In silico performance analysis of web tools for CRISPRa sgRNA design in human genes

Cristian N. Nuñez Pedrozo [b,1], Tomás M. Peralta [a,1], Fernanda D. Olea [b], Paola Locatelli [b], Alberto J. Crottogini [b], Mariano N. Belaich [c], Luis A. Cuniberti [b,*]

[a] Departamento de Biología Comparada, Celular y Molecular, FICEN. Universidad Favaloro, Buenos Aires, CABA, C1078AAI, Argentina
[b] Laboratorio de Regeneración Cardiovascular, IMETTYB – Universidad Favaloro – CONICET, Buenos Aires, CABA, C1078AAI, Argentina
[c] Laboratorio de Ingeniería Genética y Biología Celular y Molecular, Universidad Nacional de Quilmes – CONICET. Buenos Aires, Bernal B1876BXD, Argentina

ARTICLE INFO

ABSTRACT

Angiogenic gene overexpression has been the main strategy in numerous vascular regenerative gene therapy projects. However, most have failed in clinical trials. CRISPRa technology enhances gene overexpression levels based on the identification of sgRNAs with maximum efficiency and safety. CRISPick and CHOP CHOP are the most widely used web tools for the prediction of sgRNAs. The objective of our study was to analyze the performance of both platforms for the sgRNA design to angiogenic genes (VEGFA, KDR, EPO, HIF-1A, HGF, FGF, PGF, FGF1) involving different human reference genomes (GRCH 37 and GRCH 38). The top 20 ranked sgRNAs proposed by the two tools were analyzed in different aspects. No significant differences were found on the DNA curvature associated with the sgRNA binding sites but the sgRNA predicted on-target efficiency was significantly greater when CRISPick was used. Moreover, the mean ranking variation was greater for the same platform in EPO, EGF, HIF-1A, PGF and HGF, whereas it did not reach statistical significance in KDR, FGF-1 and VEGFA. The rearrangement analysis of the ranking positions was also different between platforms. CRISPick proved to be more accurate in establishing the best sgRNAs in relation to a more complete genome, whereas CHOP CHOP showed a narrower classification reordering.

© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction:

CRISPRa is a technology derived from CRISPRCas9 that uses the catalytically inactive Cas (dCas9) fused to pro-transcriptional elements as an artificial transcription factor. Recent advances include complexes that interact with multiple pro-transcriptional elements leading to high levels of gene overexpression using single guide RNA (sgRNA) molecules [1]. Indeed, CRISPRa has certain advantages over ORF-based methods such as the overexpression in its native context that allows the production of different splice isoforms, being even valuable to genes with large transcripts [2] or multiplexing [3].

Gene therapy has been originally developed as a strategy for treating inherited monogenic diseases and has obtained clinical approval for its application in various pathologies, including neuromuscular disease and hereditary blindness [4]; furthermore, it is being considered for developing treatments related to acquired diseases like vascular disease. For this, most therapies are based on the overexpression of angiogenic genes [5,6,7]. Recently, a therapy based on an intramuscular injection of a plasmid carrying a hepatocyte growth factor gene has received approval for clinical use in patients with critical limb ischemia (CLI) [8]. Although, CLI has been the target of several gene and cell therapy approaches during the last 20 years with poor results [9].

The advantages of the CRISPRa system make it very interesting as a novel alternative for its application in angiogenic gene therapy. However, on-target efficiency and potential off-target prediction still limit CRISPRa applications [10]. Regarding this, several web-based sgRNA design tools like CHOP CHOP [11] and CRISPick [12] were developed. These platforms consider the GC content, RNA secondary structure, thermodynamics, recognition sites for restriction endonucleases and nucleotide identity among other aspects in order to propose candidates contemplating their on-target and off-target scores focusing on maximizing the first one activity while minimizing the second one.

On-target scoring algorithms include Rule Set 2, Rule Set 1, and Moreno-Mateos. Off-target score is calculated by mismatch count, Cutting Frequency Determination (CFD) score and others [13] varying according to platform. One method involves studying mismatch counts by performing alignments of sgRNAs to a reference genome employing alignment tools like Bowtie and Burrows-Wheeler Aligner. On the other hand, the CFD score also considers the RNA secondary structure and its genome target location [14].

Although the reference genome is a key factor for the off-target prediction, a comparative study regarding the performance of algorithms in the design of sgRNAs using reference human genome versions of different complexity has not yet been evaluated. Currently, the genome reference consortium human build 38 (GRCH38) offers more extensive information compared to the version 37 (GRCH37). Thus, our objective was to analyze the performance of CRISPick and CHOP CHOP web tools for the sgRNA design to angiogenic genes involving different human reference genomes.

