In silico design and analysis of targeted genome editing with CRISPR

Kornel Labun

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2020

UNIVERSITY OF BERGEN

In silico design and analysis of targeted genome editing with CRISPR

Kornel Labun

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 27.01.2020

Year: 2020

Title: In silico design and analysis of targeted genome editing with CRISPR

Name: Kornel Labun

Print: Skipnes Kommunikasjon / University of Bergen

Scientific Environment

The papers presented in this thesis, as well as the thesis itself are the results of my continuous work at the Valen Group - a part of Computational Biology Unit (CBU) at the Faculty of Mathematics and Natural Sciences of the University of Bergen (UiB).

I was associated with NORBIS (National Research School in Bioinformatics, Biostatistics and Systems Biology) and MCB (Molecular and Computational Biology Research School), which allowed me to take part in relevant and high quality courses across Norway. I was further involved in teaching the following courses: INF207 (Social Network Theory), INF109 (Computer Programming for Science), and R Crash Course for MCB Research School.

My supervisor during all this time, Eivind Valen, relentlessly assured the quality of my work. I was initially co-supervised by David Fredman and at a later point by Pekka Parviainen, both of whom provided constructive feedback.

This scientific opportunity was funded by the Bergen Research Foundation, the Norwegian Research Council (FRIMEDBIO #250049) and University of Bergen core funding.

Acknowledgments

I started my scientific career rather by accident than virtue. I remember enrolling for my first scientific project many years ago, completely unaware of what I was getting into. For pulling my leg into science (twice!), my thanks go to Tomasz. I am grateful for what I have learned over those couple of years. It was a pleasure to discover how difficult it is to be a scientist.

I am grateful for the environment at CBU. It was great to be entertained by you. I wish you all success and many great adventures. I will especially remember my lab members: Adam, Adnan, Gunnar, Håkon, Kasia, Kirill, Max, Teshome and Yamila.

For accompanying me through pain and tears I have to thank my benevolent overlord Eivind, who polished me into the shining diamond that I am today.

But foremost thanks go to those who were most patient with me, my girlfriend Alicja, my mother and sisters. It took quite a while to finish this thesis; without your support it would have been far more tedious and far rainier.

Thanks to all those who entertained me here and there. Thanks to all those that spark a smile when they see me, and thank you to those that rather are, than are not.

Nomenclature and abbreviations

RNA - ribonucleic acid

DNA - deoxyribonucleic acid

CRISPR - clustered regularly interspaced short palindromic repeats

TALENs - transcription activator-like effector nucleases

DSB - double strand break

NHEJ - non-homologous end joining

HDR - homology directed repair

PAM - protospacer adjacent motif

PFS - protospacer flanking sequence

RT-qPCR - real-time quantitative polymerase chain reaction

amplicon - DNA sequence used as a source and product of the RT-qPCR

crRNA - CRISPR RNAs

tracrRNA - trans activating crRNA

gRNA - guide RNA (crRNA + tracrRNA)

protospacer - part of the crRNA that is complementary to the target

Cas9 - CRISPR associated protein 9

dCas9 - dead Cas9

NGS - next-generation sequencing

Abstract

CRISPR/Cas systems have become a tool of choice for targeted genome engineering in recent years. Scientists around the world want to accelerate their research with the use of CRISPR/Cas systems, but are being slowed down by the need to understand the technology and computational steps needed for design and analysis. However, bioinformatics tools for the design and analysis of CRISPR experiments are being created to aid those scientists.

For the design of CRISPR targeted genome editing experiments, CHOPCHOP has become one of the most cited and most used tools. After the initial publication of CHOPCHOP, our understanding of the CRISPR system underwent a scientific evolution. I therefore updated CHOPCHOP to accommodate the latest discoveries, such as designs for nickase and isoform targeting, machine learning algorithms for efficiency scoring and repair profile prediction, in addition to many others.

On the other spectrum of genome engineering with CRISPR, there is a need for analysis of the data and validation of mutants. For the analysis of the CRISPR targeted genome editing experiments, I have created ampliCan, an R package that with the use of 'editing aware' alignment and automated normalization, performs precise estimation of editing efficiencies for thousands of CRISPR experiments. I have benchmarked ampliCan to display its strengths at handling a variety of editing indels, filtering out contaminant reads and performing HDR editing estimates.

Both of these tools were developed with the idea that biologists without a deep understanding of CRISPR should be able to use them, and at the same time seasoned experts can adjust the settings for their purposes. I hope that these tools will facilitate adaptation of CRISPR systems for targeted genome editing and indirectly allow for great discoveries in the future.

List of publications

1. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. K Labun, TG Montague, JA Gagnon, SB Thyme, E Valen, 2016, Nucleic acids research 44 (W1), W272-W276
2. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. K Labun, TG Montague, M Krause, Y Torres Cleuren, H Tjeldnes, E Valen, 2019, Nucleic Acids Research, gkz365
3. Accurate analysis of genuine CRISPR editing events with ampliCan. K Labun, X Guo, A Chavez, G Church, JA Gagnon, E Valen, 2019, Genome Res. 29: 843-847

Other publications (not related to the thesis)

1. RareVariantVis: new tool for visualization of causative variants in rare monogenic disorders using whole genome sequencing data T Stokowy, M Garbulowski, T Fiskerstrand, R Holdhus, K Labun, ... 2016, Bioinformatics 32 (19), 3018-3020
2. *tailfindr*: Alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing M Krause, AM Niazi, K Labun, YN Torres Cleuren, FS Müller, E Valen 2019, RNA 2019

Contents

1. Introduction

1.1. Genome Engineering

Genomes encode the basis for all biological life on our planet. While we have yet to understand how genomes work in every detail, we have already discovered how to manipulate the genome in a variety of ways. Genome editing or genome engineering is widely defined as any kind of genome changing, whether it is to insert new deoxyribonucleic acid (DNA), remove part of the genome sequence, change bases or a mix of the above. Previously, changing of the genome was achieved in a stochastic fashion, through techniques like Agrobacterium-mediated transformation (Schell and Van Montagu 1977), transduction with viral vectors (Goff and Berg 1976), using restriction enzymes (Jeltsch et al. 1996; Schöttler et al. 1998) and mutagenesis induced with chemicals/UV (Russell et al. 1979; Kato, Rothman, and Clark 1977). Naturally, controlled and localized changes allow for more powerful experimental arrangements and have therefore been a focus of extensive research efforts over the years.

Efforts for targeted genome editing were spurred by the discovery that some biological structures or mechanisms (guiding part) can recognize specific genomic sites based on their DNA sequence and introduce a double-stranded break (DSB) using its cutting mechanism (cutting part). The cells then activate their repair pathways, mainly non-homologous end joining (NHEJ) or homology directed repair (HDR), which can repair the DSB. Although HDR is less error prone than NHEJ, both pathways sometimes make erroneous repairs. A possible change in the genome sequence prevents further binding of the guiding element. Using the above technique could lead to a loss-of-function allele (gene knock-out). Gene knock-outs can be used to determine the function of a particular gene in the cell. Furthermore, some DSBs can have ends with complementary sequences (called overhangs or sticky ends). The HDR pathway tries to fix DSBs with overhangs by using template sequence with complementary overhangs. Providing an artificial template with complementary overhangs is used in knock-in techniques to insert the

template into the genome sequence of interest. These prospects stimulated scientists to develop methods for targeted genome editing over the years.

Decades of research resulted in a handful of techniques for editing at the desired location in the genome using various effectors, such as group 2 intron (Chen et al. 2005), Thermus thermophilus Argonaute protein (Swarts et al. 2014), structure-guided endonucleases (S. Xu et al. 2016), λ-bet/exo MAGE (K. Xu, Stewart, and Porter 2015; H. H. Wang et al. 2012), single-stranded oligodeoxyribonucleotides (Aarts and te Riele 2010; Rios et al. 2012), and meganucleases (Donoho, Jasin, and Berg 1998). The most popular nucleases for targeted genome edits in chronological order has been zinc finger nucleases (ZFNs) (Y. G. Kim, Cha, and Chandrasegaran 1996), transcription activator-like effector nucleases (TALENs) (Miller et al. 2011; F. Zhang et al. 2011), and finally clustered regularly interspaced short palindromic repeats (CRISPR/Cas9) (Jinek et al. 2012; Cong et al. 2013; Mali, Yang, et al. 2013; Hsu, Lander, and Zhang 2014).

ZFNs and TALENs require the design and synthesis of the protein for each target locus, which is laborious and costly. CRISPR/Cas9 on the other hand has been demonstrated to require only its crRNA component - its guiding part - to be engineered for each locus. Other parts of the CRISPR/Cas9 system: the tracrRNA and Cas9 protein remain the same for every target of interest (Jinek et al. 2012; Cong et al. 2013; Mali, Yang, et al. 2013; Hsu, Lander, and Zhang 2014). The synthesis of RNA is currently much cheaper, faster and less strenuous than the synthesis of proteins. Thanks to its simplicity, high editing efficiencies and relatively low time cost, CRISPR has become the method of choice for precision genome editing (**Figure 1**).

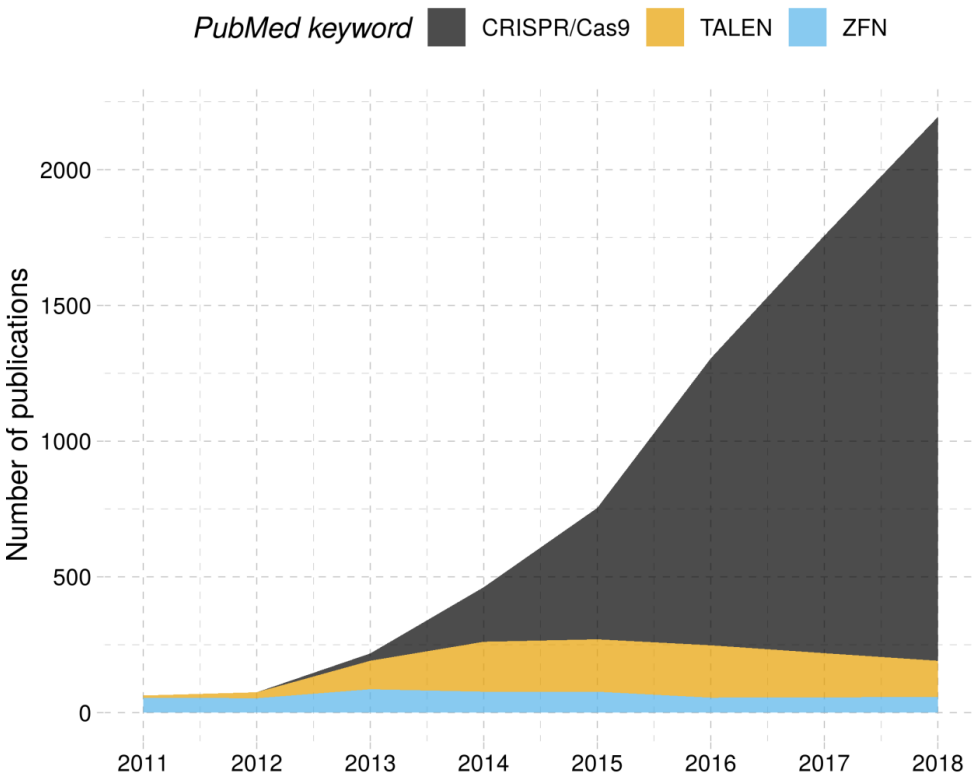


Figure 1. PubMed registered manuscripts by keyword over years. CRISPR/Cas9 genome editing is clearly the dominant strategy since 2013, even with "CRISPR/Cas9" as a strict keyword. The real number of publications that used CRISPR for genome editing applications is much higher since other terms were used.

1.2. CRISPR

1.2.1. Introduction to CRISPR/Cas9

The journey for CRISPR research started with a discovery by Mojica et al. who discovered repeats in the genome of *Haloferax mediterranei* (F. J. Mojica, Juez, and Rodríguez-Valera 1993). Today, by Mojica's suggestion, these clustered regularly interspaced short palindromic repeats are known as CRISPR array. The timeline of some of the key scientific findings leading to the CRISPR/Cas9 system are presented

in **Figure 2**. They unveil how CRISPR systems found their way into genome editing applications.

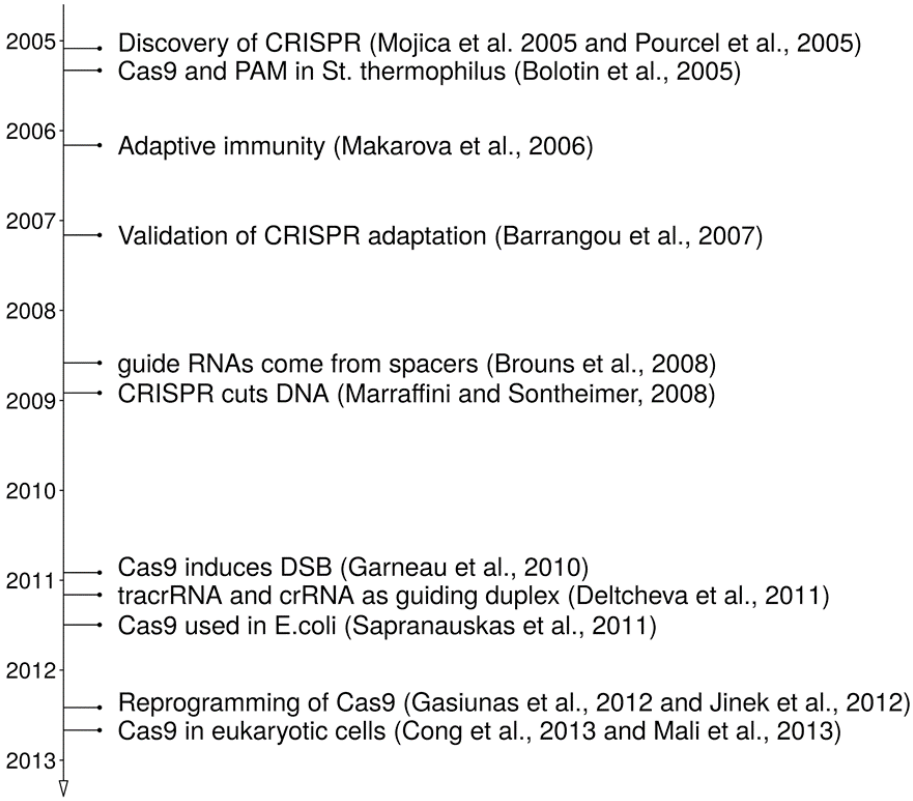


Figure 2. Timeline of CRISPR discovery (Lander 2016).

After many years of research, it became clear that CRISPR is a bacterial adaptive immune system (**Figure 3**), a genomic database to store previous viral aggressor footprints (F. J. M. Mojica et al. 2005; Pourcel, Salvignol, and Vergnaud 2005; Bolotin et al. 2005).

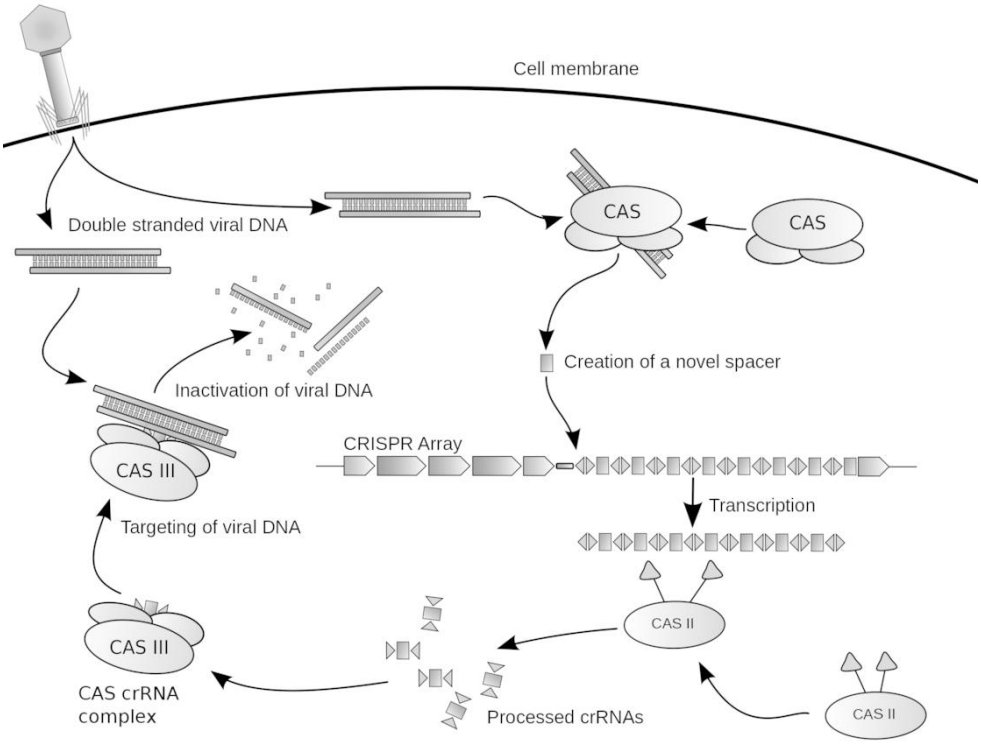


Figure 3. CRISPR as an adaptive immune system. Adapted from: James Atmos, distributed under a CC BY-SA 3.0 license.

In 2005 it was found that the effector protein that has nuclease activity is Cas9 (Bolotin et al. 2005). Interestingly, Cas9 was expressed from cas-associated genes neighboring the CRISPR repeats. Bolotin also discovered that Cas9 recognizes its genomic targets using a protospacer adjacent motif (PAM). This PAM motif must be recognized by the Cas9 protein to activate cleavage, and because it is not present in the CRISPR database, the loci are safe from self-targeting. The next groundbreaking discovery was that the spacer sequences from the CRISPR array are transcribed into crRNAs (Brouns et al. 2008). Two years later it was shown that crRNAs together with the tracrRNA form a duplex that guides Cas9 to its target (Deltcheva et al. 2011). All necessary parts of the CRISPR/Cas9 system were therefore known at this time. The Cas9 acts as the effector, cutting DNA, and the crRNAs and tracrRNA are the guiding part of the system. Reusing

this system in other prokaryotes (Sapranauskas et al. 2011) was the next major step that provided evidence that the system is transferable across species. Ideally, the guiding part of the system should be open for engineering to allow recognition of specific genomic loci. Reprogrammable guiding of Cas9 through changes in the guide RNA (gRNA) sequence was described in 2012 (Gasiunas et al. 2012; Jinek et al. 2012). Finally, using the CRISPR/Cas9 system in eukaryotic cells (human and mouse) enabled the release of the CRISPR/Cas9 system as a general genome engineering tool (Cong et al. 2013; Mali, Yang, et al. 2013).

1.2.2. Elements of CRISPR/Cas9 system

The CRISPR/Cas9 system is naturally composed of three elements: crRNA, tracrRNA and Cas9. Nowadays in genome engineering applications, crRNA and tracrRNA are not used as two distinct components, rather they are bridged by a GAAA tetraloop to form a single guide RNA (gRNA or sgRNA) (Jinek et al. 2012). The gRNA, or more precisely, its spacer part (**Figure 4**) together with the PAM, define the specificity of the system to the genomic location. The PAM is recognized by the Cas9 effector protein in the first step of target recognition. In the next step, high complementarity between the RNA spacer and the target DNA allows an R-loop to form, which in turn facilitates cleavage by Cas9 (Gasiunas et al. 2012). Cas9 uses a RuvC domain to cleave the non-target DNA strand, and an HNH domain to cleave the complementary strand. Cleavage of Cas9 is blunt (no overhangs), and is localized 3-4 bp upstream of the PAM sequence. After the cut, Cas9 releases the DNA and continues to search for the next complementary target site until the protein is degraded. If there is another locus in the genome with a similar sequence to the spacer and PAM, that locus is also cleaved. Loci not accounted for during experiment design and incidentally cleaved by Cas9 are called off-target sites. Off-target sites are dangerous and risk the integrity of the experiment, but not all off-targets are cleaved efficiently. It is understood that different spacers have different cleavage efficiencies, therefore selection of an appropriate spacer is essential for successful CRISPR/Cas9 genome editing.

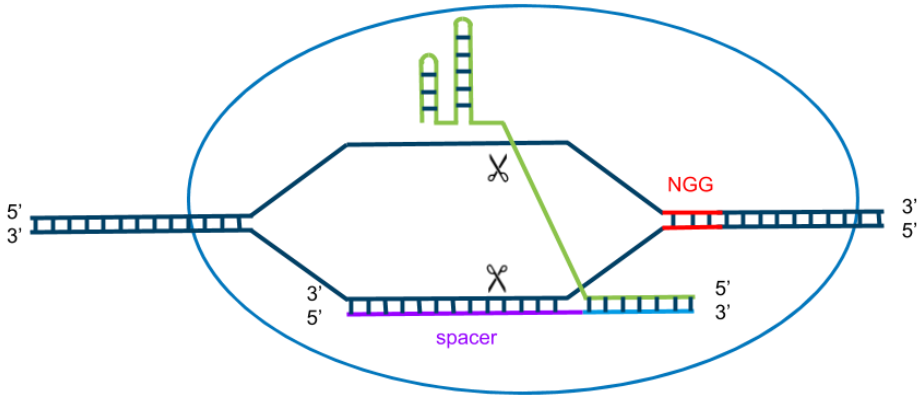


Figure 4. Elements of the CRISPR/Cas9 system. CRISPR/Cas9 protein (dark blue) searches the genome for the PAM (red NGG motif). If the spacer (purple, also called the guideRNA, gRNA, or sgRNA) and the genome sequence (protospacer) are complementary to each other, Cas9 will cleave 3-4 bp upstream of the PAM sequence using its RuvC domain for the non-complementary strand, and HNH domain for the complementary strand.

1.2.3. CRISPR effectors

CRISPR systems can be found in almost all archaea and in around 50% of bacteria (Hille et al. 2018). CRISPR/Cas systems are grouped into classes and further into types by: protein composition, effector complex structure, genome locus architecture, mechanisms of adaptation, pre-CRISPR crRNA processing and interference (Kira S. Makarova and Koonin 2013; Shmakov et al. 2015; K. S. Makarova, Wolf, and Koonin 2018). Effectors from Class 2 are the simplest and most popular system to use with genome engineering in mind. CRISPR/Cas9 (type II) and CRISPR/Cas12a (type V) come from Class 2 as they are characterized by a single large protein effector. What's more, many of the CRISPR/Cas types exist in multiple species, for example homologs with the same properties of CRISPR/Cas12a can be found in Prevotella, Francisella, Acidaminococcus or Lachnospiraceae.

During experimental design, scientists need to carefully select which CRISPR/Cas system to use, and from which species, as each species might have different PAM and gRNA requirements. For genome engineering applications, Cas protein effectors and gRNAs can be harvested and used outside of their adaptive immunity ecosystem. The most popular CRISPR effector currently used comes from Streptococcus pyogenes (**Figure 4**). Streptococcus pyogenes Cas9 is characterized by an NGG PAM and protospacer specificity towards 5'-20bp-NGG-3'. However, a plethora of other systems, with homologous Cas9 from other species exist, for example: Staphylococcus aureus defined by 5'-20bp-NNGRRT-3' (Friedland et al. 2015; Nishimasu et al. 2015), or Streptococcus thermophilus with 5'-20bp-NNAGAAW-3' (Garneau et al. 2010).

