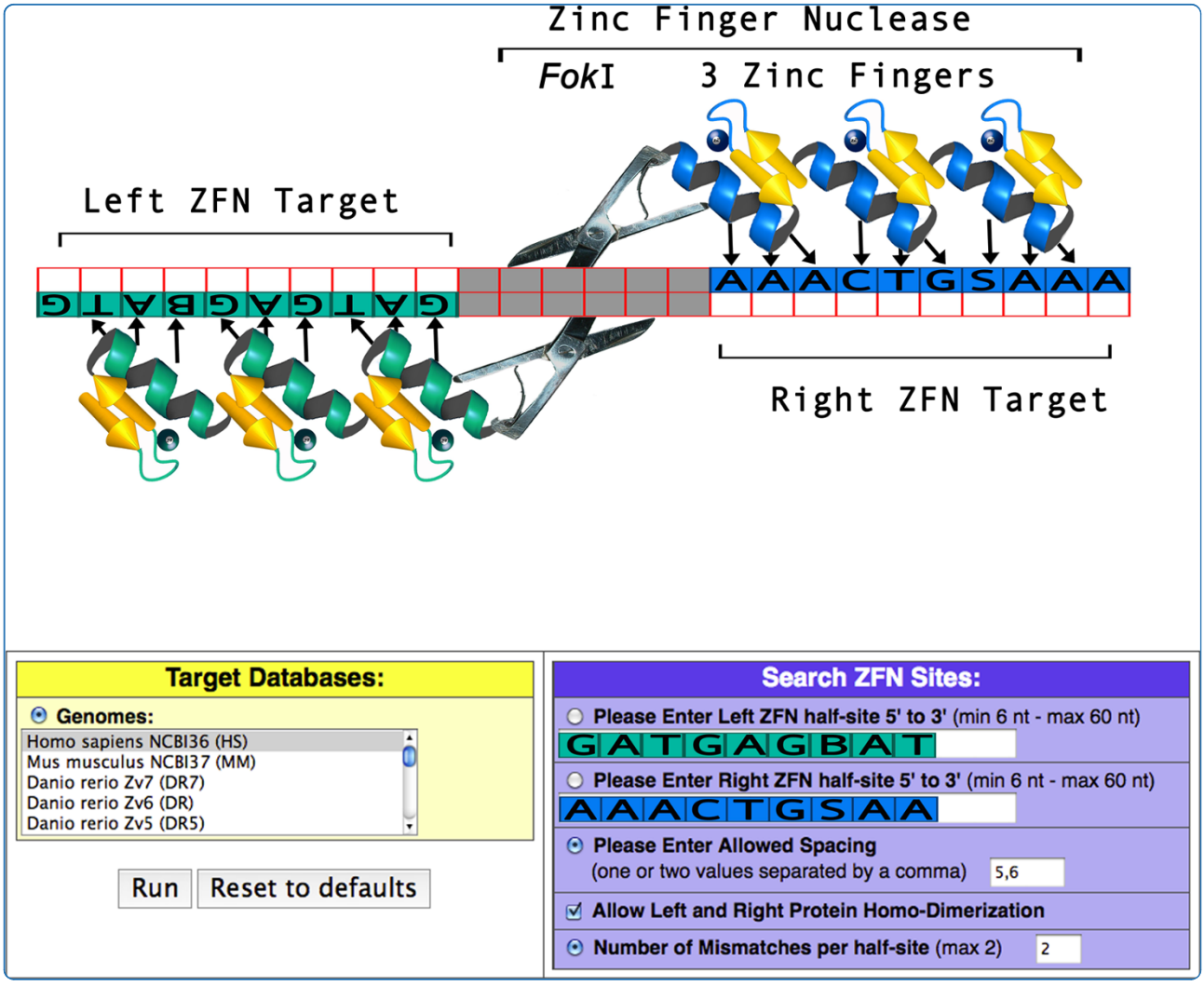
BMC Bioinformatics



ZFN-Site searches genomes for zinc finger nuclease target sites and off-target sites

Cradick *et al.*

Cradick et al. BMC Bioinformatics 2011, 12:152
http://www.biomedcentral.com/1471-2105/12/152

BMC Bioinformatics

SOFTWARE Open Access

ZFN-Site searches genomes for zinc finger nuclease target sites and off-target sites

Thomas J Cradick[1*†], Giovanna Ambrosini[2,3†], Christian Iseli[2,4], Philipp Bucher[2,3] and Anton P McCaffrey[1]

Abstract

Background: Zinc Finger Nucleases (ZFNs) are man-made restriction enzymes useful for manipulating genomes by cleaving target DNA sequences. ZFNs allow therapeutic gene correction or creation of genetically modified model organisms. ZFN specificity is not absolute; therefore, it is essential to select ZFN target sites without similar genomic off-target sites. It is important to assay for off-target cleavage events at sites similar to the target sequence.

Results: ZFN-Site is a web interface that searches multiple genomes for ZFN off-target sites. Queries can be based on the target sequence or can be expanded using degenerate specificity to account for known ZFN binding preferences. ZFN off-target sites are outputted with links to genome browsers, facilitating off-target cleavage site screening. We verified ZFN-Site using previously published ZFN half-sites and located their target sites and their previously described off-target sites. While we have tailored this tool to ZFNs, ZFN-Site can also be used to find potential off-target sites for other nucleases, such as TALE nucleases.

Conclusions: ZFN-Site facilitates genome searches for possible ZFN cleavage sites based on user-defined stringency limits. ZFN-Site is an improvement over other methods because the FetchGWI search engine uses an indexed search of genome sequences for all ZFN target sites and possible off-target sites matching the half-sites and stringency limits. Therefore, ZFN-Site does not miss potential off-target sites.