2. Methods

2.1. Gene dataset selection and sgRNA design

The Vascular Endothelial Growth Factor A (*VEGFA*), Kinase Insert Domain Receptor (*KDR*), Erythropoietin (*EPO*), Hypoxia Inducible Factor 1 Subunit Alpha (*HIF-1A*), Hepatocyte Growth Factor (*HGF*), Epidermal Growth Factor (*EGF*), Placental Growth Factor (*PGF*) and Fibroblast Growth Factor 1 (*FGF1*) gene dataset was selected from Genecard [15] according to their reported angiogenic activity and used to design CRISPRa sgRNAs applying CHOP CHOP and CRISPick based on the human reference genomes GRCH37 and GRCH38. The software setting had the following configuration: 300nt target window upstream of the transcription start site (TSS), on-target efficiency score Doench 2016, SpCas9, and the remaining options on default. The top 20 ranked sgRNAs targeting each gene were selected for further analysis. The Eucaryotic promoter database (EPD) and BLAST were used to check the TSS [16].

2.2. DNA curvature and on-target efficiency analysis

The web tool Bend.it [17] was used to predict and graph DNA curvature and GC content for each of the promoter sequences obtained from EPD for each gene. Then the regions where the sgRNAs align in GRCH38 were identified, average and compared. Moreover, each sgRNA of the top 20 designed by CRISPick and CHOP CHOP had a predicted on-target efficiency score defined by Doench 2016. This efficiency values were analyzed for each gene to compare the platforms.

2.3. Ranking variation and rearrangement analysis

First, we identified the top 20 ranked sgRNAs by CRISPick and CHOP CHOP employing GRCH37 and GRCH38. Then, the ranking position variation for each sgRNA was analyzed between both reference genomes. The position variation was expressed as an absolute value and a raw value. The absolute values were averaged for each gene to compare the platforms regarding their ability to analyze genomes of different complexity. The raw values were averaged for each gene in order to analyze if the position variation rearrangement maintains the same sgRNAs in the top 20 ranking.

2.4. Statistical analysis

Student paired T-test was applied for the DNA curvature and efficiency analysis. Wilcoxon test was applied for the ranking variation and rearrangement analysis. Prism 7.0 software (GraphPad Software Inc, La Jolla, CA, USA) was used. Statistical significance was set at $p < 0.05$. Results are expressed as mean ± standard deviation.

3. Results and discussion

The design of CRISPRa sgRNAs targeting angiogenic promoters by CRISPick and CHOP CHOP involved different versions of the human genome as shown in the workflow (Fig. 1). As expected, the same TSSs employed by CRISPick and CHOP CHOP were found in the reference genomes.

Regarding that DNA secondary structure could be important in steric impediments for the sgRNA-DNA hybridization, the DNA curvature in the sgRNA binding sites was evaluated for each considered gene of GRCH38. No significant differences were found between the candidates offer by CRISPick and CHOP CHOP (CRISPick: 3.120 ± 0.383 vs CHOP CHOP: 3.024 ± 0.583, $p > 0.05$, Student paired T-test) (Fig. 2). In addition, the predicted on-target efficiency score for the top 20 sgRNA candidates was also studied revealing that the mean value was significantly higher when used CRISPick (CRISPick: 55.5 ± 2.7 vs CHOP CHOP: 49.51 ± 2.06, $p < 0.0001$, Student paired T-test) (Fig. 3). The on-target efficiency algorithms to design these RNA molecules are shared between
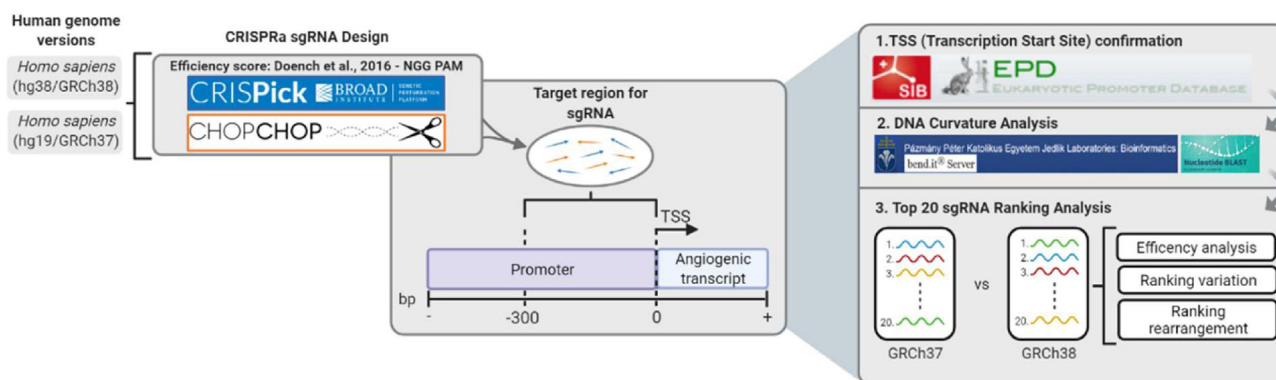