Scientists search for novel classes and types of CRISPR/Cas systems, hoping for alternative functionality or sequence context specificity. Each of those systems might have different requirements for the PAM motif, allowing targeting of the genomic loci not previously accessible. For instance, Cas12a (formerly named Cpf1) class systems harvested from Acidaminococcus and Lachnospiraceae have 5'-TTTN-23bp-3' as a PAM requirement, while creating an overhang cut - useful for homology-directed knock-in (**Figure 5)** (Zetsche et al. 2015). Additionally, targeting of RNA has also become possible with the recent discovery of Cas13a. The Cas13a system does not recognize a genomic PAM, but recognizes a PFS (protospacer flanking sequence) with the protospacer specificity defined as 5'-H-27bp-3'(Abudayyeh et al. 2017; Gootenberg et al. 2017).

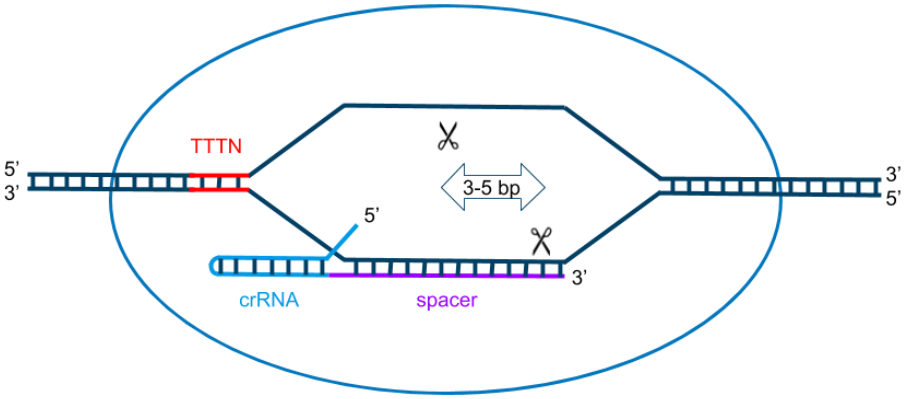


Figure 5. Elements of the CRISPR/Cas12 system. CRISPR/Cas12 (formerly Cpf1) recognizes a TTTN motif downstream of the protospacer and creates an overhang cut. This system does not require a tracrRNA (Zetsche et al. 2015).

1.2.4. Enzymatically dead Cas9

Cas9 proteins have also been engineered to be enzymatically inactive (dead Cas9), while preserving the target recognition mechanism. Dead Cas9 (dCas9) can be used for gene repression by binding to the promoter region and inhibiting the gene transcription machinery from starting transcription. Additionally, a nickase system (**Figure 6**) (Mali, Aach, et al. 2013) can be adopted where two Cas9 proteins are used, each only introducing a single-strand DNA break. In the nickase system each of the Cas9 proteins has one of the cutting domains inactivated, which creates single strand damage when the Cas9 proteins are not targeting in close proximity. Using two single nicking Cas9s restricts the number of potential off-targets significantly as both target:spacer complexes have to be bound and active in close local and temporal proximity (B. Shen et al. 2014). Nickases also result in a DSB with long terminal overhangs, a preferred type of damage for HDR repair and thus knock-in experiments.

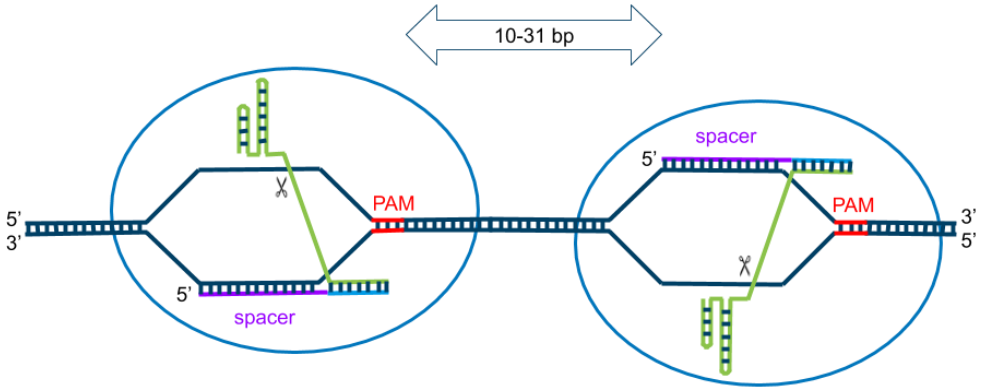


Figure 6. Elements of the CRISPR/Cas9 nickase system. Cas9 nickase is an approach where a mutated Cas9 can only create a cut on one of the two DNA strands. Consequently, two of those mutated Cas9 are necessary to create a much wider DSB, with an overhang resulting from the cut. This technique is much more resistant to off-targets, but yields lower editing efficiency.

1.2.5. CRISPR applications

The most basic utility of the CRISPR system is to guide Cas9 protein (and its homologs) to a genomic locus of interest, where it induces DSBs that are repaired (often) erroneously by the cell repair pathways. This can cause the introduction of a frameshift mutation that knocks out a gene.

A frameshift is a genetic mutation within the coding region of a gene caused by indels of a few nucleotides in a DNA sequence that is not divisible by three, which results in translation in a different frame to the original gene downstream of the mutation. If introduced towards the start of the gene the creation of a frameshift mutation is likely to render the gene non-functional and is therefore the desired outcome for gene knock-out experiments.

Introducing new sequence in the genome is also possible, which is the goal for knock-in techniques. Knock-in efficiency can be improved by overhangs (“sticky ends”) which

can be created by chance after a DSB is introduced by Cas9. The chance of creating overhangs can be increased with the use of nickases or Cas12a. The donor sequence - with arms complementary to the overhangs - is used by the HDR to repair the DSB site (Nami et al. 2018). Alternatively, some approaches use very long complementary arms together with the use of regular Cas9 (J.-P. Zhang et al. 2017).

With increased precision of genome targeting, it is now feasible to investigate the possibility of precision editing of the genome. Precision editing refers to the aim of editing a single nucleotide in the genome with high efficiency. This can for instance be achieved with the use of dead Cas9 fused with a cytidine deaminase that will alter C to U. This damage is recognized by cell-internal repair machinery and will be processed further to a T (Nishida et al. 2016; Komor et al. 2016). The field of precision editing using base editors is quickly developing: currently, it is also possible to create A to G conversions using Cas9 fused with adenosine deaminase (Gaudelli et al. 2017). In these methods, the recognition mechanism of Cas9 is unchanged, therefore the design of the gRNA for these experiments follows a similar path as for regular Cas9.

Additionally, the use of dCas9 or Cas9 nickases together with different fusion proteins, allows for guided genome effectors, for example: base editing (Komor et al. 2016; Gaudelli et al. 2017), epitranscriptome modifications (Pulecio et al. 2017), GFP tagging (Lackner et al. 2015) and many more. Based on these techniques, more specialized applications have emerged, for instance: lineage tracing (Spanjaard et al. 2018), Oxford Nanopore sequencing enrichment (Gabrieli et al. 2018), detection of Zika virus (Gootenberg et al. 2018), gene drives (Kyrou et al. 2018), and many more, including medical applications. The discovery of CRISPR systems has revolutionized biology, and in the upcoming years will transform our everyday life.

1.3. Design of CRISPR experiments

Given such broad applications of the CRISPR systems there is a need for a comprehensive design for genomic and transcriptomic targeting. There are three main

components to consider for the successful design of a gRNA: the location of interest, potential off-target sites, and the efficiency of editing.

1.3.1. Location

Before carrying out a CRISPR experiment, the scientist has to decide the target, which can be a selected gene, a promoter site, or even a whole chromosome with many gRNAs. Knowing the location of interest, these genomic regions can be scanned for the presence of a PAM (depending on the CRISPR system used), to detect whether it is possible to target these loci. Generally, for each case, additional rules apply. For instance, the choice of location and by proxy, the sequence of the spacer has an influence on efficient editing (Doench et al. 2014; H. Xu et al. 2015). This is explained more in depth at heading 1.3.2.

The choice of the precise locus to target depends on the experimental aim. For knock-out experiments, it is beneficial to target the protein coding potential of the gene close to the start codon. At the same time, there is a chance of initiation at a start codon downstream of the cut site, which can result in a truncated, but potentially still functional protein. It is also usually preferable to design gRNAs that target all transcript isoforms of the gene to ensure complete knock-out of the gene.

Knock-in experiments are highly dependent on the repair pathways and specific technique used. All knock-in methods involve preparation of microhomology arms of different lengths, which are context dependent (Nami et al. 2018). For approaches that use dCas9 for gene repression or activation, the region around the transcription start site is the preferred target, as CRISPR effectors can be used to block/unblock transcriptional machinery (Tanenbaum et al. 2014; Qi et al. 2013).

Furthermore, for RNA base editing and RNA knock-down with Cas13 it has been shown that targeting the loop regions of folded transcripts is more efficient (Abudayyeh et al. 2016). For each experimental purpose many factors have to be considered to decide on the perfect gRNA location.

1.3.2. Specificity (off-targets)

Off-targets are sites in the genome that are also targeted by the gRNA, but not intentionally. Finding all potential off-targets is important for all CRISPR experiments, as this ensures a predictable outcome of mutations in the cell. Methods such as: GUIDE-Seq (Tsai et al. 2015), GOTI (Zuo et al. 2019), DISCOVER-Seq (Wienert et al. 2019) - can detect where the CRISPR effectors have cut in the genome after the experiment, and are the most reliable methods for experimental detection of off-target sites. However, experimental detection of the off-target sites is laborious, expensive and rarely done in practice. As an alternative, computational prediction is used to minimize the number of potential off-targets.

Computational prediction can be performed by using sequence matching of the spacer to the genomic reference through alignment. It has been shown that even up to 6 mismatches between the genomic target site (protospacer) and the RNA spacer can be tolerated by Cas9, although at very low editing efficiencies (Tsai et al. 2015; Xiaoling Wang et al. 2015). Searching the genome for spacer sequences with 0-3 mismatches has been shown to capture the majority of potential off-target editing (Tsai et al. 2015; Cameron et al. 2017). Searching for more mismatches costs computational time that is leveraged against the potential gain of detecting more off-target sites (Cameron et al. 2017). A caveat with this however is that sites with many mismatches have a relatively small probability of being cut (Haeussler et al. 2016). A potential solution to this is present in a handful of off-target efficiency prediction algorithms that are able to score off-targets with the likelihood of a cut occurring at this given locus, taking into consideration the placement of mismatches on the off-target loci (Listgarten et al. 2018; Abadi et al. 2017). However, these algorithms carry the potential for false negative predictions, especially when off-target activity depends on the gRNA delivery method, cell type and duration of exposure to the effector protein (Cameron et al. 2017). Ideally, all genomic variation (e.g. single nucleotide polymorphisms, insertions) would be accounted for during the search for off-targets. This will become easier in the future with the widespread use of graph genomes for alignment, in combination with the use

of specialized aligners (Rakocevic et al. 2019). Another important aspect to consider comes from natural genomic variation. It is necessary to always include a control without CRISPR treatment for direct comparison in sequencing validation. When no controls are present, natural variation can be confused with successful targeted editing, or off-target activity.

1.3.3. Efficiency

An important feature of the CRISPR/Cas system is its overall high editing efficiency. However, the current level is not always sufficient for high precision interventions or experiments such as those needed in medical applications. Ensuring higher editing efficiencies with increased off-target fidelity than standard Cas9 with homolog/mutated effectors is therefore the focus of much active research (Moon et al. 2018; Kulcsár et al. 2017). Additionally, besides increasing efficiency through the design of new effectors, efficiency can also be computationally predicted using machine learning models allowing for the selection of highly efficient gRNAs (Doench et al. 2014).

The efficiency of CRISPR editing is influenced by many factors. State-of-the-art machine learning approaches combine locus-specific information to create more accurate predictions of efficiency. Studies have shown that important features for efficiency prediction are chromatin accessibility (Uusi-Mäkelä et al. 2018), GC content of the guide (Ren et al. 2014; T. Wang et al. 2014; Wilson, O'Brien, and Bauer 2018), thermodynamic stability (Doench et al. 2014; Horlbeck et al. 2016), sequence of the spacer and surrounding region (Doench et al. 2016; H. K. Kim et al. 2018) and self-complementarity (Thyme et al. 2016). Many of these studies provide machine learning models for predicting editing efficiency of the gRNAs.

These models generally considered to be less important than off-target prediction models. This is because in the case of false negative off-target predictions, scientists have no way of knowing that their experiments are influenced while a gRNA that is inefficient will be detected. Another reason to give these models less weight is that, a recent study showed that these models are likely overfitting to their own dataset, and

might not be as robust for different experimental setups as expected, such as the use of different cell types, promoters, or different species (Haeussler et al. 2016). Additionally, far from all Cas9 homologs have their respective efficiency models pre-trained and available. In these cases, computing simplified features (e.g. GC content, self-complementarity) might be the only possibility. Alternatively, assuming that a model trained on the close homolog will perform with similar robustness, is also possible.

In summary, maximizing efficiency of editing and minimizing off-target effects is a task for *in silico* algorithms that score gRNAs for their experimental use. The ideal software should account for all developments in the field and be continuously upgraded. The use of CRISPR in genome engineering has grown spectacularly, and the number of software tools for gRNA design is overwhelming. The most cited tools include: CHOPCHOP (Montague et al. 2014; Labun et al. 2016; Labun, Montague, et al. 2019), Cas-OFFinder (Bae, Park, and Kim 2014), CRISPR-P (Lei et al. 2014; H. Liu et al. 2017), E-CRISP (Heigwer, Kerr, and Boutros 2014), CRISPOR (Haeussler et al. 2016), CCTOP (Stemmer et al. 2017), and many more. Currently, there is a lack of comprehensive benchmarking and comparison of the tools to pinpoint which tools are good choices for each of the different experimental approaches, although some efforts are being directed there (Prykhozhij, Rajan, and Berman 2016; Bradford and Perrin 2018; Cui et al. 2018). In the future, when more data is available, the tools should undergo more considerable benchmarking. Meanwhile, the tools should evolve to further facilitate genome editing. There is still room for improvement by inclusion of new features and scientific insights in the field. Many software tools are published with minimum features and after some time become deprecated and eventually abandoned. My goal for the CHOPCHOP tool was to not follow this path, but relentlessly enhance user experience through constant updates.

1.4. Analysis of genome editing experiments

After designing and executing a CRISPR experiment, verification of the mutation is standard practice. Among other methods, highly precise identification of CRISPR edits can be achieved using targeted next-generation sequencing of amplicons (NGS)

(Sentmanat et al. 2018). NGS amplicon sequencing allows for hundreds or thousands of experiments to be run in parallel, thanks to barcode demultiplexing techniques. However, the use of NGS for CRISPR editing validation is costly, and therefore applied when it is beneficial to identify precise allelic changes or when the costs can be reduced by scaling the experiment. NGS amplicon sequencing allows scientists to see which exact bases were changed, therefore allowing them to establish heterogeneity profile of the edits for each target site. Calculating the efficiency of base editors, incorporation rate of HDR or the frameshift rate should be possible when using NGS with a good processing pipeline.

1.4.1. Editing efficiency estimation

Calculation of CRISPR editing efficiency for every locus is the most basic measure of experiment success. Nonetheless, there are multiple confounding factors in precision estimation of that value. Use of a control group, without CRISPR editing, is necessary to remove all sample-specific bias. Ignoring the control group can result in paper retraction (Schaefer et al. 2017, 2018). There can be differences between the genomic reference and the genome of the organism used for the experiment. Not accounting for this difference can confound results by confusing natural SNPs with CRISPR edits.

Another variable to consider is contaminant reads: reads that should not be considered when quantifying editing efficiency for a given locus (Lindsay et al. 2016). Contaminant reads might stem from high mosaicism, sequencing artifacts, formation of primer dimers, low quality reads, or erroneously assigned reads. For extremely precise editing efficiency estimation, sequencing noise (~ 0.1% for NGS) should also be taken into consideration. Additionally, in the case of paired-end read sequencing there can be biases connected to extracting the edited consensus from paired-end reads (Lindsay et al. 2016). Together, these confounding factors make the process of estimating editing efficiency more complicated than the matter would seem at first glance.

In summary, given the scale of the NGS experiments, as well as the complexity of the experimental problem, specialized computational tools are needed to facilitate the

analysis of genome editing experiments. Multiple tools exist for the analysis of CRISPR experiments that use amplicon sequencing data, and new ones are emerging. To name a few that have full pipeline analysis: CRISPRAnalyzeR (Winter et al. 2017), ampliCan (Labun, Guo, et al. 2019), CrispRVariants (Lindsay et al. 2016), CRISPResso (Pinello et al. 2016; Clement et al. 2019) and CRISPRMatch (You et al. 2018). The main differences between these tools come from data processing choices and visualization. As the nature of the editing efficiency estimation problem is more quantifiable than the design for CRISPR editing, considerable benchmarking can be performed on those tools (Lindsay et al. 2016). However, tools for precise estimation of editing efficiency have been shown to have significant room for improvement. Tools compared by Lindsay et al. 2016 lacked automatic normalization of the data and used aligners that are not aware of how CRISPR editing differs from normal read mapping. This benchmark also did not consider estimations of HDR efficiencies. More specialized tools will hopefully be developed in the future to match the developing field of CRISPR/Cas targeted genome editing.

2. Aim of the thesis

The potential of CRISPR has unleashed numerous new experimental approaches by use of precise genome targeting and engineering. Many scientists want to tap into what genome engineering has to offer, and use it to make a breakthrough in their own field. Current research on the CRISPR/Cas system is progressing at a frightening pace, unveiling new and unexpected applications, developing newer systems with interesting properties, as well as perfecting what is already known. Ideally, most scientists should not have to understand all of the intricacies of CRISPR to effectively use it. Computational tools come as an aid in this situation and nowadays are used for the design of CRISPR experiments, as well as the analysis of created mutants.

CHOPCHOP was one of the first tools for the design of gRNAs for CRISPR/Cas9, published in 2014. As the field progressed at a staggering pace, it became necessary to update the tool with the latest developments, especially, when the user base is constantly growing. The first goal of my PhD was to provide a continuous update of CHOPCHOP with the latest developments in the field of CRISPR, related to the gRNA design.

After the design of a gRNA - potentially using CHOPCHOP - and a successful laboratory application, the resulting data has to undergo bioinformatic analysis for editing validation. A high-throughput solution is to perform amplicon sequencing of the targeted locus. However, the experimenter needs to decide which tool to use for post-sequencing data analysis. The second aim of my PhD was to create a comprehensive benchmark of tools that analyze amplicon sequencing data from targeted genome editing experiments. The third and final aim of my PhD was to create a tool for precise estimation of editing efficiencies, that could outperform other benchmarked software.

3. Summary of Results and Discussion

Keeping CHOPCHOP up-to-date with CRISPR developments in the area of gRNA design is challenging as there are hundreds of papers published on this topic every year. The tool's user base continues to grow and therefore it contains a great need for future developments. On the other spectrum of CRISPR experiments, benchmarking data analysis tools that estimate editing efficiency of CRISPR experiments was possible thanks to simulated datasets (where true editing efficiency is known). Therefore, I directed my benchmarking efforts there. To fill the gap for highly precise estimation of efficiency editing, I created ampliCan.

Bullet points of achieved results and realized aims:

1. Updated of the CHOPCHOP tool to include latest developments in the CRISPR field.
    - 1.1. Inclusion of Cas12a (formerly Cpf1) and homolog effectors.
    - 1.2. Addition of user defined PAM sequence, and user defined gRNA length.
    - 1.3. Scoring of gRNAs for nickase targeting.
    - 1.4. Extension with the algorithms for gRNA efficiency prediction.
    - 1.5. Searching for off-targets with a less specific ruleset.
    - 1.6. Incorporation of isoform targeting with Cas13.
    - 1.7. Preparation of basic modes for standard applications for less advanced users: knock-in, knock-out, knock-down, Nanopore enrichment, gene activation and repression.
    - 1.8. Extension with the algorithms for gRNA repair profile prediction.
    - 1.9. Implementation of isoform level resolution (intersection/union modes) and selection.
    - 1.10. Visual display presentation on the website of in-frame start codons and all isoforms.
    - 1.11. Creation of batch mode (design for many genes in a streamlined fashion) and control guide creation (guides that have no targets on the genome of interest).

1.12. Preparation of queue to solve the congestion issue, arising due to growing number of users.

1.13. Maintenance of user oriented service with efficient bug fixes, user support with the addition of novel genomes.

2. Benchmarked tools for amplicon sequencing data analysis of genome editing.
   2.1. Reproduced previous benchmark from Lindsay et al. 2016.
   2.2. Characterized from where the differences between the leading tools arise when estimating true editing efficiency.
   2.3. Benchmarked how well leading tools can filter out contaminant reads.
   2.4. Determined how the type of editing event (deletion, insertion, mismatch, mixed) and its size is influencing error rates of the leading tools when estimating true editing efficiency.
   2.5. Performed separate evaluation for the estimation of the HDR editing efficiency of the leading tools.
   2.6. Discovered the depth of the precision that can be achieved when estimating the true editing efficiency on both real and simulated datasets.
3. Implemented a tool that can outperform other benchmarked tools, and also adheres to the following points.
   3.1. Tool includes automatic use of the control data.
   3.2. Tool uses specialized alignments, optimized for genome editing.
   3.3. Software is able to capture longer indels as the result of genome editing.
   3.4. The tool allows for extremely precise estimation of true editing efficiency.
   3.5. Final output of the tool is, among other formats, aggregate reports of the gRNA activity.

3.1. Updates of the CHOPCHOP tool

When performing genome editing with CRISPR, scientists need to select gRNAs that have the highest potential DNA cutting efficiencies, and minimize potential off-target sites. *In silico* tools are supposed to help users make these choices. CHOPCHOP was one of the first tools (available as a web server and a Python script) for the design of

gRNAs for CRISPR experiments (Montague et al. 2014). In late 2015, I became involved in the CHOPCHOP project as one of the maintainers.

With the attached paper Labun et al. 2016 (**Paper I, aims 1.1-1.5**), we extended CHOPCHOP with features related to the newly reported effectors. We implemented a 5' gRNA flanking type of PAM for applications with new effectors (e.g. Cas12) in addition to the 3' PAM used by Cas9. Additionally, we created a nickase mode, adjustable gRNA length, and allowance for specification of any user defined PAM to answer growing scientific interest with discoveries of novel CRISPR effectors (**Paper I**). That development of CHOPCHOP allowed users to use the tool with future discoveries of novel CRISPR effectors. For instance, gRNAs for the recently discovered CasX (J.-J. Liu et al. 2019) could be designed with CHOPCHOP since the day CasX was discovered.

CHOPCHOPs main difference in relation to other similar tools is that it is focused on integrating as much of the field knowledge as possible while maintaining flexibility of choice for more advanced users. CHOPCHOP was not created as just another efficiency scoring algorithm with a complimentary website and minimal set of features. On the contrary, CHOPCHOP incorporates published algorithms and currently supports 7 efficiency scoring models (Doench et al. 2014, 2016; Moreno-Mateos et al. 2015; Chari et al. 2015; H. Xu et al. 2015; H. K. Kim et al. 2018; T. Wang et al. 2014). What's more, this is the only tool that computes self-complementarity of the sgRNA as well as its complementarity to the backbone region that can hinder editing efficiency (Thyme et al. 2016).

GUIDE-Seq (Tsai et al. 2015) has shown that gRNAs can bind to off-target sites with up to 6 mismatches, but the majority of gRNAs that bind to potentially deleterious off-target sites have up to 3 mismatches. To account for these findings, CHOPCHOP uses bowtie alignment (Langmead et al. 2009) to find potential off-targets with up to 3 mismatches in the genome. This strategy balances computational time with sensitivity (**Paper I**).