Background

The ability to create double-stranded DNA breaks at specific genomic sequences is important for gene correction therapeutics, targeted gene integration and gene modification for research models as well as gene disruption [1]. Zinc Finger Nucleases (ZFNs) are promising candidates for such specific nucleases. ZFNs consist of the sequence-independent FokI nuclease domain fused to zinc finger proteins (ZFPs). ZFPs can be altered to change their sequence specificity. Cleavage of targeted DNA requires binding of two ZFNs (designated left and right) to adjacent half-sites on opposite strands with correct orientation and spacing, thus forming a FokI dimer [2]. The requirement for dimerization increases ZFN specificity significantly. Three or four finger ZFPs target ~9 or 12 bases per ZFN, or ~18 or 24 bases for the ZFN pair. ZFN pairs have been used for gene targeting at specific genomic loci in insect, plant, animal and human cells [3-10] (and reviewed in [11,12]). Methods are available to measure general ZFN toxicity or the amount of unrepaired DNA ends resulting from ZFN treatment [13-16]; however, determining all possible off-target cleavage sites may be challenging, as some possible cleavage sites can be missed by BLAST and similar methods. ZFN-Site determines the most probable off-target sites for further analysis or testing. Several ZFN design web tools exist that offer BLAST-based searches for potential ZFN off-target sites [17-22]. BLAST searches, which implement a local alignment search, are not optimal for finding ZFN off-target sites and may miss some sites because they utilize seed-based methods with a non-overlapping word index to search only for perfect matches, rather than longer imperfect matches. BLAST also uses an E-value threshold that does not directly correspond to a "# of mismatches" threshold. ZFN-Site is more thorough because it scans one index entry for each nucleotide in the genome, ensuring that

* Correspondence: tj@alum.mit.edu
† Contributed equally
[1]University of Iowa School of Medicine, Department of Internal Medicine, Iowa City, Iowa, 52245, USA
Full list of author information is available at the end of the article

no matches are missed. ZFN-Site was created to provide a simple, easy-to-use interface that does not require the end user to possess specialized bioinformatics or search algorithm expertise. ZFN-Site provides an interface that searches multiple genomes for sites with ambiguities, mismatches, multiple spacings, hetero-dimeric binding sites and homo-dimeric binding sites composed of two left or two right ZFN half-sites. Changing these parameters can expand the number of possible off-target sites returned to match the purpose. A larger list enables thorough screening for potential ZFN off-target sites using new methods, such as high-throughput sequencing or mutation screens.

Implementation

ZFN-Site was developed to quickly locate all possible ZFN target and off-target sites that might be cleaved. Based on the tailoring of search parameters, ZFN-Site generates sets of search strings. To ensure that all sites matching these criteria are found in the requested genomes, ZFN-Site employs the FetchGWI search engine [23]. The input can be either the nucleotide sequence of the intended target site of each ZFN (basic search) or information about each ZFN's binding specificity (relaxed specificity). The number of possible sites is expanded by choice of ZFN spacing, the possibility of ZFN homo-dimerization (see below) and the number of allowed mismatches. The output from ZFN-Site aids in the choice of ZFN pairs that minimize potential off-target sites and allows experimental testing of each ZFN pairs' off-target sites in cells or in mutated animals. Experimentally testing the list of found sites under a series of different conditions may determine the conditions favoring more specific targeting and less off-target cleavage events.

Basic Target Search

The simplest search method uses the intended target site to scan whole genomes. This type of search is valuable when choosing prospective target sites or when there is no available ZFP mismatch specificity data. ZFN-Site allows searches for off-target sites containing up to two mismatches per half-site. ZFN-Site outputs all target and off-target sites matching the selection criteria.

The genome or genomes to be searched are chosen by clicking on the species list on the left side of the ZFN-Site web page. Scrolling down reveals the full list. Use command-click (mac) or control-click (pc) to choose multiple genomes to be searched simultaneously. A click on ALL searches the entire list of genomes shown in Table 1.

Half-sites are entered without spaces, 5' to 3', as they occur on the opposite strand of a ZFN target. The following sequence is an example of the top DNA strand of a three finger ZFN pair target site: 5'-**CGGAGC-CGCTTT**aacccACTCTGTGGAAG-3'[3]. The right ZFN half-site is underlined and should be entered into the program 5'-3' as ACTCTGTGGAAG. The left ZFN half-site is the reverse complement of the bold sequence and should be entered 5'-3' as AAAGCGGCTCCG (Figure 1).

Table 1 List of Genomes Scanned by ZFN-Site

Genome Release (Code)	Species
Homo sapiens (HS)	Human
Mus musculus (MM)	Mouse
Danio rerio Zv6 (DR)	Zebrafish
Danio rerio Zv5 (DR5)	Zebrafish
Drosophila melanogaster (DM)	Fruit Fly
Apis mellifera (AME)	Bee
Bos taurus (BT)	Cow
Caenorhabditis elegans NCBIWS170(CE)	Nematode
Canis familiaris (CFA)	Dog
Pan troglodytes (PTR)	Chimpanzee
Rattus norvegicus (RN)	Rat
Saccharomyces cerevisiae (SCE)	Yeast
Tribolium castaneum (TCA)	Beetle
All genomes (ALL)	All of the above

The sequence of the DNA spacer between ZFN half-sites (lower case, above) does not greatly influence ZFN specificity, but the length of the spacer between half-sites influences how well a site is cleaved [24]. The allowed number of spacer nucleotides depends on the ZFP-to-FokI linker and is usually five or six nucleotides, although ZFNs with altered linkers have different nucleotide length preferences [25,26]. Genome searches can be run on ZFN-Site with one allowed spacing between half-sites or two spacings if entered separated by a comma (*e.g.*, 5,6). Searches can be repeated using alternate spacings if searching with more than two spacings is required.

In addition to a left ZFN and a right ZFN binding as hetero-dimers, two left or two right ZFNs can bind correctly spaced sites to form homo-dimers and cleave off-target sites [16]. If the "Allow Left and Right Protein Homo-dimerization" box is checked, ZFN-Site also searches for homo-dimeric sites. Use of modified FokI domains may prevent cleavage at most homo-dimeric sites [13,27]. However, identification of homo-dimeric sites and experimental testing for cleavage at each site on these output lists may be necessary to quantitate low levels of cleavage and generate further predictive rules for off-target cleavage events. The specificity of nuclease variants can be experimentally tested using cleavage analysis on the sites comprising the lists of possible off-target sites generated by ZFN-Site [13,25,27-29].