Fig. 1. Workflow of predicted sgRNA performance involving different human reference genomes with CRISPick and CHOP CHOP. On the right side, analysis carried out with the top 20 sgRNA. The predicted on-target efficiency score defined by Doench 2016 between platforms was considered. In addition, the ranking variation and rearrangement regarding the platform ability to adapt to genomes of different complexity were studied.
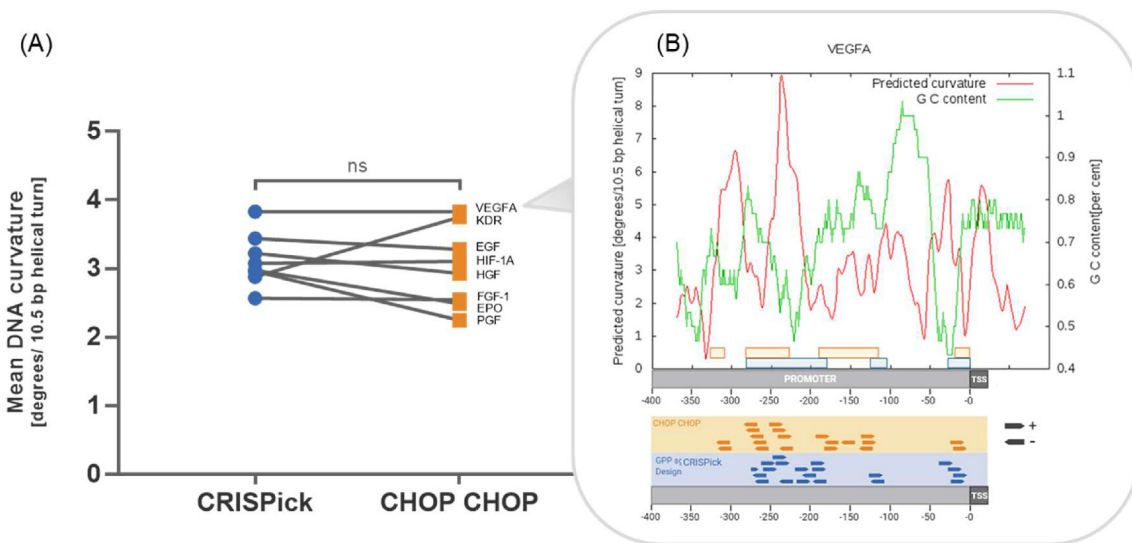
Fig. 2. Mean DNA curvature analysis of sgRNA binding sites. (A) values associated with DNA curvature at the binding sites of the 20 sgRNAs for each gene (GRCH38) were averaged, compared and plotted as blue circles (CRISPick) and orange squares (CHOP CHOP) (ns: not significant, paired Student T-test). (B) Representative figure of VEGFa gene DNA curvature analysis. The upper section shows the graph of DNA curvature and %GC for the nucleotide positions of the promoter region; the areas where the sgRNAs hybridize were shaded in blue (CRISPick) and orange (CHOP CHOP). In the lower section, the target region for each sgRNA in the promoter region is plotted as blue (CRISPick) and orange (CHOP CHOP) arrows. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
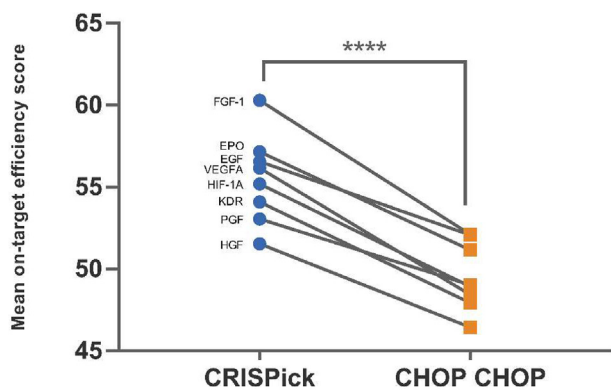