CHOPCHOP ranks gRNAs by their off-targets, efficiency, GC content and self-complementarity to deliver a full list of potential spacers for a given locus. Every parameter of CHOPCHOP can be tuned for specific applications, but basic settings are also provided for users without detailed knowledge of the current developments in the field of CRISPR/Cas editing. User interaction was the main focus for the latest CHOPCHOP version (**Paper II, aims 1.6-1.13**). The aim of this publication was to increase use of CHOPCHOP by implementing optimized parameters and output for specific experimental approaches. CHOPCHOP now supports basic modes for gene knock-out, knock-in, repression/activation, nickases, Nanopore enrichment and knock-down RNA targeting with Cas13 (**aims 1.6, 1.7**). With the latest update, CHOPCHOP also integrates prediction of the repair profile, which is one of the latest major developments in the field (M. W. Shen et al. 2018).

After the recent update (**Paper II**), CHOPCHOP now displays all isoforms of the targeted gene together with all in-frame start codons (**aim 1.10**). Thus, CHOPCHOP is (to the author's knowledge) the only tool that supports isoform-aware gRNA design, which allows users to control targeting all isoforms of the gene of interest (**aim 1.9**). The visualization in CHOPCHOP promotes simplicity in gRNA choice, efficient primer design, as well as validation of editing outcome through restriction enzymes. Furthermore, CHOPCHOP has additional wrapper scripts (**aim 1.11**) that allow more advanced functionality: 1) batch mode - design and automatic selection of gRNAs for many genes at once; 2) control mode - design of gRNAs that have no putative targets on the genome. Since the tool’s creation, the CHOPCHOP maintainers have resolved countless requests from the growing user base. Implementation of the queue (**aim 1.12**) solved congestion issues on the web server. Meanwhile, constant additions and updates of genomes and their annotations accompanied both of the papers (Labun et al. 2016; Labun, Montague, et al. 2019).

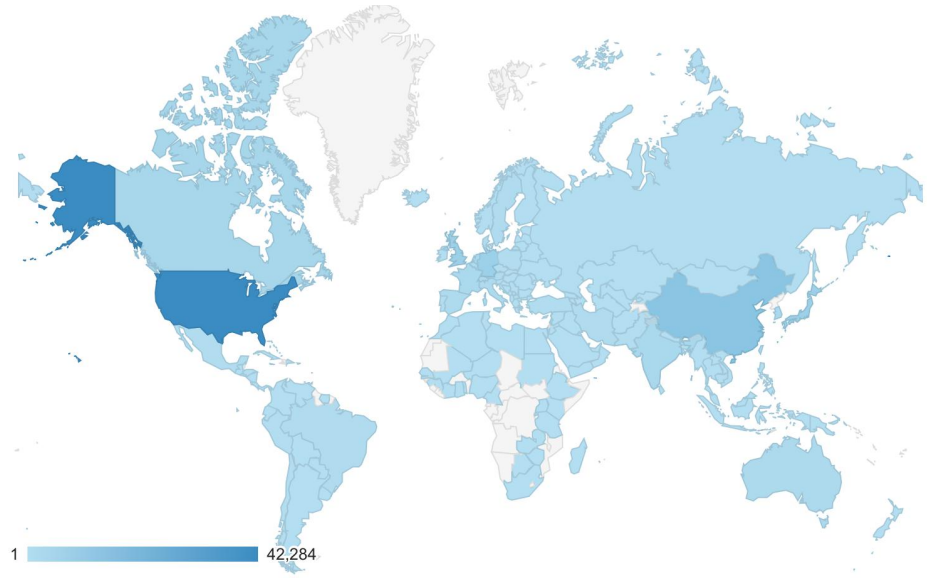


Figure 7. CHOPCHOP users come from all over the world. The image above was generated by Google analytics attached to the web server, and shows the number of unique IP addresses that accessed website between 01.09.2016 and 11.08.2019: in total, over 115,897.

These updates increased the functionality of CHOPCHOP as gRNA design tool. At the time of writing, after three major releases (Montague et al. 2014; Labun et al. 2016; Labun, Montague, et al. 2019), CHOPCHOP is one of the most cited tools with 460 citations for the first version and 245 for the second release. CHOPCHOP users come from all over the world, and there are hundreds of experiments being designed every day (**Figure 1**, **Figure 7**). To date, CHOPCHOP stands as one of the most versatile and most curated tools for the design of CRISPR editing experiments.

3.2. Analysis of CRISPR amplicon sequencing data

Pipelines that process CRISPR data from amplicon sequencing should - at the very minimum - be precise at estimating true editing efficiencies. Tools that estimate editing efficiency exist and Lindsay et al. 2016 have created an interesting approach to benchmark other tools with an artificial dataset. Synthetic data was simulated based on

distributions of real editing events. Benchmarking on real data is problematic, as the ground truth (true editing efficiency) is not known. However, synthetic datasets offer a ground truth and thus allow direct comparison of tools. I have replicated the benchmarking performed in Lindsay et al. 2016 (**Paper III, Supplemental Fig S7, aim 2.1**) to confirm their findings and include the tool that I have developed, ampliCan. The tools I compared were: ampliCan (Labun, Guo, et al. 2019), CrispRVariants (Lindsay et al. 2016), ampliconDIVider (Varshney et al. 2015), CRISPResso & CRISPResso Pooled (Pinello et al. 2016). The tools performed surprisingly unevenly on real datasets as well as on synthetic datasets (**Paper III, Supplemental Note S1**). Differences were proven to stem from processing choices: off-target detection, alignments and merging of the paired-end reads (**aim 2.2**). To highlight which method of data processing is the most robust, I simulated multiple other datasets (**Paper III, Supplemental Table S2**). With that, I established quantifiable metrics of how contaminant reads and different type of editing event can create problems with estimation of editing efficiency (**Paper III, Fig II, aims 2.3-2.4**). Since my benchmark, the developers of CRISPResso have published an updated version of their software (Clement et al. 2019). I have therefore recreated the benchmark for different types of reads with inclusion of the updated CRISPResso (**Figure 8**). CRISPResso v2 was run with docker technology and therefore there can be no mistake about improper installation of the software for the benchmark purposes. Issues with this tool are apparent in all benchmarks and it seems that the newer version has not yet addressed the pitfalls that caused the prior version to underperform. This example highlights the need for benchmarks in bioinformatics and for proper software testing.

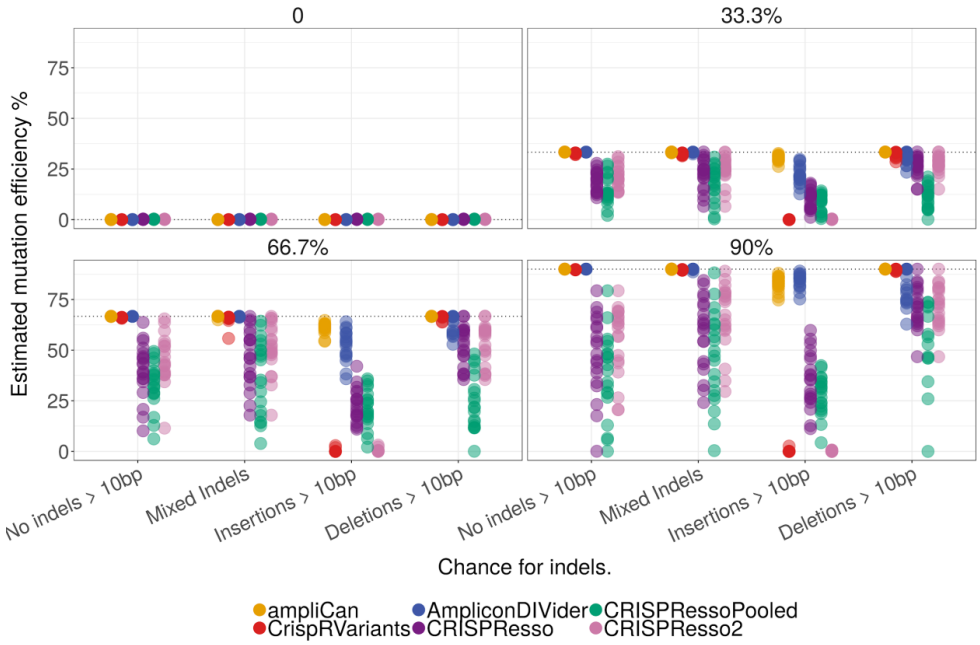


Figure 8. Benchmark of the leading tools on editing events simulated from real data. Dotted line represents true editing rate for each sample and each locus is represented by one colored dot. ampliCan is the most consistent tool at calling larger indels as well as smaller edits. This figure is an unpublished update (includes CRISPResso v2) to the Supplemental Fig S11, Labun et al. 2019.

Thanks to user feedback, ampliCan has also been extended with an HDR mode that is able to accommodate recognition and efficiency estimation of any intended editing, for example base editors and HDR insertions. Benchmarking estimation of HDR editing was a separate issue in which ampliCan was shown to outperform existing tools (**Paper III, Supplemental Note S6, aim 2.5**). Finally, I described the limit of the ampliCan approach in terms of precision. Estimating true editing efficiency in ampliCan is only restricted by the stochastic background noise of the sequencing procedure (~0.1%) as has been shown (**Paper III, Supplemental Table S1**), but even detection of events present in the frequency of 0.001% of NGS reads is possible (**aim 2.6**). With those considerable benchmarks I believe ampliCan has been proven as a comprehensive and robust tool.

It is challenging to derive true editing mutation efficiency without incorporation of the control data, especially in heterogeneous samples from many cells (**Paper III, Fig 1 B-C**). ampliCan, to the author's knowledge, was the first tool to include automatic normalization using control data. Furthermore, ampliCan does not merely subtract total editing efficiency of the control, which is common practice, but removes background events present in the control group from the treated group (**Paper III, Supplemental Note S2, aim 3.1**).

ampliCan offers ways to manipulate the analysis at a fine grained resolution of a singular read event considering all edit events: mismatch, deletion or insertion. It also provides a full pipeline with default settings for less advanced users. What differentiates ampliCan from other tools is event-level manipulation rather than estimating efficiencies at the read level. This methodology allows users to filter out some of the events from the treated group - for instance editing events found in the controls - instead of filtering out reads themselves. ampliCan features a completely new approach for filtering contaminant reads using clustering, robustly rejecting primer-dimers, and rejecting off-target reads (**Paper III, Supplemental Note S8, aim 2.3**). ampliCan alignments are optimized for CRISPR editing (**Paper III, Supplemental Note S3, aim 3.2**). Specialized alignments allow users to anticipate DSBs with the following repair, which allows them to also capture larger indels (**Paper III, Supplemental Note S5, aim 3.3**). ampliCan's data processing allows for precise estimations of editing efficiency as shown on multiple benchmarks (**Paper III, Fig 2, aim 3.4**). In addition, ampliCan prepares complete and editable reports for the user, not only basic summary metrics of the editing efficiency rates (**Paper III, Supplementary Note S7, aim 3.5**). A layer of plots and figures composing reports can be seamlessly generated from the pipeline, expediting the experiment review.

4. Paper I

CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering.

K Labun, TG Montague, JA Gagnon, SB Thyme, E Valen,

2016, Nucleic acids research 44 (W1), W272-W276

W272–W276 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* *Published online 16 May 2016*
doi: 10.1093/nar/gkw398

CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering

Kornel Labun[1,†], Tessa G. Montague[2,†], James A. Gagnon[2], Summer B. Thyme[2] and Eivind Valen[1,3,*]

[1]Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway, [2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA and [3]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway

Received February 10, 2016; Revised April 8, 2016; Accepted April 25, 2016

ABSTRACT

In just 3 years CRISPR genome editing has transformed biology, and its popularity and potency continue to grow. New CRISPR effectors and rules for locating optimum targets continue to be reported, highlighting the need for computational CRISPR targeting tools to compile these rules and facilitate target selection and design. CHOPCHOP is one of the most widely used web tools for CRISPR- and TALEN-based genome editing. Its overarching principle is to provide an intuitive and powerful tool that can serve both novice and experienced users. In this major update we introduce tools for the next generation of CRISPR advances, including Cpf1 and Cas9 nickases. We support a number of new features that improve the targeting power, usability and efficiency of CHOPCHOP. To increase targeting range and specificity we provide support for custom length sgRNAs, and we evaluate the sequence composition of the whole sgRNA and its surrounding region using models compiled from multiple large-scale studies. These and other new features, coupled with an updated interface for increased usability and support for a continually growing list of organisms, maintain CHOPCHOP as one of the leading tools for CRISPR genome editing. CHOPCHOP v2 can be found at http://chopchop.cbu.uib.no

INTRODUCTION

The discovery and adoption of the CRISPR bacterial system for genome editing has led to a revolution in biology: targeted mutations are now possible in a multitude of organisms, including many not previously amenable to genetic manipulation. This has both transformed our approach to answering biological questions and unlocked the possibility of correcting human genetic diseases.

Originally harnessed from the *Streptococcus pyogenes* type II system (1–3), CRISPR genome editing is based on a two-component system: a Cas9 nuclease and a single guide RNA (sgRNA), which directs the nuclease to a specific site in the genome. In the presence of the sgRNA, Cas9 locates the target site and makes a double-strand break (DSB). The DSB is repaired by the host non-homologous end-joining pathway, but often the repair is imperfect, creating indels and in many cases frameshift mutations. Since the technology's inception, research to improve the technology has focused on two main challenges: optimization of cutting efficiency and specificity of cutting. A substantial portion of sgRNAs designed for a given gene will produce a low or zero cutting rate, and many sgRNAs have the capacity to bind promiscuously in the genome, which can lead to off-target mutagenesis (4–10). To address these issues, research has focused on identifying the sequence features that contribute to effective (and ineffective) sgRNAs (11–16), as well as the development of new CRISPR variants that expand the targeting range and specificity of the nuclease (17–20). With the contribution of so many factors to optimum sgRNA target selection, it has become necessary to use software to aid selection of CRISPR target sites for experiments. CHOPCHOP (21) provides an intuitive online environment for target selection that optimizes efficiency and specificity according to the latest large-scale studies, as well as performing primer design and restriction site identification, all in a user-friendly, graphical interface (Figure 1). This new update of CHOPCHOP provides additional flexibility by offering new options for sgRNA design, as well as additional metrics by which sgRNA targets are scored and ranked.

IMPROVEMENTS IN THE 2016 RELEASE

CHOPCHOP accepts multiple input formats (gene identifiers, genomic coordinates and pasted sequences) for a wide range of organisms, and provides instant, visual out-

*To whom correspondence should be addressed. Tel: +47 55 584 074; Fax: +47 55 58 41 99; Email: eivind.valen@gmail.com
†These authors contributed equally to the paper as first authors.

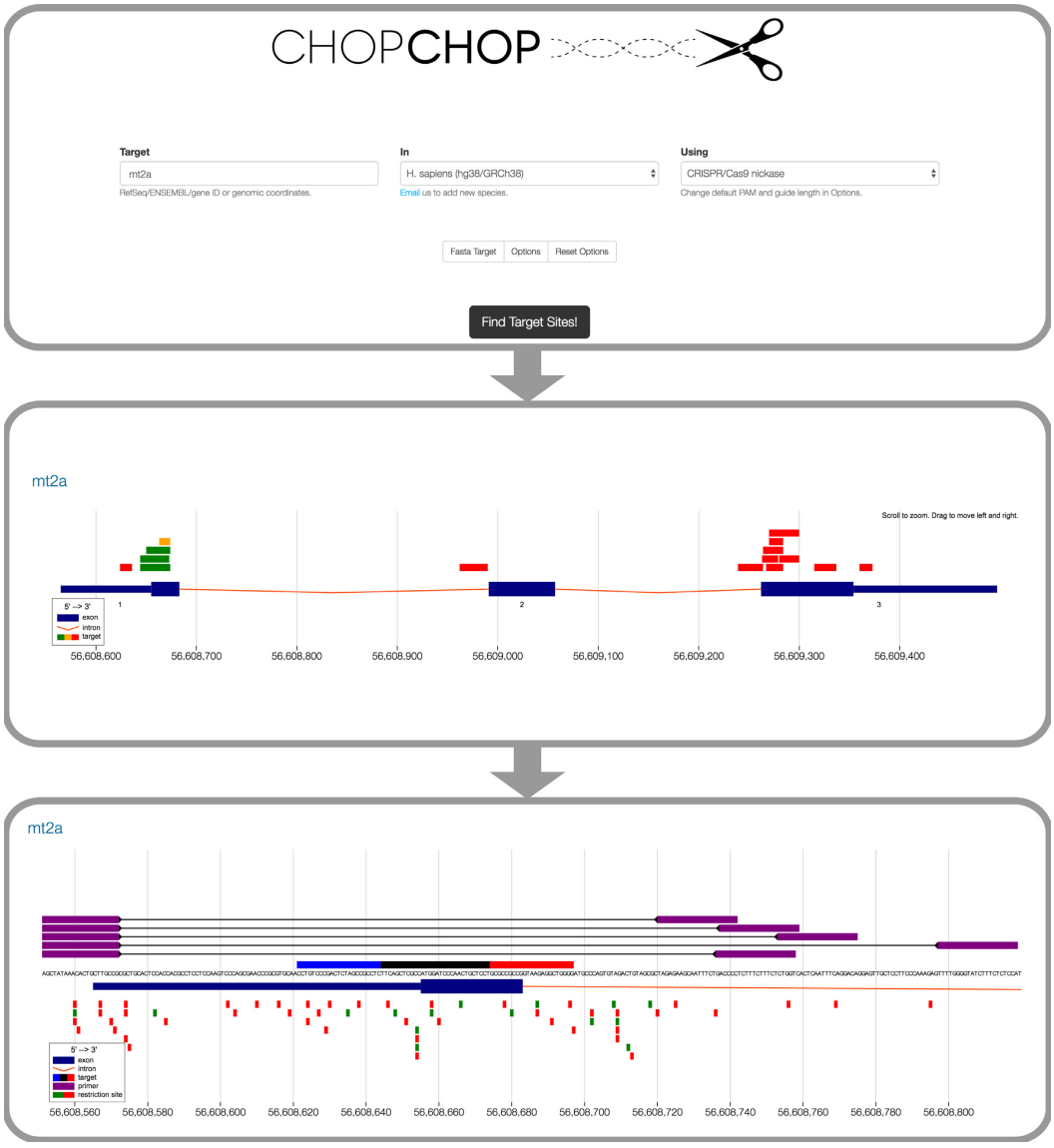


Figure 1. The workflow of CHOPCHOP in Cas9 nickase mode. The CHOPCHOP homepage (upper box) allows three types of input (DNA sequence, genomic coordinates or gene IDs) with default parameters optimized for novice users. For experienced users, a number of options for Cas9, Cas9 nickase, Cpf1 and TALEN mode can be revealed. The results of the search (middle box) are displayed across the gene, genomic region or DNA sequence, depending on the input format. The target color indicates the quality of each sgRNA or nickase pair (green [best] to red [worst]). The graphic representation of the search area is complemented by an interactive table below (not shown). Each sgRNA or nickase pair can be explored in greater detail (lower box) with annotated primer candidates and restriction sites, and information about any off-targets (not shown). Nickases are displayed in red and blue with the intermediate region in black.

put as well as downloadable data (GenBank, text tables and FASTA files). In this new version users can also view the output data in the UCSC browser (22) with a single click, enabling results to be viewed in the context of annotated genomic features, such as transcription factor binding sites and chromatin architecture and accessibility (Figure 2).

CHOPCHOP offers flexible targeting to sub-regions of protein-coding and non-coding genes, including coding regions, UTRs, splice sites and individual exons. In this new version we also offer a promoter-targeting mode (Figure 2) for experiments such as down- or upregulating gene expression using catalytically dead Cas9 (dCas9) or transcriptionally active dCas9 (e.g. dCas9-VP64), respectively (23–25). CHOPCHOP determines potential off-target sites for all sgRNAs using Bowtie (26) and automatically generates primers for target sites using Primer3 (27). The length and annealing temperature of the primers, as well as the size of the amplicon, can be specified. CHOPCHOP visualizes all elements in a dynamic visual interface that includes information about restriction sites, which can be used for downstream validation.

In addition to these improvements, the new iteration of CHOPCHOP introduces the following major new features.

Support for a new generation of CRISPR effectors

The most widely used CRISPR effector is Cas9, derived from the type II *S. pyogenes* system. While the RNA-mediated targeting of Cas9 offers great versatility in selecting a target site, a limiting factor is the requirement for an NGG protospacer adjacent motif (PAM) motif adjacent to the target. The occurrence of this motif is not rare in most genomes, but it imposes a restriction that can be inimical to achieving the high genomic precision required for certain experiments, or for targeting small genes. The new generation of CRISPR effectors vastly expands the universe of viable targets by offering alternative PAM motifs (Supplementary Table S1, Supplementary Figures S1 and 2). CHOPCHOP now provides support for alternative CRISPR effectors, including Cpf1 from *Acidaminococcus*, which utilizes an AT-rich PAM (17) and Cas9 homologs from *S. pyogenes, Streptococcus thermophilus*, *Staphylococcus aureus* and *Neisseria meningitidis* (28). In addition, CHOPCHOP also accepts user-defined custom PAMs that can be anchored to the 5′ (Cpf1) or 3′ (Cas9) end of the sgRNA. This field accepts the standard IUPAC nucleotide alphabet (29), including ambiguity codes. CHOPCHOP therefore provides support for the sequence requirements of any currently known CRISPR effector and enables immediate adoption of any new CRISPR effectors. This greatly increases the targeting range of CRISPR experiments that can be designed with CHOPCHOP, including improved targeting of AT-rich genomes such as *Plasmodium falciparum* (Supplementary Figure S2).

New rules for optimizing cutting efficiency

CRISPR sgRNAs can be ranked by 2 criteria: (i) efficiency—the likelihood that the particular sgRNA facilitates cutting, and (ii) specificity—the likelihood that the sgRNA binds off-target sites.

The initial release of CHOPCHOP provided two simple metrics for efficiency based on experimental studies. First, the GC-content of the sgRNA—ideally between 40 and 80%—and second, whether the sgRNA contains a G at position 20 (11,30). Since the initial release of CHOPCHOP, several refinements have been proposed. A study from Doench *et al*. produced a large dataset to calculate efficiencies across a wide range of sgRNAs (14), and the rules for computationally-aided sgRNA design were recently further refined by the same group (13). Moreno-Mateos *et al*. conducted similar screens and found that sgRNA stability, which depends on guanine enrichment and adenine depletion, was a major determinant of sgRNA efficiency (12). Chari *et al*. conducted a study exploiting the bias of lentiviral integration into transcriptionally active regions, which: (i) revealed that accessible DNA is more amenable to cutting with Cas9; (ii) separated the influence of DNA accessibility and sequence composition on sgRNA efficiency. CHOPCHOP users can now view results in the UCSC browser (22) in the context of DNase I hypersensitivity sites to predict accessible DNA regions (Figure 2). Finally, a meta study by Xu *et al*. compiled the sequence specificities across multiple datasets to build an aggregate model (15). We have implemented all of these metrics in the new release to give the user a broad selection of metrics to choose from (the default is the Xu *et al*. metric). Using these methods, CHOPCHOP can now score every sgRNA using position-specific scoring matrices or support vector machines that consider each individual position of the sgRNA as well as the sequence downstream of the PAM and upstream of the binding site. In the results table this score is reported as the 'efficiency score'.

Other factors also play a role in whether an sgRNA is likely to cut at its intended target. Recently, we and others showed that self-complementarity of the sgRNA can inhibit its efficient incorporation into the effector complex (12,31). CHOPCHOP now includes the basic self-complementarity score of the Thyme *et al*. study (31), which computes the number of potential 4 bp stems within the sgRNA and between the sgRNA and the backbone. The user can therefore opt to avoid sgRNAs with self-complementarity using this option.