ZFN-Site expands the query targets into a list of queries (or tags) based on the half-sites and inputs. Using increased ambiguities broadens the search.

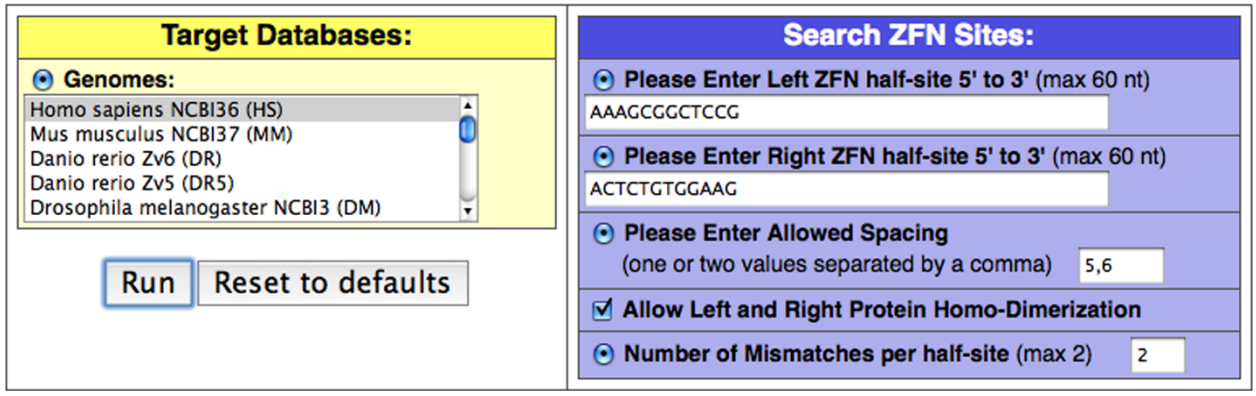


Figure 1 ZFN-Site genome scan using Basic Target Search. ZFN-Site search for Sequence 1 using the half-sites described in the text, which are the ZFN target sites found in IL2R-γ [1]. The inputs are set to search the human genome allowing five and six base pair spacing, two mismatches and homo and hetero-dimerization of the half-sites.

Degenerate nucleotides (specified by standard IUPAC codes) are allowed in the half-site queries because they are then expanded into all possible matching tags. These queries are submitted to an exact search algorithm (described in [23]). The number of such queries increases with the required mismatches and ambiguities (such as Ns and nucleotide IUPAC codes), thus increasing RAM and search time required. Very complex searches may be achieved by breaking the search into parts to speed processing and prevent stalling.

The number of mismatches per half-site (0, 1 or 2) is inputted into the last box. Use 0 to scan only for sites exactly matching the half-sites. This mode is useful for verifying the location of target sites in one or more genomes. The number of off-target sites returned can be greatly increased by allowing 1 or 2 mismatches per half-site. The use of ambiguous nucleotides in the half-sites does not count as a mismatch, and both can be used if needed. Mismatches are allowed in degenerate positions as well. If the user specifies a search with one or two mismatches, ZFN-Site will generate all possible sequence tags that match the target up to the specified number of mismatches.

Once the information above is entered, clicking run will display the query sequences on the next web page, while the genome searches are performed using the FetchGWI program (see paragraph on FetchGWI below). ZFN-Site outputs a list of half-site matches sorted by genome position. This list is scanned by a second program that extracts all combinations on each DNA strand that have the required spacing. For fast performance on the Web, we have limited the number of possible mismatches per ZFN half-site to two. The total number of degenerate nucleotides is also limited to two, such that the computational complexity is manageable.

Based on these inputs, ZFN-Site generates a list of genomic sequences that are exact or near-exact matches to the input query set, along with chromosomal coordinates (including NCBI chromosomal accession number and the start and end positions within the chromosome), DNA strand and HTML links to their exact location on ENSEMBL, UCSC and NCBI browsers [23] (Figure 2). Results are output under "WORD MATCHES" in a two-line format for each genomic sequence returned. The top line of each pair of lines depicts the genomic sequence. The lower line displays the differences from the query sequence. Spacer nucleotides are indicated in blue, and in cases where there are ambiguous nucleotides, genomic nucleotides matching an unambiguous portion of the query sequence are in blue. The number of nucleotides in the spacer is indicated by the number of green Ns in the lower line. Red nucleotides depict mismatches. The number of mismatches is displayed, not including positions with degenerate nucleotides (unless mismatches occur at degenerate positions). The next four columns list the matched sequence's "Species", "Chromosomal Coordinates [start..end]", "Strand" and "Links to Genome Browsers". Clicking on the HTML links to the right of a matched genomic sequence will open a browser in either the ENSEMBL, UCSC or NCBI genome browsers. This will direct the user to that exact location, allowing one to identify whether that targeted sequence is in an annotated gene, intron, exon or regulatory sequence.

ZFN-Site can be used to determine if ZFNs may be used to specifically target sites in multiple different genomes. ZFN-Site can scan multiple genomes simultaneously using the same settings or can be run sequentially.

Relaxed Specificity Search

Previous *in vitro* and cellular ZFP specificity studies may help determine other sequences that may be possibly