Fig. 3. Top 20 sgRNA efficiency. The individual predicted efficiency values of each sgRNA of the top 20 for each gene (GRCH38) were averaged, compared, and plotted as blue circles (CRISPick) and orange squares (CHOP CHOP) (****: p < 0.0001, paired Student T-test). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
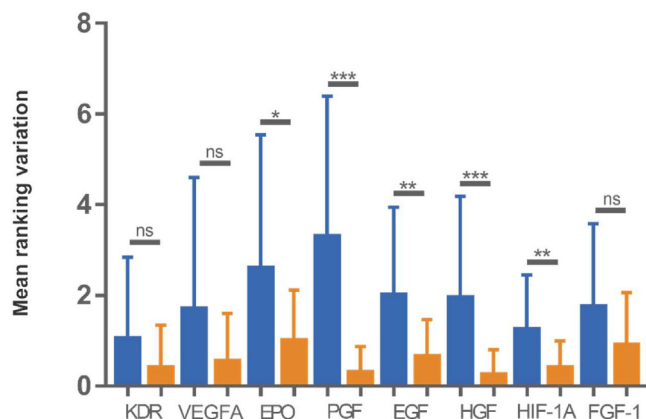

Fig. 4. Mean ranking variation analysis. The absolute values of the ranking position variation for each sgRNA were averaged for each gene to compare the platforms regarding their ability to analyze genomes of different complexity. Then, we compared and plotted as blue (CRISPick) and orange (CHOP CHOP) bars (ns: not significant, *: p < 0.05, **: p < 0.01, ***: p < 0.001, Wilcoxon test). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

both web tools; however, the mean on-target efficiency was significantly different for the top 20 sgRNAs. This situation could be associated with the algorithms used by the platforms to the off-target calculation. CHOP CHOP focusses this study on mismatch count; thus, sgRNAs displaying a good on-target efficiency but having off-targets are penalized. Meanwhile, CRISPick bases the calculation on a CFD score that considers both the mismatch count and the genomic activity in the off-target site for the sgRNA ranking.

Concerning the elaboration of the top 20 sgRNA for each platform, the ranking position variation for each sgRNA was analyzed between both reference genomes showing that it was greater when used CRISPick in EPO (p < 0.05), EGF (p < 0.01), HIF-1A (p < 0.01), PGF (p < 0.001) and HGF (p < 0.001), whereas it did not reach statistical significance in KDR, FGF-1 and VEGFA (Wilcoxon Test) (Fig. 4). This could be explained because CRISPick considers the level of genomic activity of the off-target sites in order to

classify the sgRNAs while CHOP CHOP solely regards mismatch count. Moreover, the top 3 position of the CRISPick ranking for all genes had slight variation, making the top 3 sgRNAs an optimal choice when using this platform. Then, the rearrangement dynamic within the top 20 sgRNA ranking for each web-tool were evaluated and significant differences were found (CRISPick:-0.3187 ± 0.2698 vs CHOP CHOP:-0.0437 ± 0.0563, p < 0.05, Wilcoxon test) (Fig. 5). The rearrangement of sgRNAs designed by CHOP CHOP showed an average variation very close to 0, mainly because the ranking position changes always involved the same sgRNAs of the top 20 revealing the low adaptation of the platform to a more complete genome. In contrast, CRISPick included new sgRNAs in the top 20 showing its high performance when working with new genomic data.
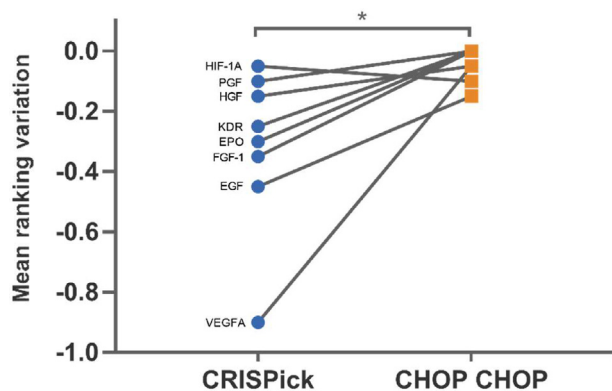
Fig. 5. Ranking position rearrangement. The raw values of the ranking position variation for each sgRNA were averaged for each gene to analyze if the rearrangement maintains the same sgRNAs in the top 20 ranking. Then, we compared and plotted as blue circles (CRISPick) and orange squares (CHOP CHOP) (*: $p < 0.05$, Wilcoxon test). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Concluding remarks

