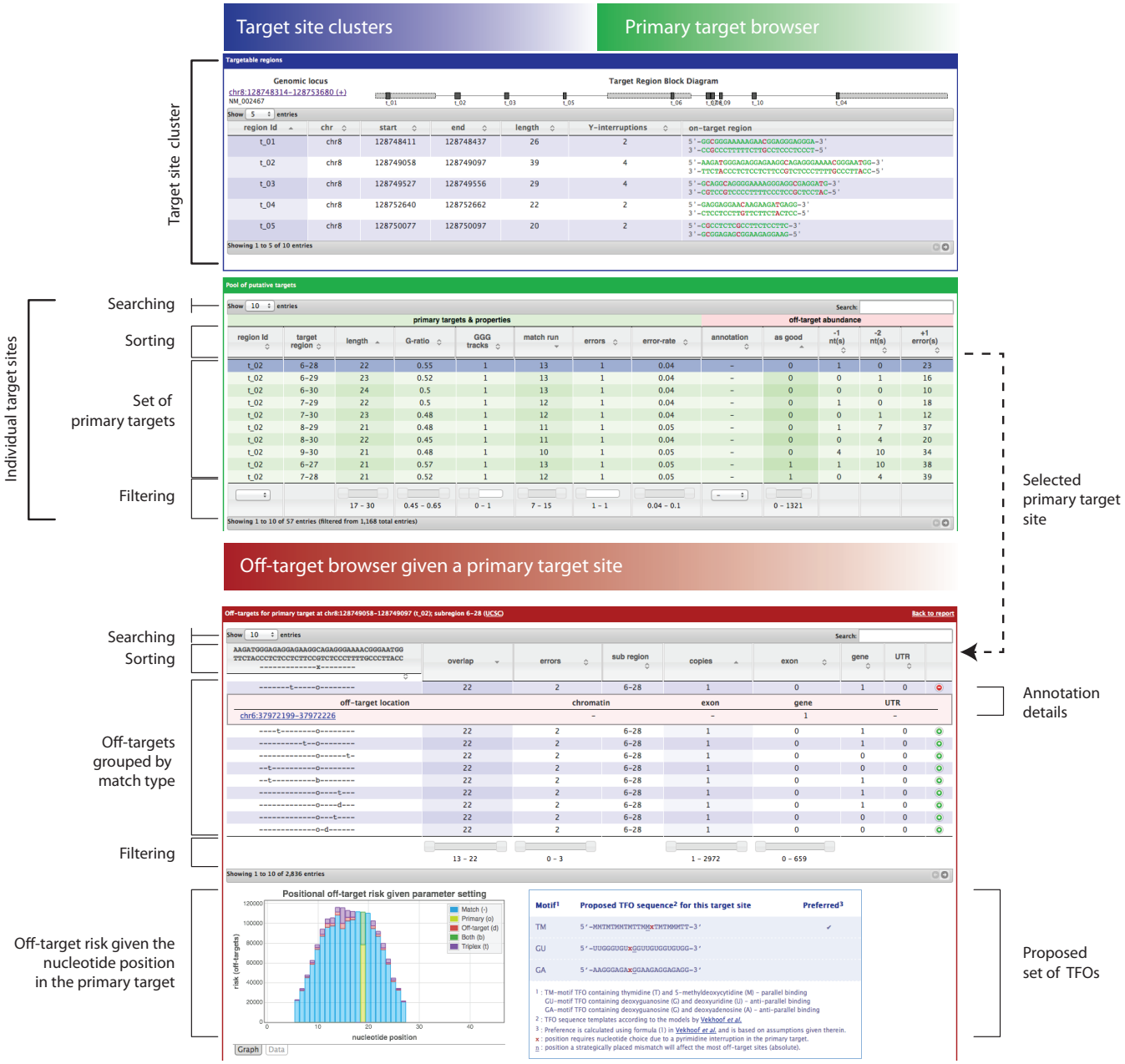

Figure 2: **Report of Triplex-Inspector for the *c-Myc* locus.**