WORD MATCHES	Mismatches	Species	Chromosomal Coordinates [start..end]	Strand	Links to Genome Browsers
CGGAGCCGCTTTAACCCACTCTGTGGAAG NNNNN	0	Human	NC_000023.9[70245862..70245890]	-	ENSEMBL UCSC NCBI
TGGAGCCTCTTTGGGGAAACTCAGAGGAAG C G NNNNNN T T	4	Human	NC_000015.8[87749357..87749386]	-	ENSEMBL UCSC NCBI
CTGCCACAGAGATGGAATCTCTGAGGAAG T TNNNNNA T	4	Human	NC_000002.10[49348139..49348167]	+	ENSEMBL UCSC NCBI
CTTCCACAGGATCCAGGGCTCTCTGGAAG AG NNNNNA G	4	Human	NC_000002.10[62504447..62504475]	+	ENSEMBL UCSC NCBI
CTTATACAGAGTTGTTGAATCTGTGTAAG CC NNNNN C G	4	Human	NC_000004.10[102936124..102936152]	+	ENSEMBL UCSC NCBI
CTGCCCCAGAGTATATAACACTGTGGCAG T A NNNNN T A	4	Human	NC_000005.8[150845059..150845087]	+	ENSEMBL UCSC NCBI
CTTCCACAGATAGTGAGACTCTCTTGAAG GTNNNNN G G	4	Human	NC_000010.9[77307499..77307527]	+	ENSEMBL UCSC NCBI
CTTCCACACAGTGGTATTCTCTATGGAAG G NNNNNA G	3	Human	NC_000011.8[127323186..127323214]	+	ENSEMBL UCSC NCBI
CTTCCATATAGTTAGAGACTCTGTGGCAG C G NNNNN A	3	Human	NC_000012.10[22563129..22563157]	+	ENSEMBL UCSC NCBI
CTTCCATATAGTGCTCTCCTCTGTGGAAC C G NNNNNA G	4	Human	NC_000012.10[94664562..94664590]	+	ENSEMBL UCSC NCBI
CTCCCACAGATTTTGTTTGTCTGTGGAAG T G NNNNNAC	4	Human	NC_000017.9[29351084..29351112]	+	ENSEMBL UCSC NCBI
CATCCACAGATTTTTTAAATCTGTGGATG T G NNNNN C A	4	Human	NC_000018.8[44900912..44900940]	+	ENSEMBL UCSC NCBI
CTACCACAGACTGCCTAACTCTGTTGAAC T G NNNNN G G	4	Human	NC_000021.7[16363930..16363958]	+	ENSEMBL UCSC NCBI
TTTCCACAGACTCCCAGAATCTCTGGAAG C G NNNNN C G	4	Human	NC_000023.9[45980147..45980175]	+	ENSEMBL UCSC NCBI
CTTTCAAAGAGTGACATGACACCGTGGAAG C C NNNNNN T T	4	Human	NC_000001.9[4044449..4044478]	+	ENSEMBL UCSC NCBI
TTTCCACAGAGGTTGTACATTCTGTGGAAT C TNNNNNN C G	4	Human	NC_000001.9[17697704..17697733]	+	ENSEMBL UCSC NCBI
CTTCCACTGTGTCCTAGAACTCTGTGCAAG A A NNNNNN G	3	Human	NC_000002.10[188445214..188445243]	+	ENSEMBL UCSC NCBI
CTTCAACAGAGCTCTGTTCCTCTGTAGAAG C TNNNNNNA G	4	Human	NC_000006.10[69350358..69350387]	+	ENSEMBL UCSC NCBI
ATTGCACAGAGTTAAATAACACTGAGGAAG C C NNNNNN T T	4	Human	NC_000007.12[46656492..46656521]	+	ENSEMBL UCSC NCBI
CTGCCAGAGAGTTTTGAAACACTGTGGAGG T C NNNNNN T A	4	Human	NC_000007.12[115962233..115962262]	+	ENSEMBL UCSC NCBI
CTTCCAGAAAGTGAACTGACTCAGTGGAAA C G NNNNNN T G	4	Human	NC_000009.10[121966107..121966136]	+	ENSEMBL UCSC NCBI
CTACCACAAAGTTTTTCTACTATGTGGATG T G NNNNNN C A	4	Human	NC_000014.7[32481511..32481540]	+	ENSEMBL UCSC NCBI
CTTCCCCAGAGAGTCCCTACTGTGGGAAG A TNNNNNN C T	4	Human	NC_000021.7[31644351..31644380]	+	ENSEMBL UCSC NCBI

Figure 2 ZFN-Site Results. ZFN-Site output listing the IL2R-γ target sequence, in row 1, and other genomic sequences matching the search criteria in Figure 1. Non-matching bases are shown in red below the correct base. Between each pair of target sequences is a spacer with its genomic sequence shown in blue. The number of nucleotides in the spacer is indicated by the number of green Ns. Each sequence row also lists the number of mismatches, chromosomal location, DNA strand and HTML links to their exact location on ENSEMBL, UCSC, NCBI and NCBI browsers. The link to results in text format provides sequences in the list ordered by increasing number of mismatches.

cleaved by a ZFN pair. This information can come from studies of individual fingers [30-32]. Without Systematic Evolution of Ligands by Exponential Enrichment (SELEX) or similar data (described below), the specificity of a ZFN can be approximated by combining the specificity of the individual fingers, even though this fails to account for the effects of adjacent fingers. There are many manuscripts detailing individual ZFP specificity; non-exhaustive examples include [30-35]. Approximating the specificity of the whole ZFN by compiling the relaxed specificity of the constituent ZFPs may provide more predictive results than using the basic target search, as the individual finger data may help determine the non-specified bases. If there are individual nucleotide positions where the ZFPs can bind several nucleotides, standard IUPAC ambiguity codes should be entered in the half-site.

More specific information comes from binding studies of full ZFPs or ZFNs using SELEX. Searches based on

experimentally determined specificity are more informative than searches with increased mismatches. If there is SELEX or similar data describing each ZFN's binding specificity, it is also entered in 5' to 3' orientation using standard IUPAC ambiguity codes (as in Figure 3). This allows relaxed specificity searches. For example, a nucleotide in a half-site that can be bound if it is either G or T can be entered as a K. Any non-specified position can be represented by an N (N=A, C, G or T). If scanning with two mismatches, the pair of half-sites should contain less than three ambiguities to prevent computational stalling (see above).

FetchGWI

ZFN-Site uses FetchGWI to perform rapid and accurate searches of the large sequence databases comprising full genomes. FetchGWI is a C program that relies on precomputed genome indices and is best used in cases where queries must be mapped very rapidly and efficiently. To get maximal search speed, FetchGWI only searches within the index files that represent the genome sequences. There is one index entry for each nucleotide in the genome. This exhaustive index also ensures that no match can possibly be missed. Other programs, such as BLAST, occasionally scan non-overlapping words and thus can miss possible off-target sites (see below) [20].

Testing Located Off-Target Sites

Predicted genomic off-target sites should be tested for cleavage. The HTML links are used to download the sequences flanking the site, for use in designing amplification primers for either mutation or sequence analysis. The listed potential off-target sites can be assayed by PCR and mutation detection [7] or deep sequencing [5] to determine ZFN specificity.