Strategies to increase specificity

A significant challenge in CRISPR experiments is the possibility of inducing cleavage at sites other than the intended target. An emerging tool to alleviate this problem is the paired nickase approach (32). Unlike natural CRISPR effectors, nickases have been modified to cut only one DNA strand. In order to create a DSB, a pair of nickases must be targeted to opposite strands and bind within 10–31 bp of each other (32). These requirements vastly reduce the likelihood of creating off-target DSBs, and CHOPCHOP has now added support for paired nickase experiments. In this mode, sites on opposite strands within a specified distance (either default or user-defined) are paired as potential nickase sites. For these sites, in addition to the default off-target search, each pair of sites is evaluated for off-targets where binding and cutting would result in a DSB. Nickase sites are

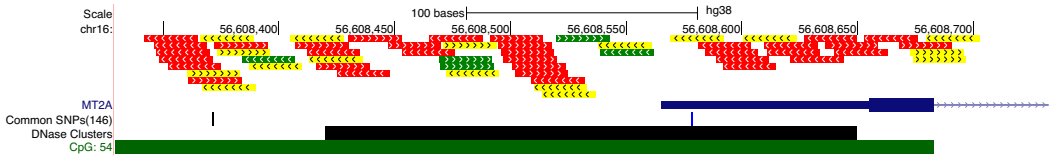


Figure 2. CHOPCHOP results can be exported to the UCSC browser with a single click. Here, the sgRNAs (in this example in promoter-targeting mode) are viewed in the context of the genome. The tracks displayed in this example are DNase sensitive regions, common SNPs and CpG islands.

visualized with two CRISPR targets surrounding a 'break' region (Figure 1).

Recent studies have highlighted the need to search for more than two mismatches when identifying off-targets (10) so CHOPCHOP now counts off-targets with up to three mismatches. While off-targets with more than three mismatches have been reported (10), evidence suggests that almost all predicted sites of four mismatches or more are not cleaved (10) and therefore the vast majority of such predicted sites would be misleading and unnecessarily time-consuming to search for during sgRNA selection.

Another strategy that has been shown to decrease off-target cleavage is the use of truncated sgRNAs (10,20). Besides increasing specificity, 5′ shortening of the customary 20 bp also increases the targeting range. The new version of CHOPCHOP therefore provides support for sgRNAs of user-defined lengths.

Thus, this version of CHOPCHOP supports a number of new features that: (i) improve the ability to target a broader range of sequences, and (ii) more thoroughly predict potential off-target sites in the genome. For an example of the increased targeting range and additions to the scoring system between the old and new versions of CHOPCHOP, see Supplementary Figure S3 and Table S2.

New genomes

In addition to a new range of features, CHOPCHOP strives to accommodate all requests for new genomes and gene annotation sets. So far we have incorporated all inquiries received, and CHOPCHOP now supports a total of 32 organisms. Furthermore, all genomes have been updated to their most recent assemblies and suggestions for new species can easily be submitted through a link on the main page.

DISCUSSION AND FUTURE DEVELOPMENTS

The overarching principle of CHOPCHOP is to provide an intuitive and powerful tool that can serve first time as well as experienced users. The basic mode offers optimized defaults for the basic user, while more advanced users can select from a wide range of options curated from the literature by their relevance and utility. All options are presented in a tabulated and organized manner to help users quickly visualize and evaluate options when designing CRISPR experiments.

This release retains the general layout of the previous release, but updates the visual profile to a modern look and to accommodate new features. The site is now mobile and tablet friendly, and to streamline the user's experience we use cookies to remember the selection of species and targeting options for subsequent searches. All reported bugs have been fixed, and the implementation is now optimized for future development to facilitate both rapid adoption of any future effectors and new targeting data from large-scale studies. This major update maintains CHOPCHOP as one of the most easy-to-use, versatile and powerful CRISPR targeting tools available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Alexander F. Schier and Michele Clamp for support and the many users who provided helpful suggestions for improving CHOPCHOP and this manuscript, in particular: Etsuko Moriyama, William C. Cheng, Miguel A. Moreno-Mateos and Antonio Giraldez. We are particularly grateful to Maximilian Haeussler for facilitating the integration of CHOPCHOP results into the UCSC Genome Browser.

FUNDING

Bergen Research Foundation (to E.V.); University of Bergen core funding (to K.L.); National Defense Science and Engineering Graduate Fellowship (to T.G.M.); American Cancer Society (to J.A.G.). Funding for open access charge: Bergen Research Foundation.
Conflict of interest statement. None declared.

REFERENCES

1. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
2. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
3. Mali,P., Aach,J., Stranges,P.B., Esvelt,K.M., Moosburner,M., Kosuri,S., Yang,L. and Church,G.M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–838.
4. Fu,Y., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,J.K. and Sander,J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.
5. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
6. Pattanayak,V., Lin,S., Guilinger,J.P., Ma,E., Doudna,J.A. and Liu,D.R. (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.*, **31**, 839–843.

7. Wang,X., Wang,Y., Wu,X., Wang,J., Wang,Y., Qiu,Z., Chang,T., Huang,H., Lin,R.-J. and Yee,J.-K. (2015) Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.*, **33**, 175–178.
8. Cradick,T.J., Fine,E.J., Antico,C.J. and Bao,G. (2013) CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.*, **41**, 9584–9592.
9. Kim,D., Bae,S., Park,J., Kim,E., Kim,S., Yu,H.R., Hwang,J., Kim,J.-I. and Kim,J.-S. (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods*, **12**, 237–243.
10. Tsai,S.Q., Zheng,Z., Nguyen,N.T., Liebers,M., Topkar,V.V., Thapar,V., Wyvekens,N., Khayter,C., Iafrate,A.J., Le,L.P. *et al.* (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.*, **33**, 187–197.
11. Gagnon,J.A., Valen,E., Thyme,S.B., Huang,P., Ahkmetova,L., Pauli,A., Montague,T.G., Zimmerman,S., Richter,C. and Schier,A.F. (2014) Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One*, **9**, e98186.
12. Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.-D., Fernandez,J.P., Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods*, **12**, 982–988.
13. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.
14. Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,I., Sullender,M., Ebert,B.L., Xavier,R.J. and Root,D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.
15. Xu,H., Xiao,T., Chen,C.-H., Li,W., Meyer,C.A., Wu,Q., Wu,D., Cong,L., Zhang,F., Liu,J.S. *et al.* (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res.*, **25**, 1147–1157.
16. Chari,R., Mali,P., Moosburner,M. and Church,G. M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods*, **12**, 823–826.
17. Zetsche,B., Gootenberg,J.S., Abudayyeh,O.O., Slaymaker,I.M., Makarova,K.S., Essletzbichler,P., Volz,S.E., Joung,J., van der Oost,J., Regev,A. *et al.* (2015) Cpf1 is a single RNA-Guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, **163**, 759–771.
18. Kleinstiver,B.P., Pattanayak,V., Prew,M.S., Tsai,S.Q., Nguyen,N.T., Zheng,Z. and Keith Joung,J. (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**, 490–495.
19. Kleinstiver,B.P., Prew,M.S., Tsai,S.Q., Topkar,V.V., Nguyen,N.T., Zheng,Z., Gonzales,A.P.W., Li,Z., Peterson,R.T., Yeh,J.-R.J. *et al.* (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, **523**, 481–485.
20. Fu,Y., Sander,J.D., Reyon,D., Cascio,V.M. and Joung,J.K. (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.*, **32**, 279–284.
21. Montague,T.G., Cruz,J.M., Gagnon,J.A., Church,G.M. and Valen,E. (2014) CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.*, **42**, W401–W407.
22. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
23. Bikard,D., Jiang,W., Samai,P., Hochschild,A., Zhang,F. and Marraffini,L.A. (2013) Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.*, **41**, 7429–7437.
24. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-Guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.
25. Hilton,I.B., D'Ippolito,A.M., Vockley,C.M., Thakore,P.I., Crawford,G.E., Reddy,T.E. and Gersbach,C.A. (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.*, **33**, 510–517.
26. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
27. Untergasser,A., Cutcutache,I., Koressaar,T., Ye,J., Faircloth,B.C., Remm,M. and Rozen,S.G. (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.
28. Fonfara,I., Le Rhun,A., Chylinski,K., Makarova,K.S., Lécrivain,A.-L., Bzdrenga,J., Koonin,E.V. and Charpentier,E. (2014) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.*, **42**, 2577–2590.
29. Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, **13**, 3021–3030.
30. Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.
31. Thyme,S.B., Akhmetova,L., Montague,T.G., Valen,E. and Schier,A.F. (2016) Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.*, **7**, 11750.
32. Ran,F.A., Hsu,P.D., Lin,C.-Y., Gootenberg,J.S., Konermann,S., Trevino,A.E., Scott,D.A., Inoue,A., Matoba,S., Zhang,Y. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.

5. Paper II

CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing.
K Labun, TG Montague, M Krause, Y Torres Cleuren, H Tjeldnes, E Valen,
2019, Nucleic Acids Research, gkz365

Published online 20 May 2019 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* ***W171–W174***
doi: 10.1093/nar/gkz365

CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing

Kornel Labun[1], Tessa G. Montague[2], Maximilian Krause[1], Yamila N. Torres Cleuren[1], Håkon Tjeldnes[1] and Eivind Valen[1,3,*]

[1]Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway, [2]Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Neuroscience, Columbia University, New York, NY 10027, USA and [3]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway

Received February 26, 2019; Revised April 16, 2019; Editorial Decision April 26, 2019; Accepted May 06, 2019

ABSTRACT

The CRISPR–Cas system is a powerful genome editing tool that functions in a diverse array of organisms and cell types. The technology was initially developed to induce targeted mutations in DNA, but CRISPR–Cas has now been adapted to target nucleic acids for a range of purposes. CHOPCHOP is a web tool for identifying CRISPR–Cas single guide RNA (sgRNA) targets. In this major update of CHOPCHOP, we expand our toolbox beyond knockouts. We introduce functionality for targeting RNA with Cas13, which includes support for alternative transcript isoforms and RNA accessibility predictions. We incorporate new DNA targeting modes, including CRISPR activation/repression, targeted enrichment of loci for long-read sequencing, and prediction of Cas9 repair outcomes. Finally, we expand our results page visualization to reveal alternative isoforms and downstream ATG sites, which will aid users in avoiding the expression of truncated proteins. The CHOPCHOP web tool now supports over 200 genomes and we have released a command-line script for running larger jobs and handling unsupported genomes. CHOPCHOP v3 can be found at https://chopchop.cbu.uib.no

INTRODUCTION

The use of CRISPR–Cas is now ubiquitous in modern molecular biology. First introduced as a tool for introducing repair-induced mutations in the genome, the emergence of catalytically dead or fused versions of the effector proteins has transformed CRISPR–Cas into a general purpose tool for targeting. For instance, CRISPR–Cas has been used to introduce new sequences into the genome (1–3), activate (4,5) or repress (6) transcription, as an enrichment tool for sequencing (7,8), for targeted hypermutation (9), as a diagnostic tool (10), to perform whole-organism lineage tracing (11), to target RNA molecules for destruction (12) or editing (13), and to track transcripts in live cells (14).

All CRISPR–Cas applications use a sgRNA to direct the CRISPR effector (Cas) protein to its target. In theory, CRISPR–Cas targeting only requires complementarity between the sgRNA and its nucleic acid target, but a number of studies have shown that efficient targeting follows more complex rules (15–21). For instance, the position of specific nucleotides in the target sequence, the accessibility of the target site, and the sequence of its flanking regions can all influence efficiency. The targeting efficiencies of Cas9 and Cas12a/Cpf1 have been measured in large-scale studies and combined with machine learning-based methods to optimize cutting (15,21).

There are other factors that can influence or prevent the generation of a null mutant. For instance, introducing a frameshift mutation too close to the start codon can permit translation initiation at a downstream ATG, leading to unintentional protein production. In other cases, targeting exons that are only present in a subset of isoforms can prevent null mutation generation. Finally, CRISPR–Cas gene editing can produce confounding phenotypes due to transcription adaptation or genetic compensation. For instance, a recent study showed that degradation of mutant mRNAs could result in the upregulation of related genes (22). In these situations, deleting the promoter can be a more robust method to produce knockouts.

Many CRISPR–Cas applications require the generation of a frameshift mutation to disrupt gene function, which requires a DNA repair event in which the number of inserted or deleted nucleotides is not a multiple of three. Surprisingly, it has been shown that double-strand break (DSB) repairs are not random: Cas9-induced DSBs using the same sgRNA often give rise to the same mutations (23). Recently, this has been incorporated into models that predict whether DSB repairs will give rise to a frameshifting mutation (24).

*To whom correspondence should be addressed. Tel: +47 55584074; Email: eivind.valen@gmail.com

In summary, the CRISPR–Cas system has been adapted for a wide selection of uses, and numerous factors influence each of these modes, necessitating the existence of intuitive software for target selection. This new update of the CHOPCHOP web tool incorporates new CRISPR–Cas targeting modes and predicts frameshift mutation frequency in an improved, user-friendly interface.

IMPROVEMENTS IN THE NEW RELEASE

Like previous versions, CHOPCHOP handles input from (i) gene and transcript identifiers, (ii) genomic coordinates and (iii) pasted sequences, and provides results in a number of output formats. The interface is simple (Figure 1) and requires the user to make four selections in the default mode: (a) target gene/isoform/region, (b) organism, (c) CRISPR effector (e.g. Cas9, CasX or Cas13), and (d) purpose (e.g. knockout, knockdown, repression). Advanced users can adjust the default settings by clicking the 'Options' button (Figure 1).

While CHOPCHOP queries typically run within a few seconds, heavy traffic can cause congestion during peak hours. We have therefore introduced a queuing system that ensures all users are prioritized. The new system retains results and calculations for up to 48 h after the initial query, permitting quick access (via caching) if an identical search is made later on and sharing results with collaborators. To increase speed for larger queries, CHOPCHOP also supports pre-filtering of sgRNA targets by (i) GC content (with a default of 10–90%) and (ii) the existence of self-complementarity within the sgRNA. This can greatly reduce computation time.

Several adjustments have been made to improve the results page visualizations. Notably, in-frame downstream ATGs are now colored in the isoform visualization (Figure 1) to help users avoid downstream translation initiation. In addition, the results table can now be sorted using any criteria.

In addition to these improvements, the latest CHOPCHOP introduces the following new major features.

Targeting the transcriptome

The most widespread use of CRISPR–Cas is to introduce mutations into DNA. However, CRISPR–Cas systems have now been engineered to target RNA. For instance, CRISPR–Cas13 has been harnessed for transcript knockdown (25), live-cell transcript imaging (14) and RNA base editing (13).

CHOPCHOP now permits CRISPR–Cas13 targeting, and implements this functionality by searching for off-targets across the complete transcriptome rather than the genome. An important aspect of RNA targeting is to avoid regions of high structure that can reduce the accessibility of Cas13 (12,14,26). CHOPCHOP calculates RNA accessibility using RNAfold from the ViennaRNA package (27) according to published recommendations (14). Briefly, accessibility is calculated in windows of 70 nucleotides, obtaining the probability that a given nucleotide position in the transcript is unpaired. For each target, we take the mean probability of structure across each position targeted by the sgRNA (14).

Similar to protospacer adjacent motifs (PAMs) in DNA-targeting modes, we support any 5′ or 3′ protospacer flanking sequences (PFS) for RNA targeting. The Cas13a PFS - an H at the 3′ end of the target - is the default. As with the PAM, the PFS should be present in the sequence of the DNA or RNA target, but not in the sgRNA. CHOPCHOP provides the appropriate sgRNA sequence, ensuring that the user does not include the PFS when ordering the sgRNA oligonucleotide.

New modes for targeting the genome

CHOPCHOP v3 expands the number of modes for DNA targeting. These include: (i) Nanopore enrichment mode. Targeted sequencing is a method used to attain high quality sequencing reads in a specific region of interest. PCR-based methods for enriching genomic regions have some limitations, for instance the maximum length of the region that can be enriched. By contrast, CRISPR–Cas provides a powerful method to excise genomic regions prior to amplification and sequencing with Oxford Nanopore Technologies (ONT) (7,8). The nanopore enrichment mode in CHOPCHOP allows users to identify pairs of gRNAs flanking large regions (up to 40kb) by excluding low-efficiency guides. The mode filters all sgRNAs with predicted self-complementarity, as this can have inhibitory effects on global Cas9 activity (8,28 and ONT personal communication); (ii) Knock-in mode. This identifies the same sgRNAs as the knockout mode, but designs homology arms up to 2 kb, which, based on recent studies (1–3) can be used for targeted insertions; (iii) Activation/repression modes. These modes are designed for use with Cas9 fusion proteins with the intention of activating or repressing a gene. Specifically, the modes target the promoter region and its flanking sites according to the guidelines specified in (29–31) in order to bring the activating/repressing domain into close proximity with the transcription start site.

For the Cas9 knockout mode, we now create a prediction of DSB repair outcomes (24). The model estimates the probability that a given sgRNA will result in a frameshift mutation. In addition, we have added efficiency scores for Cas12a/Cpf1 (21) and updated the 'Doench' efficiency score to the newest version (15), which is now the default scoring metric for Cas9 genomic targeting.

Expansion of genomes and targeting

CHOPCHOP now supports over 200 genomes and includes gene annotations for genomic targets, as well as three transcriptomes for RNA knockdown (human, mouse and zebrafish). While previous versions of CHOPCHOP required the selection of a specific isoform for targeting, this new version allows the user to target the entire gene. In its default 'intersection' mode, CHOPCHOP v3 only searches for sgRNAs present in every isoform (Figure 1). This mode can be disabled by selecting the 'union' mode, which will display all sgRNAs in all transcripts, as well as a column indicating whether the sgRNA targets a constitutive exon.

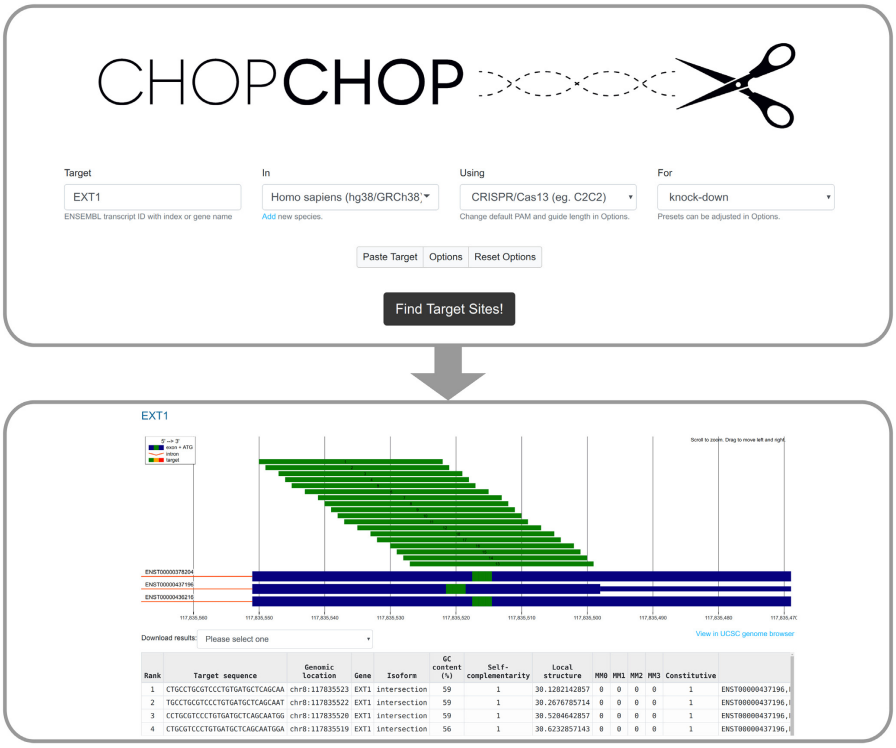


Figure 1. The workflow of CHOPCHOP when targeting RNA for knock-down. The CHOPCHOP homepage (upper box) require four types of input: (i) target, (ii) species, (iii) CRISPR effector and (iv) the purpose of the experiment. Default options will be adequate for most users, but advanced options can be revealed using the 'Options' button. The results of the search (lower box) are displayed along with all isoforms of the target gene. The target color indicates the quality of each sgRNA or nickase pair (green [best] to red [worst]). Below the graphic representation an interactive table allows for exploring each guide in greater detail.

Command-line version

In addition to the web interface, we also provide the code for the command-line version of CHOPCHOP, which can be run locally and is suited for larger queries or screens. This tool includes all of the functionality of the web interface in addition to extra functionality for larger experiments, such as the ability to design control sgRNAs that do not match any sequence in the genome. The command-line version of CHOPCHOP is compatible with ampliCan, a tool for sequencing-based assessment of mutations (32).

DISCUSSION AND FUTURE DEVELOPMENTS

Since its inception, the CRISPR field has undergone constant and rapid innovation, requiring the parallel development of bioinformatic tools that accommodate the new findings and technologies. In just a few years, CRISPR–Cas has become a powerful targeting tool for silencing and activating both DNA and RNA in a range of contexts, each of which requires the application of specific rules. The latest release of CHOPCHOP addresses this challenge by adding new functionalities that reflect the ever-expanding CRISPR toolbox. So far, we have accommodated the favorite species of over 200 research groups, and as we continue to improve the functionality of CHOPCHOP, we will continue to accommodate new transcriptomes and genomes.

In conclusion, this major update expands the CHOPCHOP toolbox, retaining its position as one of the most easy-to-use, versatile CRISPR–Cas targeting tools available.

DATA AVAILABILITY

The CHOPCHOP (version 3) server is available at https://chopchop.cbu.uib.no; the python code for local installation is available at https://bitbucket.org/valenlab/chopchop.

ACKNOWLEDGEMENTS

We would like to thank James Graham from Oxford Nanopore Technologies for his recommendations on the long-read enrichment mode.

FUNDING

Bergen Research Foundation (to E.V.); University of Bergen core funding (to K.L.); Norwegian Research Council [250049 to E.V.]; National Defense Science and Engineering Graduate (NDSEG) Fellowship (to T.G.M.). Funding for open access charge: Bergens Forskningsstiftelse (Bergen Research Foundation).

Conflict of interest statement. None declared.