The CRISPRa technology is a great opportunity for novel regenerative vascular therapies. The correct design of sgRNAs is crucial to achieve safe strategies that do not affect non-target genomic regions. The development of bioinformatic methods and user-friendly platforms have transformed the design of sgRNAs for CRISPR/dCas approaches in the recent years. Nevertheless, a consensus for sgRNA design that considers genome information dynamics is still needed. Although CRISPick showed better in silico performance than CHOP CHOP, further in vitro efficiency analysis will be required to validate the present computational results.

Funding

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] La Russa MF, Qi LS. The new state of the art: Cas9 for gene activation and repression. Mol Cell Biol 2015;35(22):3800–9. https://doi.org/10.1128/mcb.00512-15.
[2] Kampmann M. CRISPRi and CRISPRa screens in mammalian cells for precision biology and medicine. ACS Chem Biol 2017;13(2):406–16. https://doi.org/10.1021/acschembio.7b00657.
[3] McCarty NS, Graham AE, Studená L, Ledesma-Amaro R. Multiplexed CRISPR technologies for gene editing and transcriptional regulation. Nat Commun 2020;11(1). https://doi.org/10.1038/s41467-020-15053-x.
[4] Bulaklak K, Gersbach CA. The once and future gene therapy. Nat Commun 2020;11(1). https://doi.org/10.1038/s41467-020-19505-2.
[5] Ylä-Herttuala S, Rissanen TT, Vajanto I, Hartikainen J. Vascular endothelial growth factors. J Am Coll Cardiol 2007;49(10):1015–26. https://doi.org/10.1016/j.jacc.2006.09.053.
[6] Giménez CS, Castillo MG, Simonin JA, Núñez Pedrozo CN, Pascuali N, del Bauzá M, et al. Effect of intramuscular baculovirus encoding mutant hypoxia-inducible factor 1-alpha on neovasculogenesis and ischemic muscle protection in rabbits with peripheral arterial disease. Cytotherapy 2020;22 (10):563–72. https://doi.org/10.1016/j.jcyt.2020.06.010.
[7] Olea FD, Janavel GV, Cuniberti L, Yannarelli G, Meckert PC, Cors J, et al. Repeated, but not single, VEGF gene transfer affords protection against ischemic muscle lesions in rabbits with hindlimb ischemia. Gene Ther 2009;16(6):716–23. https://doi.org/10.1038/gt.2009.30.
[8] MHLW Panel OKs AnGes' gene therapy collategene for conditional approval. PHARMA JAPANFebruary 21, 2019. https://pj.jiho.jp/article/239483.
[9] Ylä-Herttuala S. Gene therapy of critical limb ischemia enters clinical use. Mol Ther 2019;27(12):2053. https://doi.org/10.1016/j.ymthe.2019.11.001.
[10] Liu G, Zhang Y, Zhang T. Computational approaches for effective CRISPR guide RNA design and evaluation. Comput Struct Biotechnol J 2020;18:35–44.
[11] Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res 2019;47(W1):W171–4. https://doi.org/10.1093/nar/gkz365.
[12] Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 2016;34(2):184–91. https://doi.org/10.1038/nbt.3437.
[13] Hanna RE, Doench JG. Design and analysis of CRISPR–Cas experiments. Nat Biotechnol 2020;38(7):813–23. https://doi.org/10.1038/s41587-020-0490-7.
[14] Listgarten J, Weinstein M, Kleinstiver BP, Sousa AA, Joung JK, Crawford J, et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat Biomed Eng 2018;2(1):38–47. https://doi.org/10.1038/s41551-017-0178-6.
[15] Safran M, Rosen N, Twik M, BarShir R, Stein TI, Dahary D, et al. The GeneCards suite. Pract Guide Life Sci Databases 2021;27–56. https://doi.org/10.1007/978-981-16-5812-9_2.
[16] Dreos R, Ambrosini G, Groux R, Cavin Périer R, Bucher P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic Acids Res 2016;45(D1):D51–5. https://doi.org/10.1093/nar/gkw1069.
[17] Vlahovicek K. DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic Acids Res 2003;31(13):3686–7. https://doi.org/10.1093/nar/gkg559.