If ZFN-Site locates more sites that match the selected criteria than can be tested, the criteria may be narrowed by using less mismatches or using less ambiguous nucleotides for relaxed searches. The list of found sites can also be narrowed using the text output. If the text output link is clicked, the found sites are outputted in another screen in order of increasing number of total mismatches. If a search is conducted using two mismatches per half-site, the output can be greatly narrowed by selecting the genomic sequences at the top of the list with three or fewer total mismatches.

This list of possible target sequences can be further analysed using other computer programs. For example, the output can be ranked using an excel spreadsheet containing a positional weight matrix based on experimentally determined specificity data as described below.

Results

ZFN-Site was validated by comparing our results to a previously published study by Perez *et al.* [7]. Perez *et al.* looked for off-target cleavage by a pair of ZFNs specific for the gene coding for human C-C chemokine receptor type 5 (CCR5). This study used an unpublished algorithm to identify potential off-target sites by scanning the human genome using *in vitro* SELEX selection specificity data [7]. Their sequencing of the identified off-target sites revealed that a site in the related CCR2 gene was also cleaved at a low frequency. The left and right ZFN half-sites, including ambiguities suggested from their SELEX data, were compiled and entered into ZFN-Site (Figure 3). ZFN-Site found the CCR5 target site and each of the off-target sites on their list, including the experimentally verified CCR2 off-target cleavage site (Figure 4). Additional file 1, Figure S1 contains ZFN-Site output with less than three total mismatches.

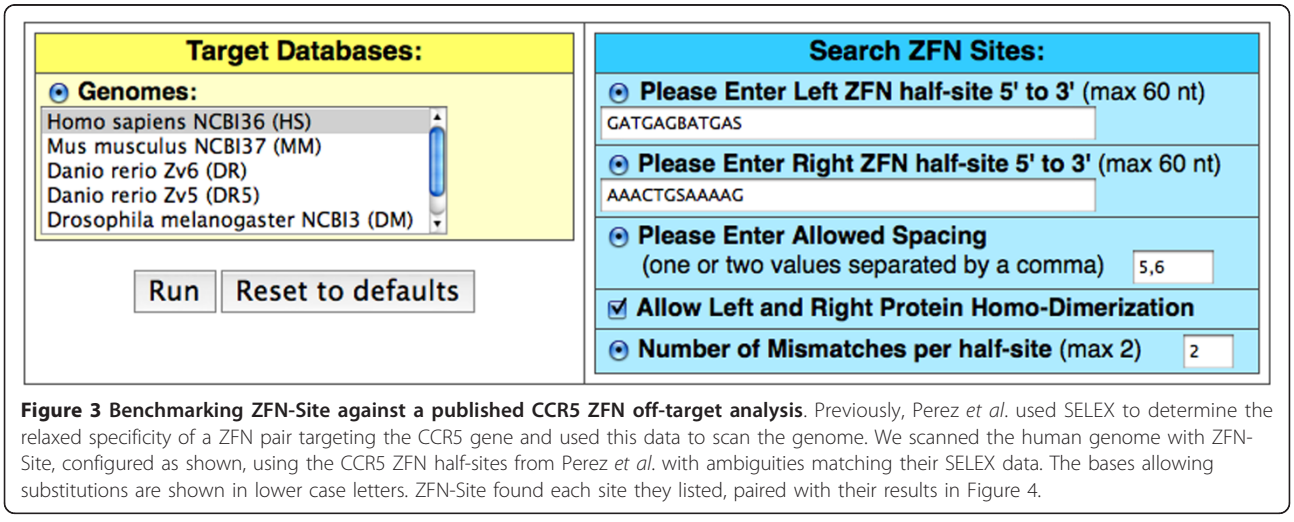


Figure 3 Benchmarking ZFN-Site against a published CCR5 ZFN off-target analysis. Previously, Perez *et al.* used SELEX to determine the relaxed specificity of a ZFN pair targeting the CCR5 gene and used this data to scan the genome. We scanned the human genome with ZFN-Site, configured as shown, using the CCR5 ZFN half-sites from Perez *et al.* with ambiguities matching their SELEX data. The bases allowing substitutions are shown in lower case letters. ZFN-Site found each site they listed, paired with their results in Figure 4.

WORD MATCHES/Mismatches		Chromosomal Coordinates [start..end]	Strand	REF
GTCATCCTCATCCTGATAAACTGCAAAAG S V NNNNN Y S	0	NC_000003.10[46389548..46389576]	+	CCR5
GTCATCCTCATCTCACGGATGAGGATGCC S V NNNNN B AS	1	NC_000004.10[8165376..8165404]	+	1
GTTTTGCAGTTTCACCTCAAACTGCAAAAG C S R NNNNNN Y S	1	NC_000014.7[87308760..87308789]	+	2
CTTTTGCTGTTGCACCTCAAACTGCAAAAG S R TNNNNNN Y S	1	NC_000017.9[64617754..64617783]	+	3
GGCCTCCTCATCATAGCAGGTGAGGATGAC ST A V NNNNNN A R S	3	NC_000016.8[46391387..46391416]	+	4
GTCATCCTCAGCGCCATCGATGAACATGAC S V T NNNNNN GB S	2	NC_000001.9[64870982..64871011]	+	5
GTTTTGCTGTTTCAGCTTAAACTGCAAAAG C S R NNNNNN Y S	1	NC_000021.7[32319967..32319996]	+	6
GTTTTGCTGTTTCACCTAAAACTGCAAAAG C S R NNNNNN Y S	1	NC_000008.9[78368434..78368463]	+	7
GTCGTCCTCATCTTAATAAACTGCAAAAA S A V NNNNN Y S G	2	NC_000003.10[46374209..46374237]	+	8
CTTTTCCAGTTCTAACAAAACTAGAAAAG S R TNNNNN YGS	2	NC_000005.8[4950370..4950398]	+	9
TTTTTCCAGTTTGAATAGAATCTGCAAAAG C S R NNNNNN A Y S	2	NC_000018.8[49408236..49408265]	+	10
CTTATGCAGTTTGTCTATAAACTGGAAAAA T S R NNNNNN Y S G	2	NC_000008.9[120585181..120585210]	+	11
ATCATCCTCAGCAAACTAAAACAGGAAAAG S V T NNNNNN Y S	1	NC_000023.9[136017981..136018010]	+	12
GTTATCCTCAGCAAACTAAAACTGGAACAG S C V T NNNNNN Y S A	3	NC_000007.12[70557254..70557283]	-	13
CTCATCCTCATCCATGCACAATGCAAAAG S V NNNNN A CY S	2	NC_000023.9[50149961..50149989]	+	14
GTCATCCTCAGCATGGGAAACAGCAGAAG S V T NNNNN Y S A	2	NC_000002.10[154567664..154567692]	-	15