REFERENCES

1. Nakamae,K., Nishimura,Y., Takenaga,M., Nakade,S., Sakamoto,N., Ide,H., Sakuma,T. and Yamamoto,T. (2017) Establishment of expanded and streamlined pipeline of PITCh knock-in - a web-based design tool for MMEJ-mediated gene knock-in, PITCh designer, and the variations of PITCh, PITCh-TG and PITCh-KIKO. *Bioengineered*, **8**, 302–308.
2. Sakuma,T., Nakade,S., Sakane,Y., Suzuki,K.-I.T. and Yamamoto,T. (2016) MMEJ-assisted gene knock-in using TALENs and CRISPR–Cas9 with the PITCh systems. *Nat. Protoc.*, **11**, 118–133.
3. Suzuki,K., Tsunekawa,Y., Hernandez-Benitez,R., Wu,J., Zhu,J., Kim,E.J., Hatanaka,F., Yamamoto,M., Araoka,T., Li,Z. *et al.* (2016) In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature*, **540**, 144–149.
4. Maeder,M.L., Linder,S.J., Cascio,V.M., Fu,Y., Ho,Q.H. and Joung,J.K. (2013) CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods*, **10**, 977–979.
5. Perez-Pinera,P., Kocak,D.D., Vockley,C.M., Adler,A.F., Kabadi,A.M., Polstein,L.R., Thakore,P.I., Glass,K.A., Ousterout,D.G., Leong,K.W. *et al.* (2013) RNA-guided gene activation by CRISPR–Cas9-based transcription factors. *Nat. Methods*, **10**, 973–976.
6. Gilbert,L.A., Larson,M.H., Morsut,L., Liu,Z., Brar,G.A., Torres,S.E., Stern-Ginossar,N., Brandman,O., Whitehead,E.H., Doudna,J.A. *et al.* (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, **154**, 442–451.
7. Gabrieli,T., Sharim,H., Fridman,D., Arbib,N., Michaeli,Y. and Ebenstein,Y. (2018) Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH). *Nucleic Acids Res.*, **46**, e87.
8. Gilpatrick,T., Lee,I., Graham,J.E., Raimondeau,E., Bowen,R., Heron,A., Sedlazeck,F.J. and Timp,W. (2019) Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants and mutations. bioRxiv doi: https://doi.org/10.1101/604173, 11 April 2019, preprint: not peer reviewed.
9. Halperin,S.O., Tou,C.J., Wong,E.B., Modavi,C., Schaffer,D.V. and Dueber,J.E. (2018) CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature*, **560**, 248–252.
10. Gootenberg,J.S., Abudayyeh,O.O., Lee,J.W., Essletzbichler,P., Dy,A.J., Joung,J., Verdine,V., Donghia,N., Daringer,N.M., Freije,C.A. *et al.* (2017) Nucleic acid detection with CRISPR–Cas13a/C2c2. *Science*, **356**, 438–442.
11. McKenna,A., Findlay,G.M., Gagnon,J.A., Horwitz,M.S., Schier,A.F. and Shendure,J. (2016) Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, **353**, aaf7907.
12. Abudayyeh,O.O., Gootenberg,J.S., Konermann,S., Joung,J., Slaymaker,I.M., Cox,D.B.T., Shmakov,S., Makarova,K.S., Semenova,E., Minakhin,L. *et al.* (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, **353**, aaf5573.
13. Cox,D.B.T., Gootenberg,J.S., Abudayyeh,O.O., Franklin,B., Kellner,M.J., Joung,J. and Zhang,F. (2017) RNA editing with CRISPR–Cas13. *Science*, **358**, 1019–1027.
14. Abudayyeh,O.O., Gootenberg,J.S., Essletzbichler,P., Han,S., Joung,J., Belanto,J.J., Verdine,V., Cox,D.B.T., Kellner,M.J., Regev,A. *et al.* (2017) RNA targeting with CRISPR–Cas13. *Nature*, **550**, 280–284.
15. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. *Nat. Biotechnol.*, **34**, 184–191.
16. Chari,R., Mali,P., Moosburner,M. and Church,G.M. (2015) Unraveling CRISPR–Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods*, **12**, 823–826.
17. Xu,H., Xiao,T., Chen,C.-H., Li,W., Meyer,C.A., Wu,Q., Wu,D., Cong,L., Zhang,F., Liu,J.S. *et al.* (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res.*, **25**, 1147–1157.
18. Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.-D., Fernandez,J.P., Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR–Cas9 targeting in vivo. *Nat. Methods*, **12**, 982–988.
19. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
20. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
21. Kim,H.K., Min,S., Song,M., Jung,S., Choi,J.W., Kim,Y., Lee,S., Yoon,S. and Kim,H.H. (2018) Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nat. Biotechnol.*, **36**, 239–241.
22. El-Brolosy,M., Rossi,A., Kontarakis,Z., Kuenne,C., Guenther,S., Fukuda,N., Takacs,C., Lai,S.-L., Fukuda,R., Gerri,C. *et al.* (2019) Genetic compensation is triggered by mutant mRNA degradation. *Nature* , **568**, 193–197.
23. Gagnon,J.A., Valen,E., Thyme,S.B., Huang,P., Akhmetova,L., Ahkmetova,L., Pauli,A., Montague,T.G., Zimmerman,S., Richter,C. *et al.* (2014) Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One*, **9**, e98186.
24. Shen,M.W., Arbab,M., Hsu,J.Y., Worstell,D., Culbertson,S.J., Krabbe,O., Cassa,C.A., Liu,D.R., Gifford,D.K. and Sherwood,R.I. (2018) Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, **563**, 646–651.
25. Konermann,S., Lotfy,P., Brideau,N.J., Oki,J., Shokhirev,M.N. and Hsu,P.D. (2018) Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell*, **173**, 665–676.
26. Smargon,A.A., Cox,D.B.T., Pyzocha,N.K., Zheng,K., Slaymaker,I.M., Gootenberg,J.S., Abudayyeh,O.A., Essletzbichler,P., Shmakov,S., Makarova,K.S. *et al.* (2017) Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28. *Mol. Cell*, **65**, 618–630.
27. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
28. Thyme,S.B., Akhmetova,L., Montague,T.G., Valen,E. and Schier,A.F. (2016) Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.*, **7**, 11750.
29. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.
30. Cheng,A.W., Wang,H., Yang,H., Shi,L., Katz,Y., Theunissen,T.W., Rangarajan,S., Shivalila,C.S., Dadon,D.B. and Jaenisch,R. (2013) Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res.*, **23**, 1163–1171.
31. La Russa,M.F. and Qi,L.S. (2015) The new state of the art: Cas9 for gene activation and repression. *Mol. Cell Biol.*, **35**, 3800–3809.
32. Labun,K., Guo,X., Chavez,A., Church,G., Gagnon,J.A. and Valen,E. (2018) Accurate analysis of genuine CRISPR editing events with ampliCan. *Genome Res.*, **29**, 843–847.

6. Paper III

Accurate analysis of genuine CRISPR editing events with ampliCan.

K Labun, X Guo, A Chavez, G Church, JA Gagnon, E Valen,

2019, Genome Res. 29: 843-847

Accurate analysis of genuine CRISPR editing events with ampliCan

Kornel Labun,[1] Xiaoge Guo,[2,3] Alejandro Chavez,[4] George Church,[2,3] James A. Gagnon,[5] and Eivind Valen[1,6]

[1]*Department of Informatics/Computational Biology Unit, University of Bergen, Bergen 5008, Norway;* [2]*Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts 02115, USA;* [3]*Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA;* [4]*Department of Pathology and Cell Biology, Columbia University, New York, New York 10032, USA;* [5]*Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA;* [6]*Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen 5008, Norway*

We present ampliCan, an analysis tool for genome editing that unites highly precise quantification and visualization of genuine genome editing events. ampliCan features nuclease-optimized alignments, filtering of experimental artifacts, event-specific normalization, and off-target read detection and quantifies insertions, deletions, HDR repair, as well as targeted base editing. It is scalable to thousands of amplicon sequencing–based experiments from any genome editing experiment, including CRISPR. It enables automated integration of controls and accounts for biases at every step of the analysis. We benchmarked ampliCan on both real and simulated data sets against other leading tools, demonstrating that it outperformed all in the face of common confounding factors.

[Supplemental material is available for this article.]

With the introduction of CRISPR (Jinek et al. 2012; Cong et al. 2013), researchers obtained an inexpensive and effective tool for targeted mutagenesis. Despite some limitations, CRISPR has been widely adopted in research settings and has made inroads into medical applications (Courtney et al. 2016). Successful genome editing relies on the ability to confidently identify induced mutations after repair through nonhomologous end-joining (NHEJ) or homology directed repair (HDR). Insertions or deletions (indels) are often identified by sequencing the targeted loci and comparing the sequenced reads to a reference sequence. Deep sequencing has the advantage of both capturing the nature of the indel, readily identifying frameshift mutations or disrupted regulatory elements, and characterizing the heterogeneity of the introduced mutations in a population. This is of particular importance when the aim is allele-specific editing or the experiment can result in mosaicism.

The reliability of a sequencing-based approach is dependent on the processing and interpretation of the sequenced reads and is contingent on factors such as the inclusion of controls, the alignment algorithm, and the filtering of experimental artifacts. To date, no tool considers and controls for the whole range of biases that can influence this interpretation and, therefore, distort the estimate of the mutation efficiency and lead to erroneous conclusions. Here we introduce a fully automated tool, ampliCan, designed to determine the true mutation frequencies of CRISPR experiments from high-throughput DNA amplicon sequencing. It scales to genome-wide experiments and can be used alone or integrated with the CHOPCHOP (Montague et al. 2014; Labun et al. 2016) guide RNA (gRNA) design tool.

Corresponding author: eivind.valen@gmail.com
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.244293.118. Freely available online through the *Genome Research* Open Access option.

Results

ampliCan accurately determines the true mutation efficiency

Estimation of the true mutation efficiency depends on multiple steps all subject to different biases (Lindsay et al. 2016). Following sequencing, reads have to be aligned to the correct reference and filtered for artifacts, and then the mutation efficiency has to be quantified and normalized (Fig. 1A). In most existing tools, many of the choices made during these steps are typically hidden from the user, leading to potential misinterpretation of the data. These hidden steps can lead to widely different estimates of mutation efficiency (in up to 67% of all experiments) when run on data from real experiments (Supplemental Note S1; Supplemental Fig. S1). Furthermore, steps are frequently relegated to other tools that have not been optimized for CRISPR experiments. ampliCan instead implements a complete pipeline from alignment to interpretation and can therefore control for biases at every step.

Despite being arguably the most important step in any experiment, the use of controls is frequently overlooked in CRISPR assays. Discrepancies between a reference genome and the genetic variation in an organism of interest often lead to false positives and the false impression that mutations have been introduced (Gagnon et al. 2014). Although the use of controls is (in principle) possible with any tool, it commonly requires running the treated and control samples separately followed by a manual inspection and comparison. In ampliCan, controls are an integrated part of the pipeline, and mutation frequencies are normalized and estimated automatically. ampliCan accomplishes this by normalizing at the level of editing events (insertion, deletion, or mismatch) rather than at the level of whole reads. This means that any putative editing event detected in the reads from the target sample that also occurs in the reads from the control sample, above the level of

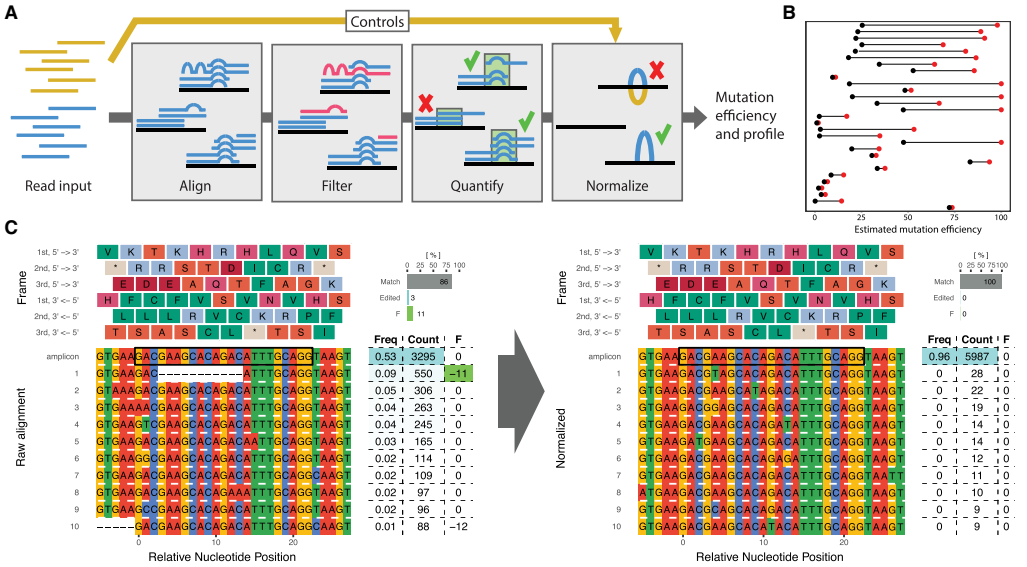


Figure 1. Overview of ampliCan pipeline and normalization. (*A*) Estimation of mutation efficiency consists of multiple steps. At each of these steps, biases can be introduced. Controls are processed identically to the main experiment and used for normalization. (*B*) Overview of the change in estimated mutation efficiency on real CRISPR experiments when using controls that account for natural genetic variance in 29 experiments (mean change of 30%). Red dots show initial estimates based on unnormalized data, whereas black dots show the values after normalization. (*C*) Alignment plot showing the top 10 most abundant reads in a real experiment. The table shows relative efficiency (Freq) of read, absolute number of reads (Count), and the summed size of the indel(s) (F), colored green when inducing a frameshift. The bars (*top right*) show the fraction of reads that contain no indels (Match), those having an indel without inducing frameshift (Edited), and frameshift-inducing indels (F). The *left* panel shows the estimated mutation efficiency from raw reads, which is 14% (11% with frameshift, 3% without). The *right* panel shows the same genomic loci after normalization with controls, resulting in a mutation efficiency of 0%. The deletion of 11 bp in 9% of the reads could not be found in the GRCz10.88 Ensembl Variation database and would, in the absence of controls, give the impression of a real editing event.

noise, is ignored when calculating mutation frequencies. Importantly, this normalization process does not remove any reads from the calculation; it only refrains from counting the specific editing events that are also present in the controls (Supplemental Figs. S2–S4; Supplemental Table S1). Therefore, it also does not filter any genuine editing events that may co-occur on the same read as a normalized event (see Supplemental Note S2). This process is blind to the source of the event, which may include genetic variance as well as experimental and sequencing artifacts. To assess the impact of controls, we generated 112 CRISPR data sets and pooled them with data we previously generated (Gagnon et al. 2014) for a total of 263 experiments (Methods; Supplemental Note S1; Supplemental Table S2). These consisted of pools of CRISPR-injected zebrafish using wild-type fish as a control. This experimental setup presents a challenging task to pipelines because the genetic background may not be identical across all fish and because the injected fish can be highly mosaic in their mutational outcomes. This benchmark revealed that accounting for the genetic background in the wild-type fish reduced the estimated mutation frequencies substantially in several experiments and is a necessary step to ensure accurate results (Fig. 1B,C; Supplemental Fig. S5).

Estimating mutation efficiency starts with the alignment of the sequenced reads (Fig. 1A). A common strategy is to use standard genomic alignment tools. However, these tools do not align using knowledge about the known mechanisms of CRISPR-induced double-stranded breaks and DNA repair. Genome editing typically results in a single deletion and/or insertion of variable length. Hence, correctly aligned reads will often have a low number of events (optimally one deletion and/or one insertion after normalization for controls) overlapping the cut site, whereas misaligned reads will result in a high number of events throughout the read owing to discrepancies to the correct loci. Therefore an alignment strategy that penalizes multiple indel events (see Methods) is more consistent with DNA repair mechanisms and the CRISPR mode of action. ampliCan uses the Needleman–Wunsch algorithm with tuned parameters to ensure optimal alignments of the reads to their loci and models the number of indel and mismatch events to ensure that the reads originated from that loci (see Methods; Supplemental Note S3). In contrast, nonoptimized aligners can create fragmented alignments, resulting in misleading mutation profiles and possible distortion of downstream analyses and frameshift estimation (Supplemental Fig. S6). In assessments, ampliCan outperforms the tools CrispRVariants, CRISPResso, and ampliconDIVider on the synthetic benchmarking previously used to assess these tools (Lindsay et al. 2016), in which experiments were contaminated with simulated off-target reads that resemble the real on-target reads but have a mismatch rate of 30% per base pair (Supplemental Fig. S7). A cause for concern is that the mapping strategy used in the pipelines of several tools (Supplemental Table S3) is not robust to small perturbations of this mismatch rate, and when we simulated contaminant off-target

data with varying degrees of mismatches to the on-target loci (see Supplemental Note S4), it led to a significant reduction in performance (Fig. 2, left). In contrast, ampliCan's strategy of modeling editing events to ascertain whether a read originated from the on-target or the off-target loci resulted in consistently high performance across a broad range of mismatch rates (Fig. 2, left; Supplemental Figs. S7, S8).

ampliCan can detect long indels and estimate HDR efficiency

Targeted insertion of shorter fragments through co-opting of the homology directed repair (HDR) pathway is becoming increasingly popular (Lackner et al. 2015; Kuscu et al. 2017). This, together with long indels occurring in regular CRISPR experiments (Supplemental Figs. S9, S10), presents a challenge for most CRISPR analysis tools. To assess the ability of the leading tools in recognizing long indels, we simulated data using the strategy from Lindsay et al. (2016), but restricted to indels of ≥10 bp. This revealed an inability of current pipelines to process these longer events (Fig. 2, right), typically stemming from alignment strategies that are unable to assign reads with long indels to the correct loci. In previous assessments, simulated data have often been restricted to short indels in which this weakness would not be apparent (Supplemental Note S5). By using a localized alignment strategy, based on primer matching (see Methods), ampliCan knows a priori which loci the reads are supposed to originate from. This alignment strategy therefore outperforms all other tools and robustly handles these longer indels (>10 bp) when they occur unintentionally (Fig. 2, right; Supplemental Fig. S11).

Intentional introduction of specific edits using donor templates is supported in ampliCan through an HDR mode in which it first aligns the donor template to the reference in order to identify editing events that are expected to take place in a successful integration. The presence of these success-events is then quantified in the edited samples, obtaining the frequency of integration. To assess this strategy, we simulated experiments with different levels of donor integration (a result of HDR) in the presence of different levels of cut loci but with donor introduction (a result of nonhomologous end-joining [NHEJ]). This revealed that only ampliCan can consistently recover both the true HDR and NHEJ efficiency (Supplemental Note S6; Supplemental Fig. S12). An identical strategy also makes it possible to quantify the efficiency of base editors (Komor et al. 2016; Gaudelli et al. 2017) by supplying ampliCan with templates in which the target bases have been altered.

ampliCan summarizes and aggregates results over thousands of experiments

To aid analysis of heterogeneous outcomes, ampliCan quantifies the heterogeneity of reads (Supplemental Fig. S13), the complete mutation efficiency for an experiment, and the proportion of mutations resulting in a frameshift (Fig. 1C, top right). It also aggregates and quantifies mutation events of a specific type if a particular outcome is desired (Supplemental Fig. S14). In addition, ampliCan provides overviews of the impact of all filtering steps (Supplemental Figs. S15, S16). Reports can be generated in several formats (Supplemental Tables S4, S5) and aggregated at multiple levels such as sequencing barcodes, gRNA, gene, loci, or any user-specified grouping (Supplemental Note S7). This enables exploration of questions beyond mutation efficiency such as the rules of gRNA design, whether a particular researcher is better at designing gRNAs than others (Supplemental Fig. S17), whether a given barcode is not working, or determining the stochasticity in the mutation outcome from a given gRNA (Supplemental Fig. S18).

Discussion

ampliCan offers a complete pipeline for genome engineering controlling for biases at every step of evaluation. When used with CRISPR, it can be integrated with the CHOPCHOP tool for gRNA design to incorporate all computational steps necessary for a CRISPR experiment. It scales from a single experiment to genome-wide screens and can be run with a single command. For more advanced users, it provides a complete and adaptable

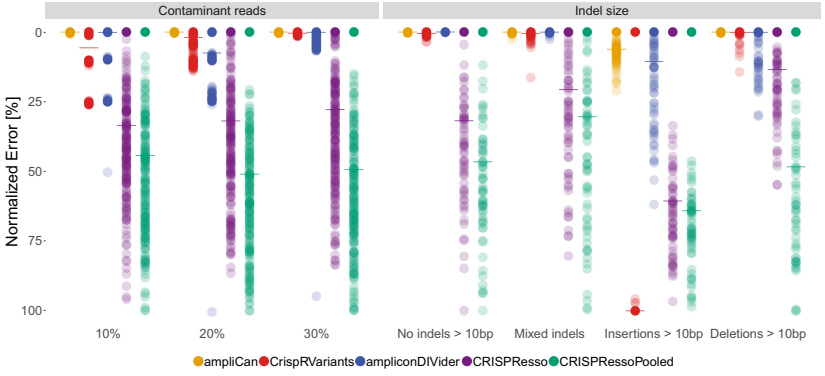


Figure 2. Benchmark of leading tools when estimating mutation efficiency under different data set conditions. Each dot shows the error of the estimate to the correct value for a single experiment normalized to a 0–100 scale. The median performance (mixed indels) is indicated by the horizontal line. The *left* panel shows comparison of tools when data sets contain contaminant reads (see text and Methods). The *x*-axis denotes how dissimilar the contaminant reads are to the correct reads. In cases in which the contaminants are from homologous regions, this may be low (10%); for other contaminants, this is likely to be higher (30%). The *right* panel shows performance of tools as a function of the length of indel events. The sets in the first column contain no indels >10 bp; the second column (Mixed indels) contains a mix of shorter and longer events; the sets in the third and fourth columns contain insertions and deletions >10 bp, respectively.

framework, enabling further exploration of the data. Collectively, these advances will minimize misinterpretation of genome editing experiments and allow effective analysis of the outcome in an automated fashion.

Methods

ampliCan pipeline

ampliCan is completely automated and accepts a configuration file describing the experiment(s) and FASTQ files of sequenced reads as input. The configuration file contains information about barcodes, gRNAs, forward and reverse primers, amplicons, and paths to corresponding FASTQ files (Supplemental Table S6). From here, ampliCan generates reports summarizing the key features of the experiments.

In the first step, ampliCan filters low-quality reads that have either ambiguous nucleotides, an average quality, or individual base quality under a default or user-specified threshold (Supplemental Note S8). After quality filtering, ampliCan assigns reads to the particular experiment by searching for matching primers (default up to two mismatches, but ampliCan supports different stringency) (Supplemental Note S9). Unassigned reads are summarized and reported separately for troubleshooting. After read assignment, ampliCan uses the Biostrings (https://bioconductor.org/packages/release/bioc/html/Biostrings.html) implementation of the Needleman–Wunsch algorithm with optimized parameters (gap opening = −25, gap extension = 0, match = 5, mismatch = −4, no end gap penalty) to align all assigned reads to the loci/amplicon sequence. Subsequently, primer dimer reads are removed by detecting deletions larger than the size of the amplicon, subtracting the length of the two primers and a short buffer. Additionally, sequences that contain a high number of indels or mismatch events compared with the remainder of the reads are filtered as these are potential sequencing artifacts or originate from off-target amplification (Supplemental Note S8; Supplemental Fig. S19). Mutation frequencies are calculated from the remaining reads using the frequency of indels that (Supplemental Fig. S14) overlap a region (± 5 bp) around the expected cut site. If paired-end sequencing is used, ampliCan follows consensus rules for the paired forward and reverse read, generally picking the read with the best alignment in case of disagreement (for description, see Supplemental Figs. S20, S21). The alternative strategy of merging the paired reads is supported by ampliCan but has been shown to be detrimental to performance (Lindsay et al. 2016). The expected cut site can be specified as a larger region for nickase or TALEN experiments in which the exact site is not known. Any indel or mismatch also observed above a 1% threshold in the control is removed. Frameshifts are identified by summing the impact of deletions and insertions on the amplicon.

A series of automated reports is prepared in form of ".Rmd" files, which can be converted to multiple formats but also immediately transformed into HTML reports with knitr (https://yihui.name/knitr/) for convenience. There are six different default reports prepared by ampliCan with statistics grouped at the corresponding level: identifier, barcode, gRNA, amplicon, summary, and group (user-specified, but typically signifies the researcher conducting the experiment, treatment of sample, or other grouping of interest). In addition to alignments of top reads (Fig. 1C; Supplemental Fig. S5), reports contain plots summarized over all deletions, insertions, and variants (Supplemental Fig. S14). In addition, a number of plots showing the general state of the experiments is shown, including the heterogeneity of reads to investigate mosaicism or sequencing issues (Supplemental Figs. S13, S22, S23) and overviews of how many reads were filtered/assigned at each step (Supplemental Fig. S24). In addition to the default plots, ampliCan produces R objects that contain all alignments and read information; these can be manipulated, extended, and visualized through the R statistical package.

ampliCan provides a versatile tool that can be used out-of-the-box or as a highly flexible framework that can be extended to more complex analysis. The default pipeline consists of a single convenient wrapper, amplicanPipeline, which generates all default reports. More advanced users can gain complete control over all processing steps (Supplemental Fig. S25) and produce novel plots for more specialized use cases. Compatibility with the most popular plotting packages ggplot2 (https://ggplot2.tidyverse.org) and ggbio (Yin et al. 2012), as well as the most popular data processing packages dplyr (https://dplyr.tidyverse.org) and data.table, provides a full-fledged and elastic framework. Output files are encoded as GenomicRanges (Lawrence et al. 2013) tables of aligned read events for easy parsing (Supplemental Table S5) and human-readable alignment results (Supplemental Table S4) and FASTA. We would like to encourage users to communicate their needs and give us feedback for future development.