Figure 4 ZFN-Site returns sites found in previous CCR5 ZFN off-target analysis. The sequences returned by ZFN-Site were matched to the sequences found by Perez *et al*. For clarity of presentation, the ZFN-Site output was arranged to match the order of Perez *et al*. ZFN-Site found all the sites found by the unpublished algorithm of Perez *et al*., thus validating ZFN-Site. We replaced the column containing the genome browser links with a column referencing the order listed in the first column of Perez, *et al*. Columns detail the human sequences matching the query, the chromosomal coordinates and the strand. The first row consists of the exact match to the CCR5 genomic sequence.

Multiple BLAST searches sometimes accomplish the same function as ZFN-Site if one inputs all possible permutations of homo/heterodimers, spacings and relaxed specificities. This can be labor-intensive. For example, six BLAST queries for permutations of the Perez *et al*. ZFNs could replace one ZFN-Site search without ambiguities (Figure 5). However, in contrast to ZFN-Site, BLAST does not allow ambiguous bases. While BLAST could return these sites, user intervention would be required to distinguish these from true mismatches. ZFN-Site thus simplifies the process of searching for ZFN off-target sites.

ZFN-Site locates every site matching the specified search criteria. In contrast, it has been noted that the BLAST methodology may not find every ZFN site [20]. Because BLAST searches implement a local alignment

```
ZFN-L half site = GATGAGGATGAC
ZFN-R half site = AAACTGCAAAAG

Query Strings With 5bp Spacing:
L+R GTCATCCTCATCnnnnnAAACTGCAAAAG
L+L GTCATCCTCATCnnnnnGATGAGGATGAC
R+R CTTTTGCAGTTTnnnnnAAACTGCAAAAG

Query Strings With 6bp Spacing:
L+R GTCATCCTCATCnnnnnnAAACTGCAAAAG
L+L GTCATCCTCATCnnnnnnGATGAGGATGAC
R+R CTTTTGCAGTTTnnnnnnAAACTGCAAAAG
```

Figure 5 ZFN half-sites and resulting query sequences with 5 or 6 bp spacing. ZFN-Site generates six query sequences based on the two half-sites entered for homo-dimerization, and two different spacings are allowed between the half-sites. The left and right half-sites are listed followed by the resulting query sequences if the half-sites are separated by 5 or 6 bps. Each list includes a query string made of one left and one right half-site (L + R), two left half-sites (L + L) and two right half-sites (R + R). Each of these would need to be searched individually if using BLAST.

search, they are incapable of reproducing the same type of results as ZNF-Site. To compare results to the single ZFN-Site search above, six sets of BLAST searches for the CCR5 ZFN pair were done to include homo- and hetero-dimerization at both 5 bp and 6 bp spacing. Some of the sites found by Perez *et al*. and by ZFN-Site were not found using BLAST, although the BLAST parameters were optimized to attempt to return all matches (Additional file 2, Figure S2). The BLAST search for the right homo-dimer pair with six base spacing failed to return two sequences found by Perez *et al*. and ZFN-Site (numbers 10 and 11). This search returned 474 genomic sequences, many of which were too dissimilar to be likely off-target sites. Because BLAST outputs the matching portion of the sequences with the ends truncated, further user intervention was required to verify the total similarity of these sequences.

In some cases, ZFN-Site may return a large number of sequences. The degree to which one may wish to narrow a list of ZFN-Site outputs depends on the experimental means used to search for off-target cleavage and the resources for scanning multiple sites. The use of deep sequencing may require less narrowing of the list because one can quantitatively test hundreds of sites. Until more information is available on the actual prevalence of ZFN off- target cleavage, it would be desirable to test as many potential off-target sites as experimentally feasible.

A post-processing step using positional weight matrices (PWM) can be used to rank the output of ZFN-Site. Additional file 3 is an example of a spreadsheet used to rank ZFN-Site output using PWMs based on the graph of nucleotide frequencies in Perez *et al*. [7]. The top putative target sites could then be tested experimentally.

Conclusions

ZFN-Site is applicable to genome searches for pairs of half-sites in nucleases or other types of DNA binding proteins. Here, we have presented a user friendly interface allowing a directed search of multiple genomes and have validated its use for finding ZFN sites and off-target sites in the human genome. Experimental testing for ZFN cleavage at the potential sites found by ZFN-Site using large scale sequencing or mutation detection may provide a more thorough understanding of the determinants of ZFN specificity and allow optimization for decreased off-target cleavage events. These results can also be compared with results from other methods for detecting off-target cleavage and toxicity [13-16].

Recently, other nucleases, such as TALE nucleases, have been used for genome alteration [36-39]. While ZFN-Site was tailored to locate ZFN off-target sites, it can also be used to find targets for TALE nucleases. A spreadsheet for creating PWMs and ranking output for TALE nucleases is available upon request.

Availability and requirements

ZFN-Site is available freely on our web site, http://ccg.vital-it.ch/tagger/targetsearch.html[40], and the FetchGWI source code is also available at Source Forge, http://sourceforge.net/projects/tagger/[41].