Running parameters

Supplemental Code S1 and https://github.com/valenlab/amplican_manuscript both contain all code related to reproducibility of benchmark and analyses. For benchmarking, all the tools were used with their default options; specific versions of the tools and software can be found in the description file.

Software availability

ampliCan is developed as an R package (R Core Team 2018) under GNU General Public License version 3 and is available through Bioconductor under http://bioconductor.org/packages/amplican or https://github.com/valenlab/amplican. Supplemental Code S2 contains ampliCan source for installation, version 1.5.6.

Data access

All real data sets from this study come from the zebrafish TLAB strain and have been submitted to the NCBI BioProject database (BioProject; https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA245510 (run 1 and run 5). Other data sets used in this study, published previously, are described in the Supplemental Material. Descriptions, treatments, and other details of those data sets were previously described (Gagnon et al. 2014). Synthetic data sets can be reconstructed with the use of code from https://github.com/valenlab/amplican_manuscript (Supplemental Code S1). Synthetic data sets were created in a similar fashion to the sets previously described (Lindsay et al. 2016) using 20 different loci edited at variable efficiency (0%, 33.3%, 66.7%, and 90%) and with the possibility of adding HDR. Further details can be found in the Supplemental Material.

Acknowledgments

We thank Jason Rihel, Tessa Montague, and Alex Schier for support and many useful comments and the members of the Schier laboratory for their contributions. The project was supported by the Bergen Research Foundation and the Norwegian Research Council (FRIMEDBIO #250049; E.V.), University of Bergen core funding (K.L.), and the American Cancer Society and University of Utah start-up funding (J.A.G.).

Author contributions: E.V. conceived and supervised the project. J.A.G. performed wet-lab experiments and prepared data sets. X.G., G.C., and A.C. assisted in data interpretation and writing the manuscript. K.L. developed the R package and performed all computational work.

References

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339:** 819–823. doi:10.1126/science.1231143

Courtney DG, Moore JE, Atkinson SD, Maurizi E, Allen EHA, Pedrioli DML, McLean WHI, Nesbit MA, Moore CBT. 2016. CRISPR/Cas9 DNA cleavage at SNP-derived PAM enables both *in vitro* and *in vivo KRT12* mutation-specific targeting. *Gene Ther* **23:** 108–112. doi:10.1038/gt.2015.82

Gagnon JA, Valen E, Thyme SB, Huang P, Akhmetova L, Pauli A, Montague TG, Zimmerman S, Richter C, Schier AF. 2014. Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* **9:** e98186. doi:10.1371/journal.pone.0098186

Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, Liu DR. 2017. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551:** 464. doi:10.1038/nature24644

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337:** 816–821. doi:10.1126/science.1225829

Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533:** 420–424. doi:10.1038/nature17946

Kuscu C, Parlak M, Tufan T, Yang J, Szlachta K, Wei X, Mammadov R, Adli M. 2017. CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nat Methods* **14:** 710–712. doi:10.1038/nmeth.4327

Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E. 2016. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res* **44:** W272–W276. doi:10.1093/nar/gkw398

Lackner DH, Carré A, Guzzardo PM, Banning C, Mangena R, Henley T, Oberndorfer S, Gapp BV, Nijman SMB, Brummelkamp TR, et al. 2015. A generic strategy for CRISPR-Cas9-mediated gene tagging. *Nat Commun* **6:** 10237. doi:10.1038/ncomms10237

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118

Lindsay H, Burger A, Biyong B, Felker A, Hess C, Zaugg J, Chiavacci E, Anders C, Jinek M, Mosimann C, et al. 2016. CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nat Biotechnol* **34:** 701–702. doi:10.1038/nbt.3628

Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. 2014. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res* **42:** W401–W407. doi:10.1093/nar/gku410

R Core Team. 2018. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Yin T, Cook D, Lawrence M. 2012. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol* **13:** R77. doi:10.1186/gb-2012-13-8-r77

Received September 16, 2018; accepted in revised form February 25, 2019.

Supplemental Text and Figures

Table of contents

Availability of code and data

All code and configuration files used for the analyses in this manuscript are available at https://github.com/valenlab/amplican_manuscript where they can be used for independent verification. Copy of this repository is also available as **Supplementary Code S1**. Data analysed is publicly available and accessible at their respective accession numbers. ampliCan is available from Bioconductor as the R package at http://bioconductor.org/packages/amplican. To obtain the newest development version visit https://github.com/valenlab/amplican.

Data Overview

In this manuscript we have used multiple datasets, both real and simulated. Real datasets from Gagnon et al. 2014, runs 6-10 (accessible under E-MTAB-6310, E-MTAB-6355, E-MTAB-6356, E-MTAB-6357, E-MTAB-6358) were supplemented with our experiments for a total of 263 loci. All experiments had a control sample where no guideRNA was injected (accessible under BioProject PRJNA245510, run 1 and run 5). This dataset was used to highlight importance of controls as well as differences between the tools. The data is further discussed in **Supplemental Note S1**. Additional real data from ~ 1400 loci from Chari et al. 2015 was used for assessment whether large deletions can happen in real experiments.

Simulated datasets were created using a strategy similar to Lindsay et al. 2016 where CRISPR editing is emulated based on distributions of events (mismatches, deletions and insertions) from real experimental data (20 loci from (Shah et al. 2016)). FASTQ files were created using ART (Huang et al. 2012) where qualities were set to be uniformly high. Configuration files were

created using base loci sequences, same as used for CRISPR editing simulation. This data is further detailed in **Supplemental Note S4**. Complete overview of all datasets is presented in **Supplemental Tab S6**.

Supplemental Note S1: Tools differ in estimating editing efficiency on real data.

We assessed how tools estimate editing efficiency on 263 real CRISPR experiments, of which 151 were previously published by us (Gagnon et al. 2014), datasets from run 1 and run 5 available at BioProject under accession number PRJNA245510), and 112 novel experiments from 5 sets for this study (datasets from run 6-10 available at ArrayExpress: E-MTAB-6310, E-MTAB-6355, E-MTAB-6356, E-MTAB-6357, E-MTAB-6358). All experiments were conducted by injection into 1 cell zebrafish embryos and sequenced 2 days post-fertilization (Gagnon et al. 2014). Due to the rapid cell division and development these experiments are likely to result in highly heterogeneous mutational efficiencies from mosaicism. For these experiments the true mutation efficiency is not known and we can therefore not assess how precise the tools are in their estimates. Instead, we quantified how much the tools differ in their estimates (**Supplemental Fig S1A)** and, to qualitatively assess the underlying reason for their discrepancy, we plotted the estimated mutation efficiency values of the tools relative to the non-normalized ampliCan result (**Supplemental Fig S1B**). This showed that discrepancies are likely to originate from different causes. Some, those above the normalized ampliCan estimate, likely stem from a failure to consider control experiments. In our data the experiments impacted by controls is about 5%, but this will depend heavily on the reference genome, heterogeneity of the region and organism under study. Specific examples of the importance of normalization are shown in **Supplemental Fig S5**. The discrepancies of the other experiments

are due to the steps in the processing pipeline, e.g. off-target detection, primer dimer filtering, alignment strategy and read merging. To investigate the specific sources of these discrepancies and quantitatively assess the performance of the tools we created several synthetic benchmark datasets.

Supplemental Note S2: Automatic normalization using control reads

By default ampliCan normalizes through the strict removal of all editing events (insertions, deletions, mutations) that are also found above a threshold in the control sample. The default threshold value is a frequency of 0.01 and was chosen based on the typical frequency of low-abundance editing events ('background noise') present in control experiments (**Supplemental Fig S4**). These events are assumed to be technical or experimental artifacts present in most experiments. The threshold is also selected to be well above the expected Illumina error rate (Ross et al. 2013).

The threshold can be adjusted to increase the precision of indel detection when sequencing depth is high or account for a higher error rate in low-depth/precision experiments. The threshold can be set for instance based on the background error in the user's own control experiments or in the absence of controls by inspecting the error rate outside of the target site (see for instance in **Supplemental Fig S14A**). Alternatively, if the user has information about the level of variance or noise expected in the case and/or the controls (e.g. genetic variance restricted to 100% or 50%) the threshold can be raised for increased stringency. This may in particular be useful if the user expects high frequency background events in the control experiments. One such use case is if index hopping is likely to be an issue. Index hopping may cause reads from the edited sample to erroneously be assigned to the control sample. We

therefore offer a second pipeline 'amplicanPipelineConservative' for experiments where this is expected to be a problem featuring a more stringent threshold of 0.15. For more information about index hopping and how to mitigate the problem see Illumina's webpage (https://www.illumina.com/science/education/minimizing-index-hopping.html, Accessed January 22, 2019).

Given that the sequencing depth is sufficient the default settings should allow detection of indels as low as 0.01% (**Supplemental Tab S5**). If higher accuracy is necessary this threshold can be lowered and the sequencing depth increased. In the extreme case of setting the normalization threshold to 0% any event found in control would be removed from the case sample. This may result in removal of real edits and consequently the underestimation of real editing events.

Due to the stochasticity in sequencing data this approach is better suited to handle more heterogeneous cases than the subtraction method where indel frequency is simply normalized by subtraction of the control indel frequency. In the latter case variation in the levels of indel frequencies in the control versus CRISPR treated samples can lead to partially normalized data. Both normalization methods are outlined in **Supplemental Fig S2** while **Supplemental Fig S3** shows examples of highly heterogeneous data where without normalization estimated mutation efficiency would be biased. Normalization becomes even more important when the exact nature of the indel event matters, for instance whether it induces a frameshift or not (**Supplemental Fig S5**). Examples in **Supplemental Fig S5** can be recreated with the use of "make_comparison_normalized.Rmd" in the amplican_manuscript repository.

As a test for the normalization method we calculated the (non-existent) mutation frequency of the control samples normalized by the CRISPR-edited samples in order to see if false positives were generated. This was accomplished by 1) only considering the region between the sequence complementary to the guide RNA and the primer-matching sequence, and 2) excluding all large deletion events (>10bp) which could extend beyond the boundaries of this region. This resulted in a mean estimated mutation frequency of 0.0038 (median of 0.0012), well below the detection limit for the standard settings of 0.01.

Supplemental Note S3: ampliCan utilizes optimized alignments

CRISPR genome editing events typically result in a single break at a single site and by extension produce a single deletion and/or insertion. Sequence read aligners are generally not optimized for this type of genome editing event which can lead to the aligner fragmenting the indels and creating multiple events (example in **Supplemental Fig S6**). In the worst case fragmented alignments could shift the indel events resulting in a distortion of the mutation efficiency for those tools that only allow events within a certain distance from the expected site. A more likely outcome however, is the misinterpretation of the nature of the mutation.

Under certain assumptions the theoretically optimal alignment can be obtained by the Needleman-Wunsch algorithm. ampliCan uses this algorithm with optimized parameters to reflect the expectation that a CRISPR experiment should result in one deletion and/or insertion event, of unknown length (match = 5, mismatch = -4, gap opening = 25, gap extension = 0, no end gap penalties). With these parameters the Needleman-Wunsch algorithm performs well over a broad range of test cases (data not shown). ampliCan uses these optimized parameters

by default, but also allows for supervision of the alignments through human readable output of individual alignment results (**Supplemental Table S4**).

Supplemental Note S4: Synthetic data set evaluation

The latest available versions was used for all tools and packages. The assessment set from Lindsay et al. (Lindsay et al. 2016) paper (Synthetic Dataset 2, Supplemental 4) was replicated with the same settings and seed values as described (**Supplemental Fig S7**). The script from Lindsay et al. 2016 was used for parsing, but a small bug in the code was fixed for the CRISPResso output. In Lindsay et al. only NHEJ estimation of mutation efficiency was considered for CRISPResso, skipping HDR and "mixed" mutation frequencies. However, fixing this error did not influence CRISPResso's overall performance in any significant way. Versions of tools, scripts and details needed for replication are available in the https://github.com/valenlab/amplican_manuscript repository. ampliCan used the same amplicon sequences as CRISPResso.

It should be noted that Synthetic Dataset 2 from Lindsay et al. 2016 (used for **Supplemental Fig S7**) is not a good approximation of a real life situation. First, the sequence matching the primers can not be very divergent as they would then fail to amplify. Second, several experiments are badly designed in that the target sites are very close to the sequencing end of the reads. This makes it difficult to correctly call indels with support from both paired reads. Third, paired-end sequencing of 200bp or longer is somewhat expensive and error-prone and most labs would seek to restrict this to shorter reads. To account for this we created an additional set, Synthetic Dataset 3, in a similar fashion to Synthetic Dataset 2, but with with the following minor modifications. First, the length of amplicons and reads (150 bp) were adjusted.

Second, gRNA target sites were designed to be covered by both reads. Third, PCR off-target reads were created without mutating the primer sequences. Finally, mutation efficiency was tested across a range of mismatch rates, 10%, 20% and 30% (**Fig 2**, **Supplemental Fig S8**), to reflect different levels of similarity to the contaminant reads.

For Synthetic Dataset 2 ampliCan matches the perfect score of CrispRVariants and AmpliconDivider. However, on Synthetic Dataset Dataset 3 ampliCan is more consistent at estimating the known mutation efficiencies within the dataset (**Fig 2**). AmpliconDIVider has no filtering step and is confused by the contaminating reads. CrispRVariants has a filtering step, but is unable to discern divergent off-target sequences (e.g. homologous regions) that are still able to align to the correct target site. As in the benchmark from Lindsay et al. 2016 CRISPResso performs poorly on all benchmarks. When increasing mismatch rate (from 10% of all bases to 20% and 30%), AmpliconDIVider and CrispRVariants get closer to the correct estimated indel rate, but in all cases ampliCan obtains the highest precision and shows the most robust performance (**Fig 2**, **Supplemental Fig S8**).

Supplemental Note S5: ampliCan is able to correctly call longer indels

We have found that even without targeted insertion CRISPR mutagenesis can frequently result in some proportion of longer indels (**Supplemental Fig S9**). In particular, we have observed unintended insertions from lentiviral vectors used to introduce the guides and Cas9 (**Supplemental Fig S10**, (Chari et al. 2015).

Current tools primarily rely on either global mapping (CrispRVariants, AmpliconDIVider) (**Supplemental Tab S1**) that can have problems identifying the correct loci in the presence of a

larger insertion or have certain processing steps that are incompatible with longer events (see below for CrispRVariants). This mitigates primer dimer contamination problems (which can be identified by too large deletion gaps after alignments), but ignores bona fide large indels. These tools are therefore often unable to handle longer indels whether unintended or targeted. Long deletions are also a problem for some tools. For instance, CrispRVariants filters out any deletion that does not start or end within the gRNA complementary sequence plus a buffer of 5 bp. Any deletion spanning this region is ignored. This can be used as a strategy to filter primer-dimers, but also has the side-effect of ignoring any bona fide longer deletions. ampliCan uses a local alignment strategy that can detect these longer indels and a more realistic model of primer-dimer artifacts (**Supplemental Note S8**).

We noticed that in the Synthetic Dataset 2 from Lindsay et al. 2016 (CrispRVariants benchmark dataset) large indels (>10 bp) were disabled. To assess the capabilities of leading tools in handling longer indels we created Synthetic Dataset 4. We made three subsets: 1) with no indels > 10bp, 2) with a mix of indels by simply removing the line disabling longer reads in the Lindsay et al. script. 3) To check explicitly how tools handle experiments with planned shorter insertion of donor sequence we created a third scenario described as “insertions > 10bp” on the figure. 4) For completeness we also created a set of large deletions. ampliCan match the best competitors on the the set with no long indels and consistently outperform the other tools on the mixed set and the set only containing long indels (**Fig 2**, **Supplemental Fig S11**).

Supplemental Note S6: ampliCan consistently recovers the true HDR efficiency when faced with diverse donor templates

ampliCan takes into account the donor template and the original genomic sequence to define the set of events that corresponds to a correct HDR editing experiment, but allowing for some background sequencing noise (currently 3 mismatches by default). This is unlike CRISPResso, the other CRISPR tool that can handle HDR events, which do not model events but simply align reads against donor and original sequence picking the best-scoring instance. The advantage of ampliCan's approach is that it accounts for alignment imperfections in a more robust fashion, allowing for complex donor-amplicon relations and sequencing errors.

We designed a dataset for benchmarking the HDR calling capabilities of the most popular tools. Using the same loci as in **Supplemental Note S1** we tested 20 different donor templates for each of three kinds of donor types: with point mutations, insertions or deletions of variable length from 5bp to 70bp introduced into the amplicon sequences. We simulated 2000 reads with different levels of HDR efficiency rate (0, 33, 66, 90). In this benchmark set only ampliCan makes no errors (**Supplemental Fig S12**).

Supplemental Note S7: Visualization and aggregation of the complete activity of gRNAs

While the default alignment plot shows the most abundant reads across the expected cut site it doesn't provide an overview over all editing events. ampliCan therefore also produces multiple plots that aggregate and visualize editing events (**Supplemental Fig S14**). Unlike the alignment plots these show the complete activity of the gRNA allowing for comparison of gRNAs by

manual inspection. The pipeline in ampliCan treats forward and reverse reads separately which, after visualization, makes it possible to spot read-related problems immediately (**Supplemental Figs S21, S22**, **S23**). In addition, ampliCan provides meta plots that aggregate information across groups of experiments allowing for visualization of deletions, mismatches or insertions across groups of gRNAs, amplicons or any other set. This can for instance show the combined activity of a single guide across multiple experiments.

ampliCan builds on top of ggplot2 (Wickham 2016) package and provides higher level functions that automatically group event data (eg. collapse on start and end of deletion) and plot results. Users can extend those plot objects and treat them like any other ggplot2 object. ampliCan supports multiple types of meta plots to facilitate comparison of not only the gRNAs, but also any group that a user wants e.g. barcode, amplicon, type of treatment (**Supplemental Figs S3**, **S17**, **S18**).

Supplemental Note S8: Filtering of noise

Multiple sources of noise can confound the estimation of cut rates, low quality reads, primer-dimers, PCR off-target amplification and sequencing artifacts. ampliCan has three filters to remove noise from different sources: 1) low quality reads, 2) primer-dimers, and 3) erroneously assigned reads and sequencing artifacts.

Low quality reads

ampliCan offers basic read quality overview with the use of ShortRead (Morgan et al. 2009) package and filters for minimum base quality (default: 0) in a read, average base minimum quality (default avg min: 30) and the presence of ambiguous (N) letters.

Primer-dimers

Filtering primer-dimers is a balance between getting rid of erroneous reads and allowing for longer deletions. For instance, CrispRVariants ignores all alignments with deletions larger than 33 bp (the guide plus a buffer of 5 bp) and is frequently unable to map long insertions (**Fig 2B**, **Supplemental Fig S11**). This effectively removes all primer dimers, but also ignores any bona fide longer indels. ampliCan instead tries to estimate the likely length of a primer dimer deletion by taking the length of the amplicon, subtracting the primer lengths with a small buffer (30 bp) to arrive at maximally allowed deletion length. This results in a more realistic estimate of the length of artificial deletions that would result from primer-dimers. As an example, for an amplicon of size 150 with primers of length 20, the maximum deletion length would be 80 bp (150-(2*20+30)).

Erroneously assigned reads and sequencing artifacts

An assumption of ampliCan is that a CRISPR editing event will result in a low number of indel events resulting in a good alignment with few discrepancies to the reference sequence or (if available) the control experiment. To accomplish this ampliCan uses a two dimensional clustering method based on sequence alignment score and sequence alignment indel events to filter out erroneous reads and sequencing artifacts. This takes all alignments and performs k-means clustering with different 1-3 clusters. It then uses the silhouette criterion to determine the optimal number of clusters. In the case of 1 cluster, all reads are either edited or perfectly matching the reference/control. In the case of 2 you have both edited and unedited reads. In the case of 3 clusters, cluster with center that has the biggest number of events and lowest alignment score (in normalized relation on 0-1 scale) means that you in addition have a group

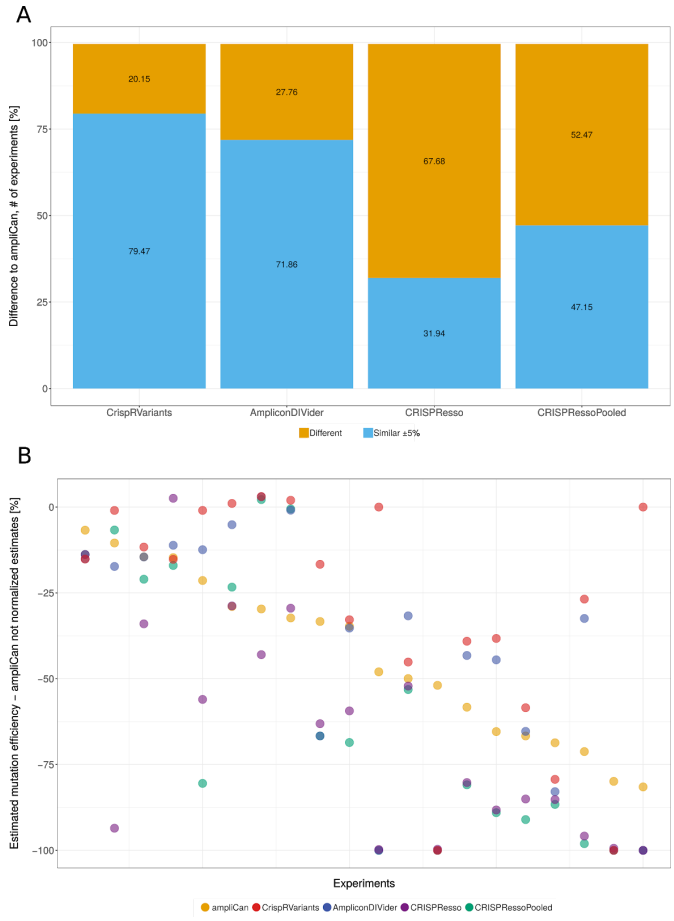
of sequencing artifacts or reads that poorly align to the loci (example in **Supplemental Fig S19**).

ampliCan provides plots that shows the impact of each of the filtering steps, across the whole library for read quality (**Supplemental Fig S16**) and for each experiments for primer-dimer and assignment/artifacts issues (**Supplemental Fig S15**).

Supplemental Note S9: Read assignment

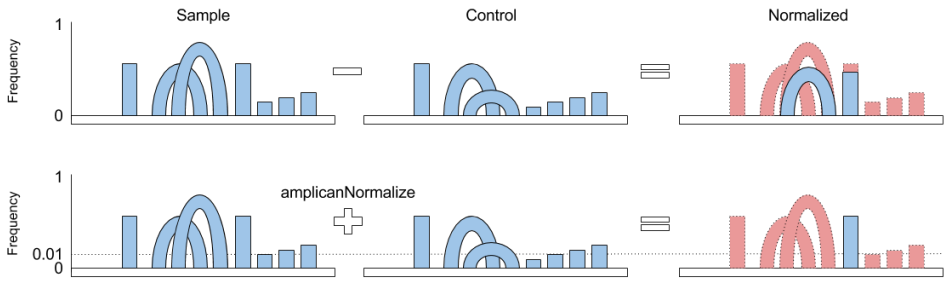
ampliCan assigns reads to the respective experiment by matching primers used in the amplification of the loci. These region should be immutable and match the reads since an indel spanning a primer would either result in failure to amplify the locus or be "corrected" by the primer when it amplifies the target site. However, since small sequencing and primer synthesis errors could potentially occur in the primer part ampliCan allows for up to 2 mismatches (user customizable) between the primers and reads. During this process it is possible that some reads will be unassigned and not match any of the experiments. While these reads are typically noise from the high-throughput nature of the sequencing experiment, they could in some cases be helpful in troubleshooting failed experiments. ampliCan therefore provides human readable alignments of the top 5 most abundant forward and reverse read pairs aligned to each other (**Supplemental Fig S24**). In some cases these correspond to off-target PCR amplicons and close homologous regions as well as errors in the specification of the experiment.

Supplemental Figures

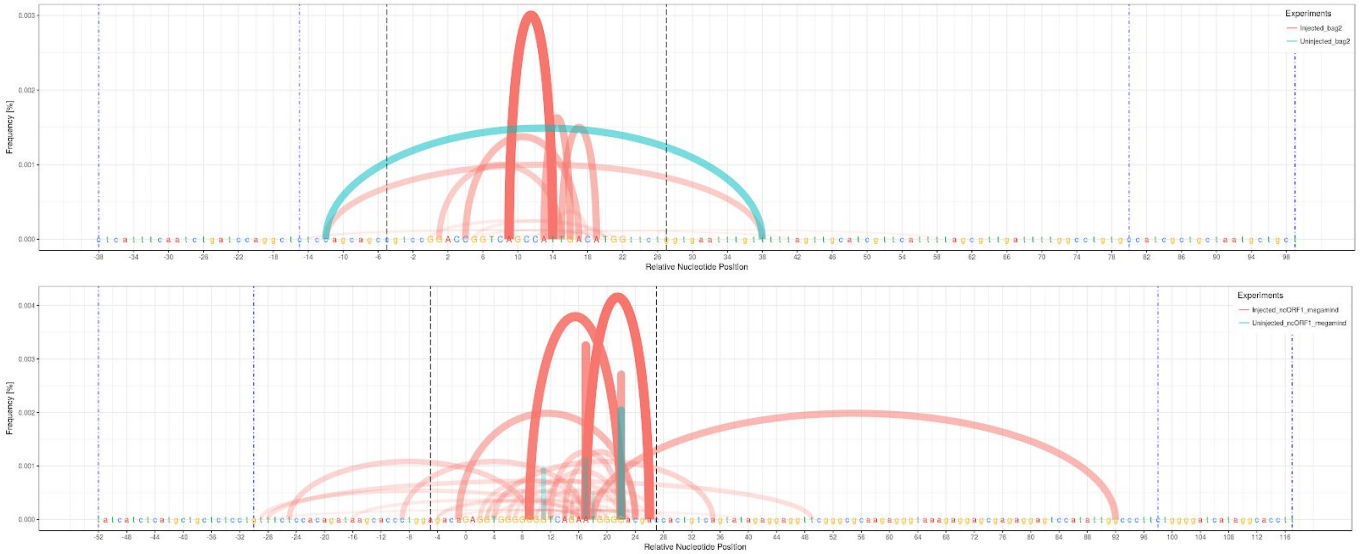


Supplemental Fig S1. Comparison of leading tools on real CRISPR experiments. **A.** Summary of differences between ampliCan and other tools. CrispRVariants reports similar editing efficiencies to ampliCan in ~80% out of 263 experiments. The remaining experiments are due to controls (~5%) and processing (~15%) **B.** Experiments (x axis, sorted) where estimated mutation efficiency differs by at least 5% from non-normalized data. y-axis shows differences in relation

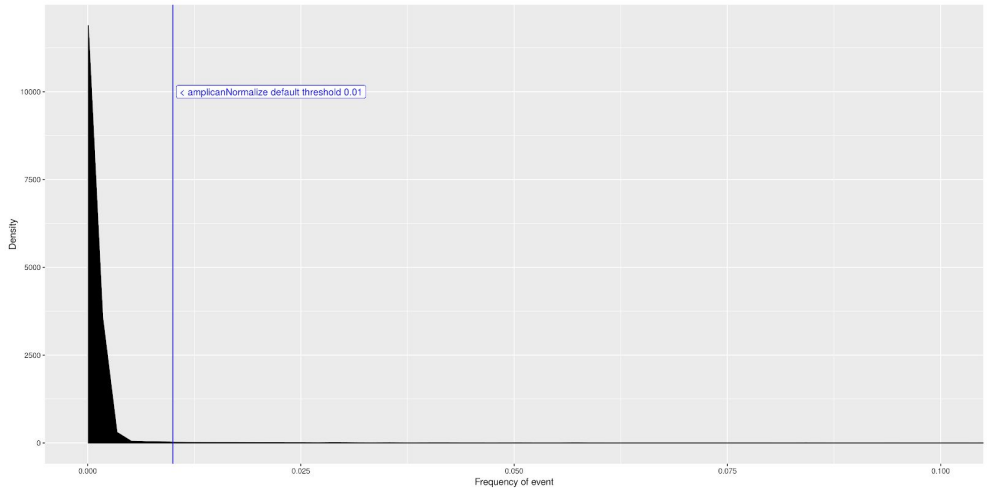
to non-normalized ampliCan estimates. Differences between tools predictions staying above line created by ampliCan prediction are likely to be due to lack of normalization, while the predictions below the ampliCan are likely due to the alignments, processing and filtering of data.



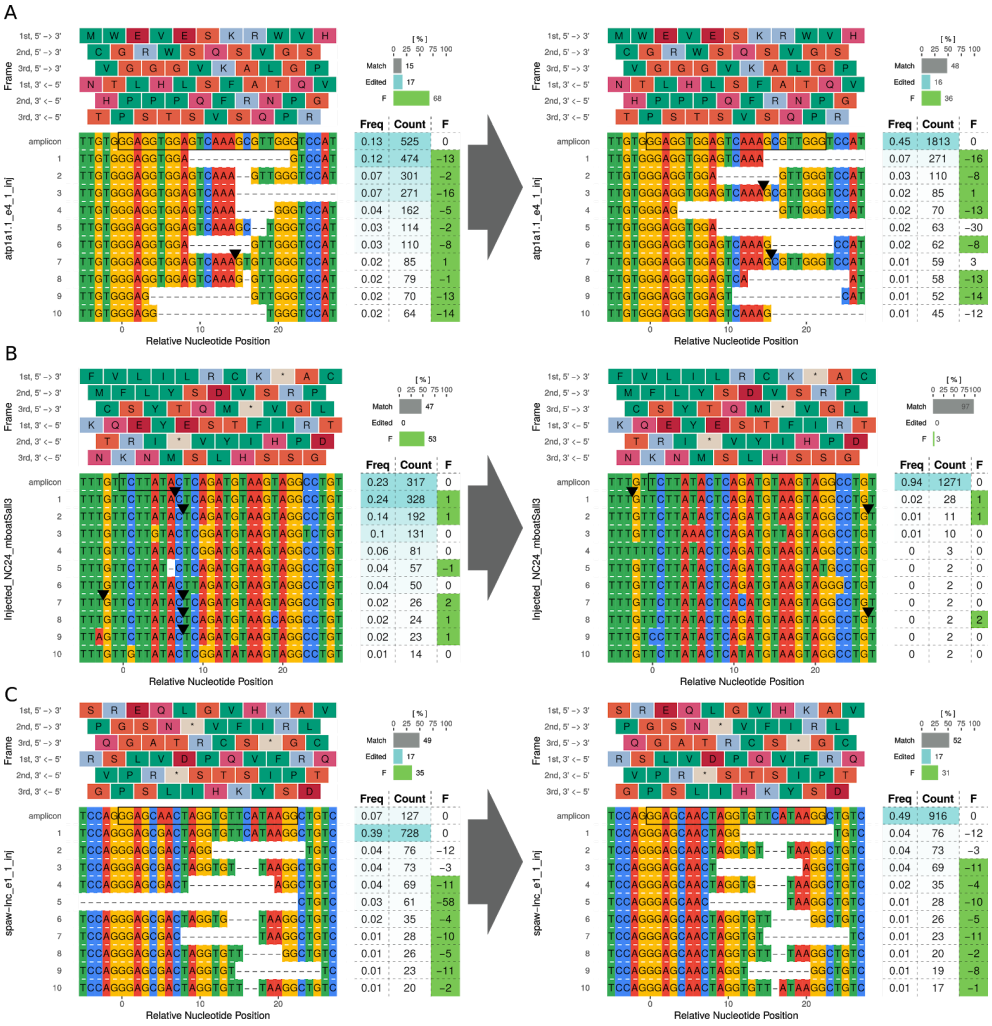
Supplemental Fig S2. Two common methods for normalization using controls. In the first, the frequencies from events in the control sample are subtracted from the frequencies in the treated sample (subtraction method). In the second method, all events occurring in the control above a frequency threshold (ampliCan default: 1%) are removed from the treated sample. The latter is more robust to stochastic differences.



Supplementary Fig S3. Deletion plots produced by ampliCan for two experiments showing deletions in Cas9 injected (red) versus control (cyan) samples. The archs indicate deletions (x-axis, start to end of arch) present in the samples at a frequency indicated by the y-axis and transparency. The blue, vertical dotted lines shows the start and end of the primers. In the top panel experiment normalization using the subtraction method would filter the big deletion also present in the control, but it would not completely get rid of the 1bp long deletion in the lower panel experiment. Both cases would get expunged with the default ampliCan normalization.

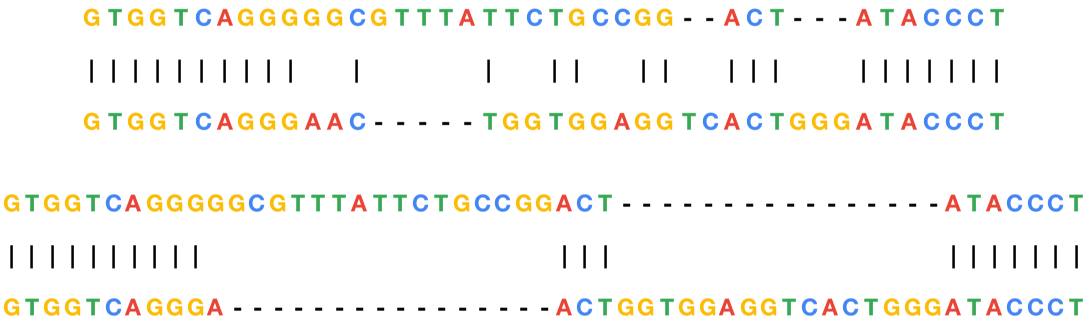


Supplemental Fig S4. Distribution of the frequency of events in the control sample of real experiments. The vast majority are low frequency events likely to be technical and experimental artifacts. ampliCan uses a threshold to exclude these so they are not considered for normalization. The default threshold is a frequency of 0.01 which excludes more than 99% of these events as noise. These low-abundance events are therefore not considered when normalizing the target loci. If the user know the level of variance expected in the controls they can raise the threshold for increased stringency.

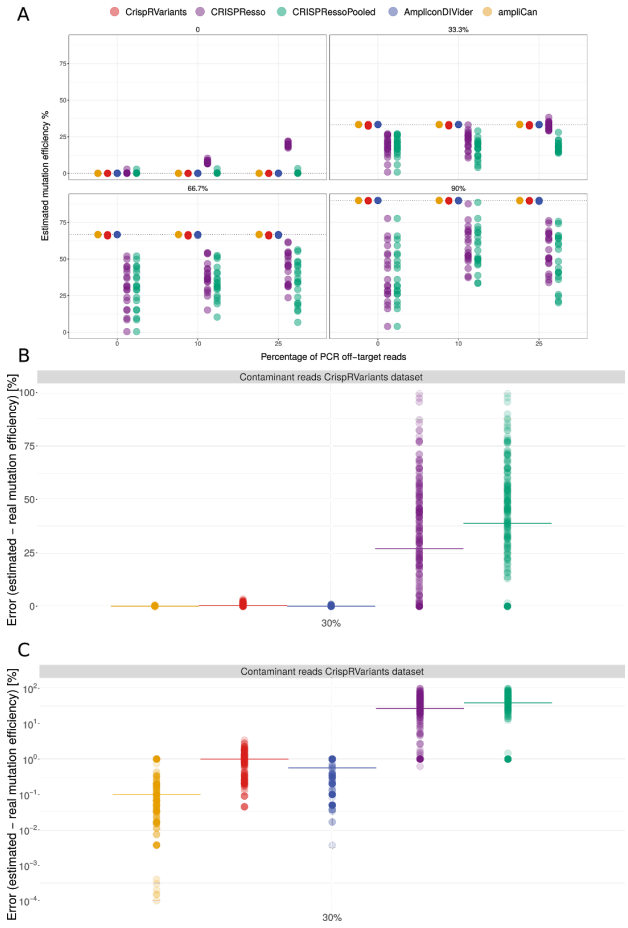


Supplemental Fig S5. Example of normalization with controls on 3 real experiments. In heterogeneous data it can be challenging to derive the true mutation profile. **A.** ampliCan automatically removes wild type mutations and reduce the number of frameshift inducing mutations from 68% to 36%. **B.** A large number of insertions could be mistaken for CRISPR activity reducing the total indel rate from 53 % to 3 %. **C.** A large fraction of reads (39%) carry a G instead of an A. This mismatch also occur at high frequency in the control and is therefore

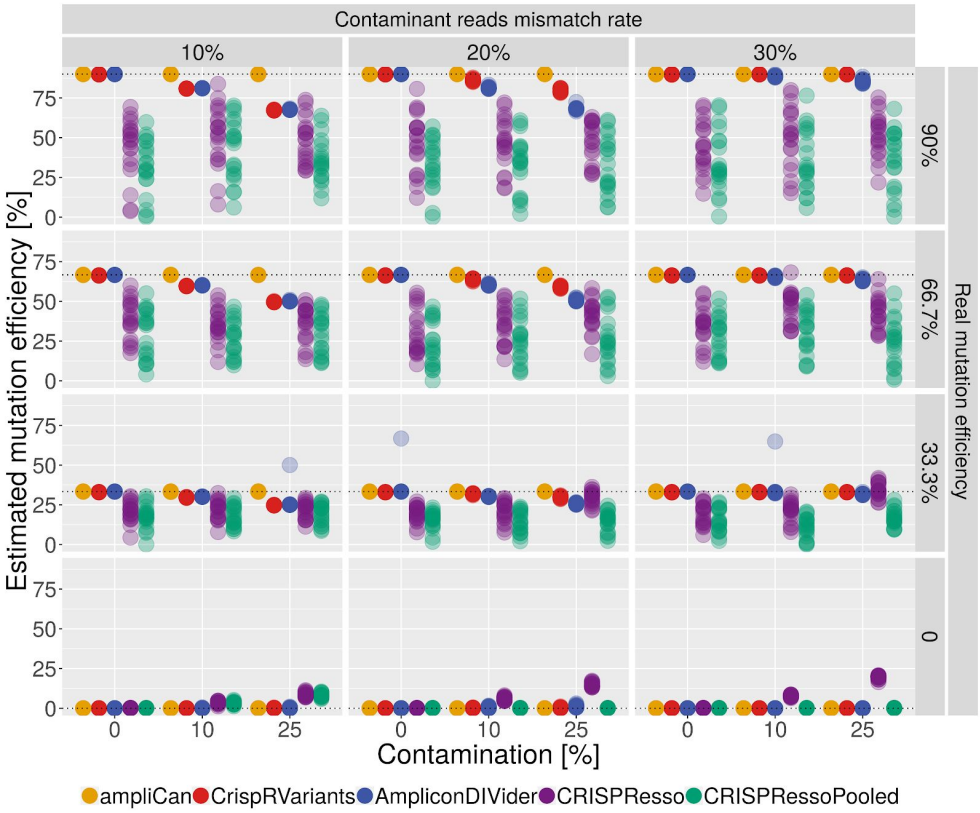
assumed to represent genetic variance. The mismatch is ignored and the reads are merged with the counts of the other unedited reads.



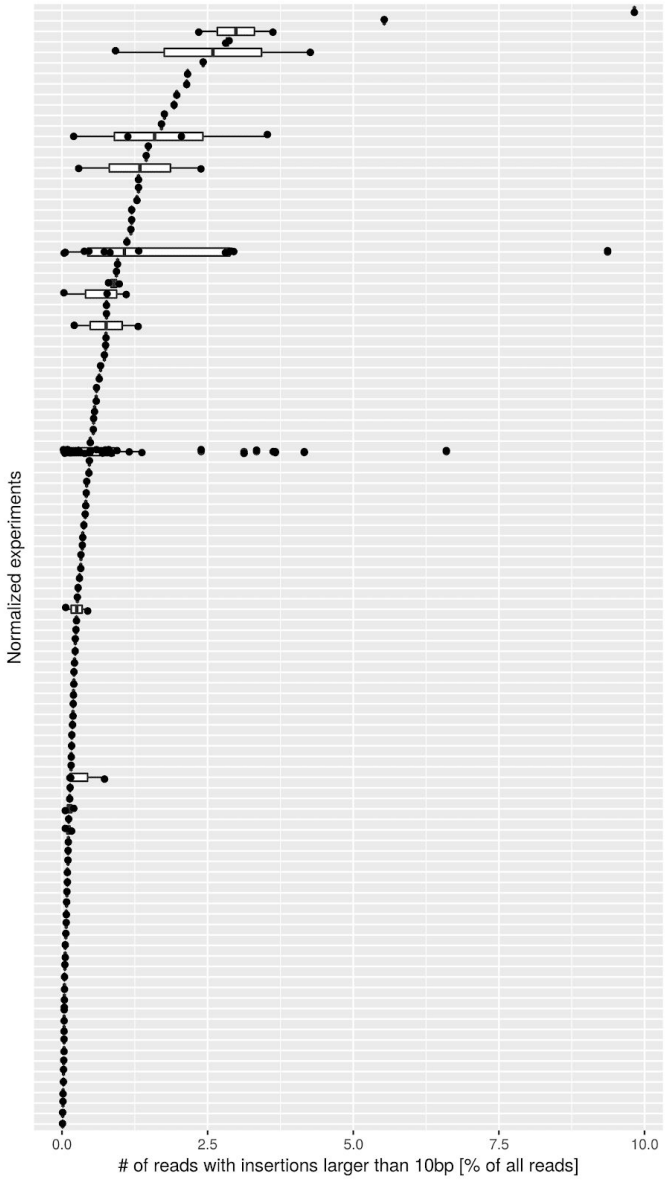
Supplemental Fig S6. Alignments dependent on multiple parameters and are often a tradeoff between the cost of gaps and the cost of mismatches. The top alignment shows a typical alignment for many aligners that are not optimized for genome editing (gap opening 10, gap extension 2, match 5, mismatch -4 and no end gap penalty). The bottom alignment, illustrates the result with ampliCan parameters (gap opening -25, gap extension 0, match 5, mismatch -4 and no end gap penalty) that are optimized for few events as is expected for CRISPR activity.



Supplemental Fig S7. ampliCan outperforms CrispRVariants on its own synthetic dataset. **A.** Performance of leading tools when facing different mutation efficiencies and fractions of contaminating reads. The sets are obtained from the benchmark data in (Lindsay et al. 2016) (CrispRVariants), Synthetic Dataset 2 (Supplemental Fig. 15 in (Lindsay et al. 2016). Each dot in the plot correspond to the estimated mutation efficiency calculated by a single tool for a single experiment, while the dotted line shows the true mutation efficiency The fraction of contaminant reads varies on the x-axis. **B.** Error rates (**C.** log10 scaled) of the same experiments. The median error is indicated by the horizontal line.

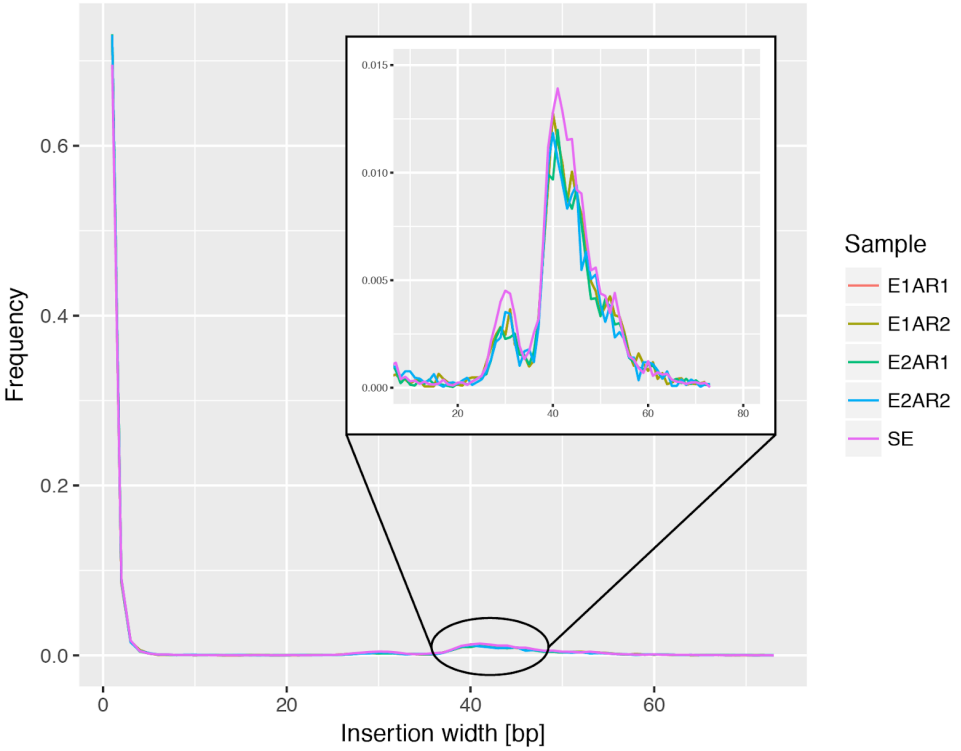


Supplemental Fig S8. The data from left panel of Fig. 2 split by the true mutation efficiency and with a baseline of 0% true mutation efficiency added for comparison. The dots shows estimated mutation efficiency by leading tools on reads with increasing contamination (as 0, 10 and 25 percentage of all reads) with different mismatch rate of the off-target reads (10%, 20%, 30%). Each point corresponds to one experiment. True mutation efficiency is indicated with dotted lines, and labelled to the right of the charts. Contamination is simulated by introducing random mismatches (Contaminant read mismatch rate) in reads mapping to the loci, similar to the benchmark in (Lindsay et al. 2016). The mismatch rate is indicated at the top. Only ampliCan shows robustness to the whole range of different mismatch rates and mutational efficiencies.