Project name: ZFN-Site

Project home page: http://ccg.vital-it.ch/tagger/targetsearch.html

Operating system(s): Platform independent

Programming language: C and Perl

Other requirements: None

License: GNU General Public License (GPL), version 2

Any restrictions to use by non-academics: No

Additional material

Additional File 1: Figure S1 - Genomic sites located by ZFN-Site with up to three mismatches. ZFN-Site was run using two mismatches and two ambiguities per half-site as in Figure 3. Genomic sites were located that matched each site found by Perez *et al*. [7] as shown in Figure 4. This comparison provides validation for ZFN-Site. Numerous other sites not described in Perez *et al*. were also found by ZFN-Site, and these can be analyzed experimentally in order to determine if they are actual off-target sites. The text output was sorted by increasing number of mismatches for each genomic location. This is the full list of genomic sequence with three or fewer total mismatches from the half-sites. Mis, # of mismatches; Ch, chromosome; strand, DNA strand

Additional File 2: Figure S2 - BLAST search failed to return the full list of potential off-target sites. Because BLAST searches implement a local alignment search, they are incapable of reproducing the same type of results as ZNF-Site; this is demonstrated by the output of one BLAST

search that failed to find some off-target sites but returned many irrelevant sites. BLAST searches were run using each of the six half-site combinations from Perez *et al*. [7]. This figure shows the results from the BLAST search consisting of two right half-sites separated by six bases (CTTTTGCAGTTT nnnnnn AAACTGCAAAAG). To increase the likelihood of returning all relevant sequences, the EXPECT parameter was raised to a low stringency value of 100, and the penalty for a nucleotide mismatch was dropped to -1. Of the six sequences of this type previously located by Perez *et al*. and by ZFN-Site (Figure 4), BLAST did not locate two sequences (sequences 10 and 11 from Perez *et al*. [7]). BLAST did locate four of the six sequences (sequences 2, 3, 6 and 7) and six similar sequences but also returned 474 sequences that were dissimilar enough to be unlikely to mediate ZFN cleavage. BLAST returns matches in both the forward and reverse DNA strand as indicated in the far right column. The fifth column contains a comparison of the BLAST result to the reference sequence. Mismatches are indicated by an A, C, G or T. A mismatched base not returned by BLAST is shown by an X. Bases truncated at the end of the query sequence are show by a "?", as the user would have to refer back to the genomic sequence to determine if the bases indicated by "?" matched the query sequence, unlike in ZFN-Site. Because BLAST uses a strictly local alignment algorithm, non-matching ends are automatically truncated from the query in order to keep the total number of mismatches low. With the mismatch penalty used in this search, the percent difference threshold for truncating ends is 50%. This figure shows that potential off-target sites can be found using BLAST, but BLAST misses some potential off-target sites. BLAST returns many extraneous sites requiring evaluation through additional processing steps and is much more cumbersome to use than ZFN-Site. Chrom, chromosomal location

Additional File 3: Example of a spreadsheet for ranking ZFN-Site output. The first tab in the spreadsheet provides instruction for ranking ZFN-Site text output based on specificity data. SELEX data are converted to positional weight matrices (PWM). Where base frequencies were zero, a very small, arbitrary pseudocount factor was used to prevent multiplication by zero. PWM were used to score the half-sites of each genomic sequence, assign the matching target sequence and compute the ranking score. The genomic sequences were ranked based on these numbers. Sorting by these scores allowed the choice of sequences most similar to the specificity data.

Acknowledgements
The authors would like to thank Ramona McCaffrey for editorial assistance. This work was supported by the National Institutes of Health [grant number R01 5R01AI068885-03] (TC & AM). Conflict of Interests: none declared.

Author details
[1]University of Iowa School of Medicine, Department of Internal Medicine, Iowa City, Iowa, 52245, USA. [2]Swiss Institute of Bioinformatics (SIB), Bâtiment Génopode, Université de Lausanne, 1015 Lausanne, Switzerland. [3]Ecole Polytechnique Federale de Lausanne (EPFL), Swiss Institute for Experimental Cancer Research (ISREC), 1015 Lausanne, Switzerland. [4]Ludwig Institute for Cancer Research (LICR), Bâtiment Génopode, Université de Lausanne, 1015 Lausanne, Switzerland.

Authors' contributions
TJC provided the initial concept, methods and pseudo-code. GA redesigned the querying methods and implemented the Web interface. TJC tested and benchmarked early versions and provided spreadsheet and supplemental files. CI developed the FetchGWI interface with contribution from GA. TJC & APM wrote the manuscript with contributions from GA and PB.

Received: 11 September 2010 Accepted: 13 May 2011
Published: 13 May 2011

References
1. Porteus MH, Baltimore D: **Chimeric nucleases stimulate gene targeting in human cells.** *Science* 2003, **300(5620)**:763.
2. Bitinaite J, Wah DA, Aggarwal AK, Schildkraut I: **FokI dimerization is required for DNA cleavage.** *Proc Natl Acad Sci USA* 1998, **95(18)**:10570-10575.
3. Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC: **Highly efficient endogenous human gene correction using designed zinc-finger nucleases.** *Nature* 2005.
4. Beumer K, Bhattacharyya G, Bibikova M, Trautman JK, Carroll D: **Efficient gene targeting in Drosophila with zinc-finger nucleases.** *Genetics* 2006, **172(4)**:2391-2403.
5. Meng X, Noyes MB, Zhu LJ, Lawson ND, Wolfe SA: **Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases.** *Nat Biotechnol* 2008, **26(6)**:695-701.
6. Morton J, Davis MW, Jorgensen EM, Carroll D: **Induction and repair of zinc-finger nuclease-targeted double-strand breaks in Caenorhabditis elegans somatic cells.** *Proc Natl Acad Sci USA* 2006, **103(44)**:16370-16375.
7. Perez EE, Wang J, Miller JC, Jouvenot Y, Kim KA, Liu O, Wang N, Lee G, Bartsevich VV, Lee YL, *et al*: **Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases.** *Nat Biotechnol* 2008, **26(7)**:808-816.
8. Cai CQ, Doyon Y, Ainley WM, Miller JC, Dekelver RC, Moehle EA, Rock JM, Lee YL, Garrison R, Schulenberg L, *et al*: **Targeted transgene integration in plant cells using designed zinc finger nucleases.** *Plant Mol Biol* 2009, **69(6)**:699-709.
9. Geurts AM, Cost GJ, Freyvert Y, Zeitler B, Miller JC, Choi VM, Jenkins SS, Wood A, Cui X, Meng X, *et al*: **Knockout rats via embryo microinjection of zinc-finger nucleases.** *Science* 2009, **325(5939)**:433.
10. Hockemeyer D, Soldner F, Beard C, Gao Q, Mitalipova M, Dekelver RC, Katibah GE, Amora R, Boydston EA, Zeitler B, *et al*: **Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases.** *Nat Biotechnol* 2009.
11. Carroll D: **Progress and prospects: zinc-finger nucleases as gene therapy agents.** *Gene Ther* 2008, **15(22)**:1463-1468.
12. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD: **Genome editing with engineered zinc finger nucleases.** *Nat Rev Genet* 2010, **11(9)**:636-646.
13. Szczepek M, Brondani V, Buchel J, Serrano L, Segal DJ, Cathomen T: **Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases.** *Nat Biotechnol* 2007, **25(7)**:786-793.
14. Cornu TI, Thibodeau-Beganny S, Guhl E, Alwin S, Eichtinger M, Joung JK, Cathomen T: **DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases.** *Mol Ther* 2008, **16(2)**:352-358.
15. Pruett-Miller SM, Reading DW, Porter SN, Porteus MH: **Attenuation of zinc finger nuclease toxicity by small-molecule regulation of protein levels.** *PLoS Genet* 2009, **5(2)**:e1000376.
16. Radecke S, Radecke F, Cathomen T, Schwarz K: **Zinc-finger nuclease-induced gene repair with oligodeoxynucleotides: wanted and unwanted target locus modifications.** *Mol Ther* 2010, **18(4)**:743-753.
17. **Zinc Finger Tools.** [http://www.scripps.edu/mb/barbas/zfdesign/zfdesignhome.php].
18. Mandell JG, Barbas CF: **Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases.** *Nucleic Acids Res* 2006, **34(Web Server issue)**:W516-523.
19. **ZiFit.** [http://zifit.partners.org/ZiFiT/].
20. Sander JD, Maeder ML, Reyon D, Voytas DF, Joung JK, Dobbs D: **ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool.** *Nucleic Acids Res* 2010, **38(Suppl)**:W462-468.
21. **ZiFDB.** [http://bindr.gdcb.iastate.edu:8080/ZiFDB].
22. Fu F, Sander JD, Maeder M, Thibodeau-Beganny S, Joung JK, Dobbs D, Miller L, Voytas DF: **Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays.** *Nucleic Acids Res* 2009, **37(Database issue)**:D279-283.
23. Iseli C, Ambrosini G, Bucher P, Jongeneel CV: **Indexing strategies for rapid searches of short words in genome sequences.** *PLoS ONE* 2007, **2(6)**:e579.
24. Bibikova M, Carroll D, Segal DJ, Trautman JK, Smith J, Kim YG, Chandrasegaran S: **Stimulation of homologous recombination through targeted cleavage by chimeric nucleases.** *Mol Cell Biol* 2001, **21(1)**:289-297.
25. Handel EM, Alwin S, Cathomen T: **Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity.** *Mol Ther* 2009, **17(1)**:104-111.
26. Shimizu Y, Bhakta MS, Segal DJ: **Restricted spacer tolerance of a zinc finger nuclease with a six amino acid linker.** *Bioorg Med Chem Lett* 2009.

27. Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I, Beausejour CM, Waite AJ, Wang NS, Kim KA, *et al*: **An improved zinc-finger nuclease architecture for highly specific genome editing.** *Nat Biotechnol* 2007, **25(7)**:778-785.
28. Fajardo-Sanchez E, Stricher F, Paques F, Isalan M, Serrano L: **Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences.** *Nucleic Acids Res* 2008, **36(7)**:2163-2173.
29. Guo J, Gaj T, Barbas CF: **Directed evolution of an enhanced and highly efficient FokI cleavage domain for zinc finger nucleases.** *J Mol Biol* 2010, **400(1)**:96-107.
30. Liu Q, Xia Z, Zhong X, Case CC: **Validated zinc finger protein designs for all 16 GNN DNA triplet targets.** *J Biol Chem* 2002, **277(6)**:3850-3856.
31. Segal DJ, Dreier B, Beerli RR, Barbas CF: **Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences.** *Proc Natl Acad Sci USA* 1999, **96(6)**:2758-2763.
32. Choo Y, Klug A: **Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions.** *Proc Natl Acad Sci USA* 1994, **91(23)**:11168-11172.
33. Dreier B, Segal DJ, Barbas CF: **Insights into the molecular recognition of the 5'-GNN-3' family of DNA sequences by zinc finger domains.** *J Mol Biol* 2000, **303(4)**:489-502.
34. Dreier B, Beerli RR, Segal DJ, Flippin JD, Barbas CF: **Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors.** *J Biol Chem* 2001, **276(31)**:29466-29478.
35. Dreier B, Fuller RP, Segal DJ, Lund CV, Blancafort P, Huber A, Koksch B, Barbas CF: **Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors.** *J Biol Chem* 2005, **280(42)**:35588-35597.
36. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF: **Targeting DNA double-strand breaks with TAL effector nucleases.** *Genetics* 2010, **186(2)**:757-761.
37. Li T, Huang S, Jiang WZ, Wright D, Spalding MH, Weeks DP, Yang B: **TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain.** *Nucleic Acids Res* 2010, **39(1)**:359-372.
38. Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ, *et al*: **A TALE nuclease architecture for efficient genome editing.** *Nat Biotechnol* 2010, **29(2)**:143-148.
39. Mahfouz MM, Li L, Shamimuzzaman M, Wibowo A, Fang X, Zhu JK: **De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks.** *Proc Natl Acad Sci USA* 2011, **108(6)**:2623-2628.
40. **ZFN-Site.** [http://ccg.vital-it.ch/tagger/targetsearch.html].
41. **Source Forge.** [http://sourceforge.net/projects/tagger/].

doi:10.1186/1471-2105-12-152
Cite this article as: Cradick *et al.*: ZFN-Site searches genomes for zinc finger nuclease target sites and off-target sites. *BMC Bioinformatics* 2011 **12**:152.