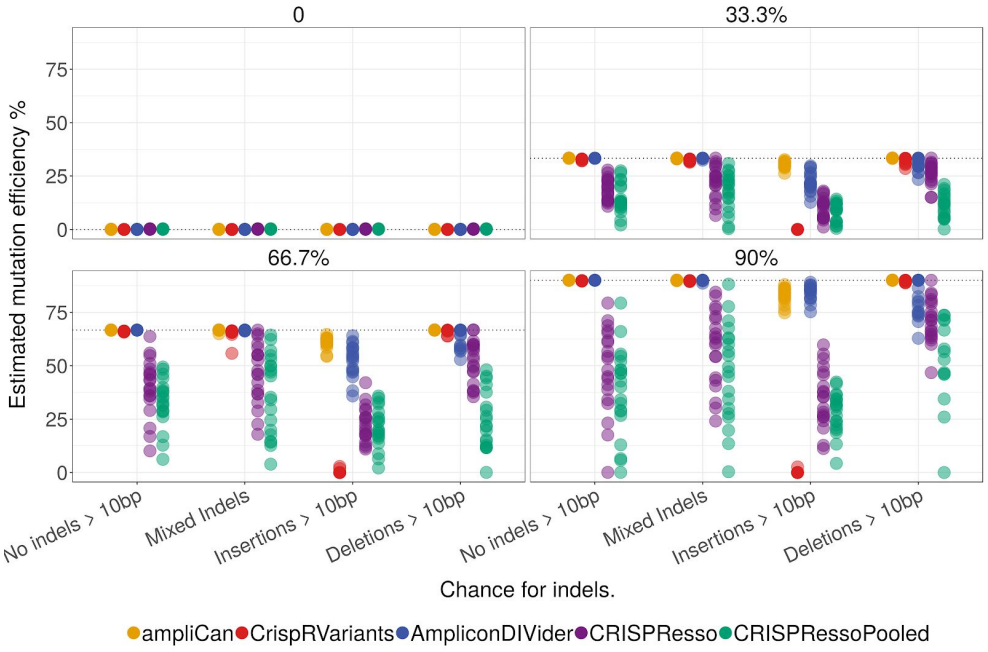


Supplemental Fig S9. Fraction of reads having indels greater than 10 bp across 176 experiments. Each dot represents one experiment and are grouped in rows by having the same gRNA

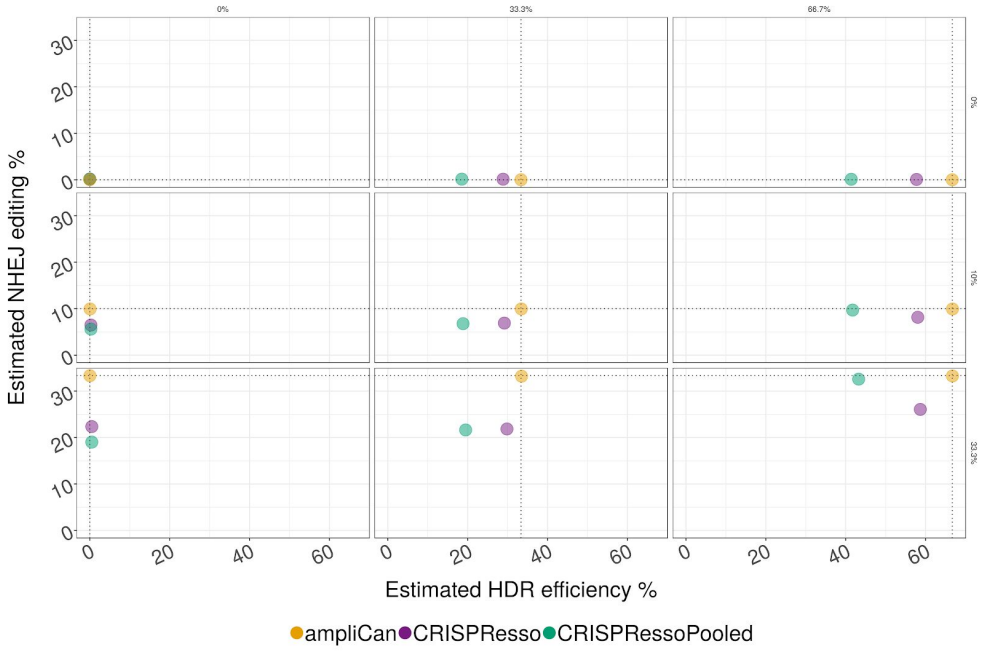
(replicates). All experiments are normalized using wild type controls ensuring that these are real indel events. The higher mean for some of the replicated experiments suggests that some gRNAs have a higher chance of resulting in long indels.



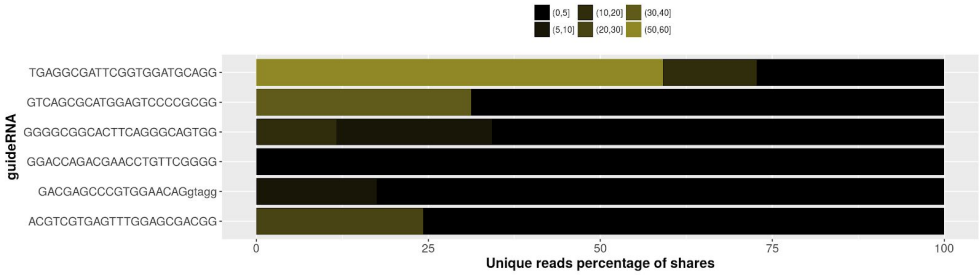
Supplemental Fig S10. Distribution of insertions width for lentiviral samples on Chari et al. 2015 datasets. Shows the proportion (y-axis) of reads with a given insert length (x-axis). The insert shows a population of unintended larger insertions. Around 90% of these originate from the lentiviral vector used to transfect the cells.



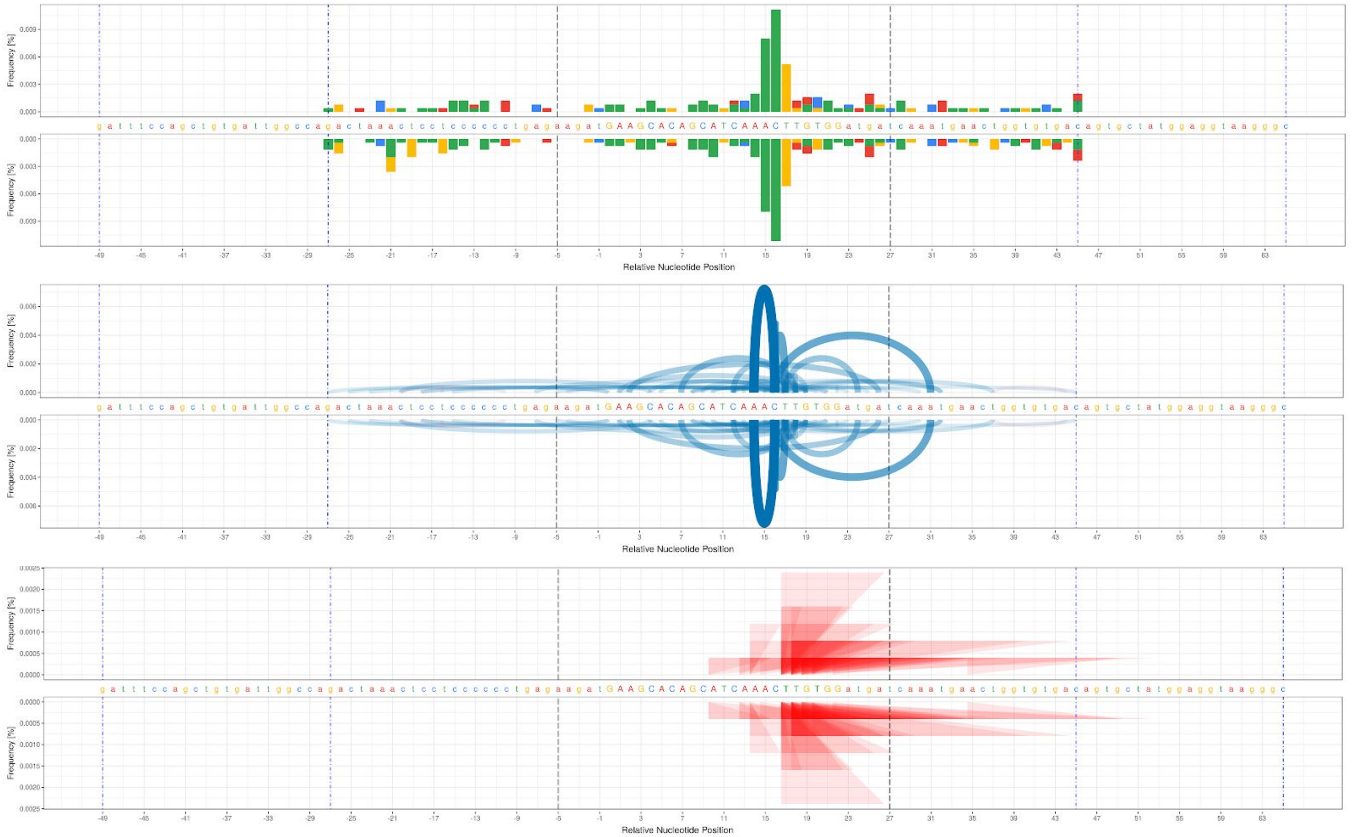
Supplemental Fig S11. Performance of leading tools grouped on simulated data with large indels. Some tools find larger indels challenging. ampliCan consistently returns values closest to the real mutation efficiency.



Supplemental Fig S12. Performance of ampliCan, CRISPResso and CRISPRessoPooled on simulated data with variable donor templates, variable true HDR read rate (0, 33, 66 %) and variable NHEJ editing (0, 10, 33 %). The dotted lines represents the real HDR (vertical) and NHEJ (horizontal) efficiencies and their intersection the correct estimate of both. The performance on 20 donors of different types (mismatch, insertions and deletions) and of variable length (from 5-70bp) were averaged into a single dot for visibility. While ampliCan handles donor templates with high accuracy, CRISPResso has difficulties with estimation in situations where alignment is imperfect.

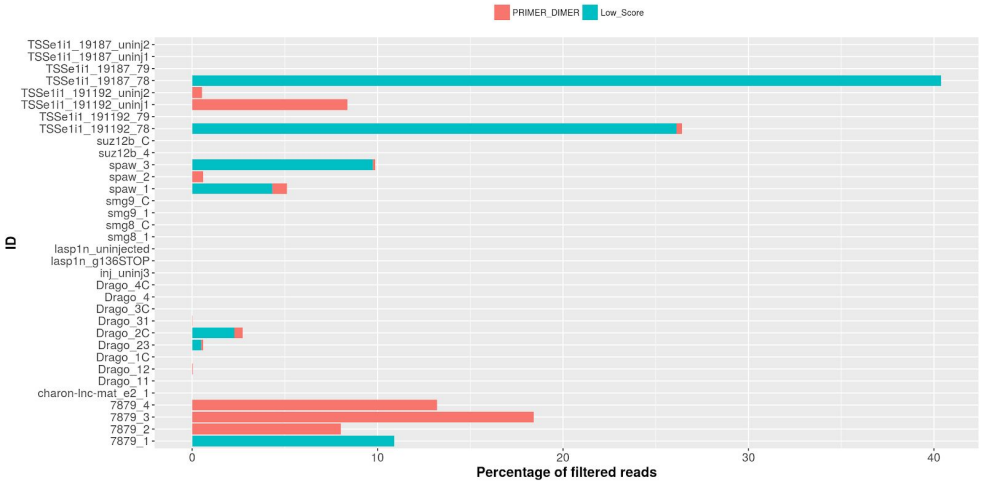


Supplemental Fig S13. Example of a heterogeneity plot produced by ampliCan. In this, identical reads are collapsed together and grouped by gRNA. A stronger shade of yellow indicates a large group of homogeneous reads. This plot can give insight into the heterogeneity of the outcome. A high level of heterogeneity can indicate sequencing problems or mosaicism. Reads can also be aggregated on experiments or any other user selected grouping rather than guides.

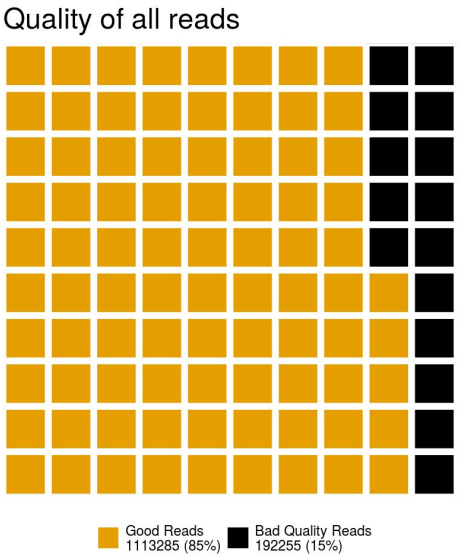


Supplemental Fig S14. A number of plots are produced by ampliCan to investigate editing events: mismatches to the reference (**A**), aggregated deletions (**B**) and aggregated insertions (**C**). Dashed blue vertical lines indicate where the primers start and end. The black

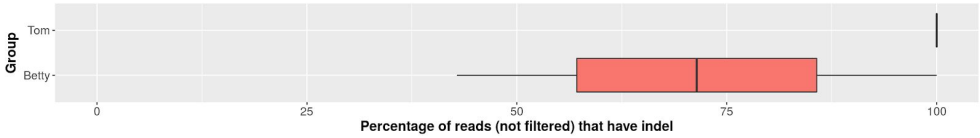
dashed lines mark the region around the cut site where we require a mutation event to overlap to be considered a CRISPR event. The top half of the plots shows the forward reads, while the bottom shows the reverse reads. The amplicon sequence is shown in the middle, with uppercase letters indicating the gRNA. The y axis shows the efficiency of the events summed over all reads. This allows for immediately discerning which bases are edited above the background noise from sequencing.



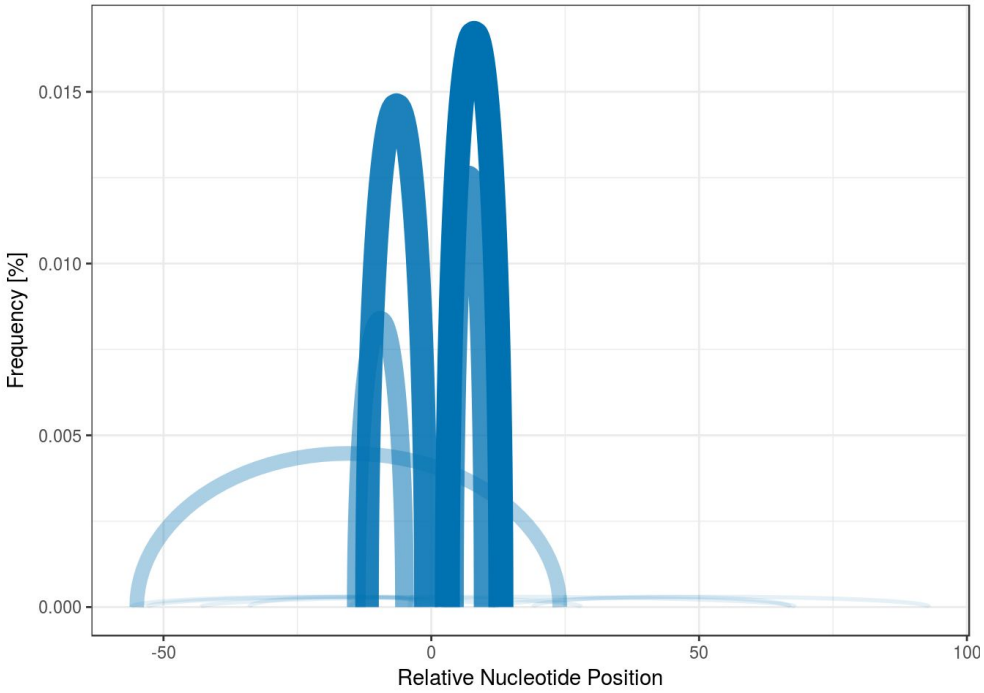
Supplemental Fig S15. Example of bar plot showing fraction of reads that were filtered out of the experiments. Red bars correspond to a primer-dimer filter, and blue bars indicate low quality reads.



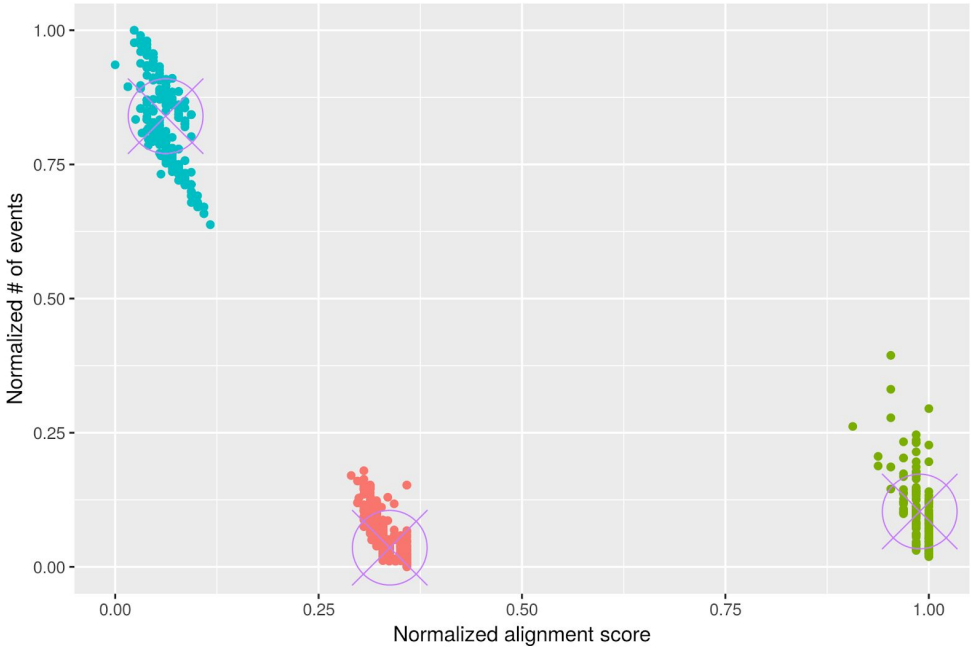
Supplemental Fig S16. Waffle plot of the quality of reads across all experiments.



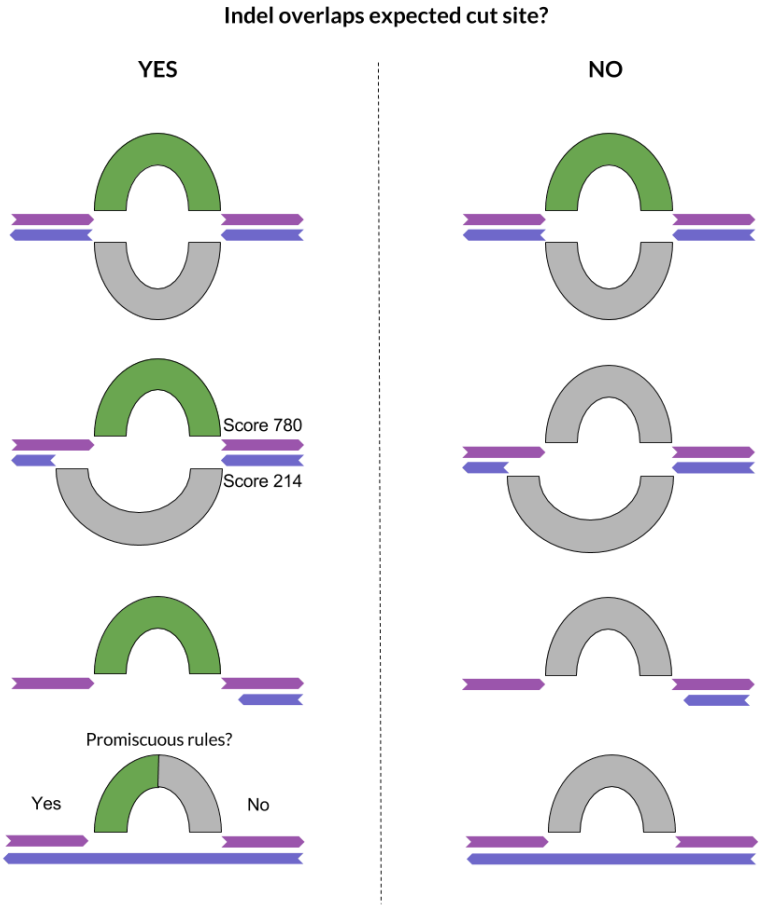
Supplemental Fig S17. Example of comparison plots in ampliCan test data where Indel rates are grouped by the researcher performing the experiments. These can be grouped on any user specified criteria.



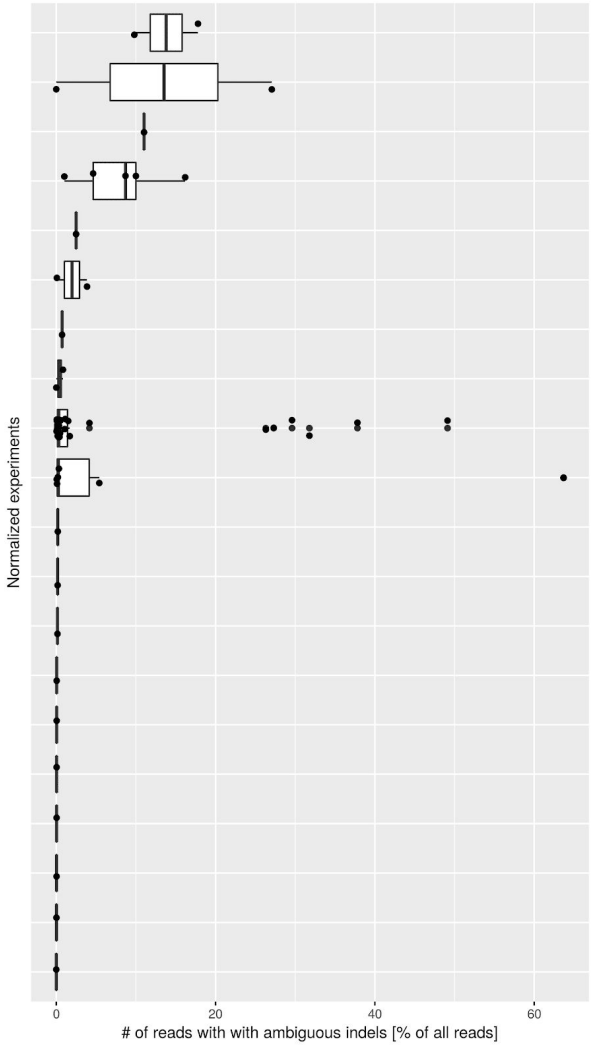
Supplemental Fig S18. Example of a deletion metaplot produced by ampliCan summarizing multiple experiments. Here, an aggregation of editing events from many experiments (many targets) using the same gRNA is presented giving an overall gRNA cut profile. Position 0 is relative to the first 5' base of gRNA.



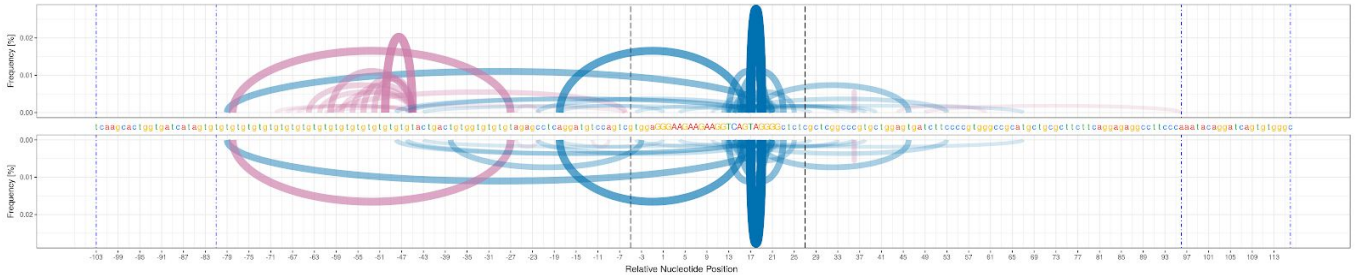
Supplemental Fig S19. Example of k-means clustering of reads during filtering of contaminant reads. A low alignment score (x axis) combined with a high number of events (y axis) indicate erroneous reads. Silhouette criterion is used to determine whether data should be clustered into two (read with no edits and reads with editing events) or three clusters (a noise cluster). In the case of three clusters, the cluster with its center (purple) closest to the upper left corner is filtered.



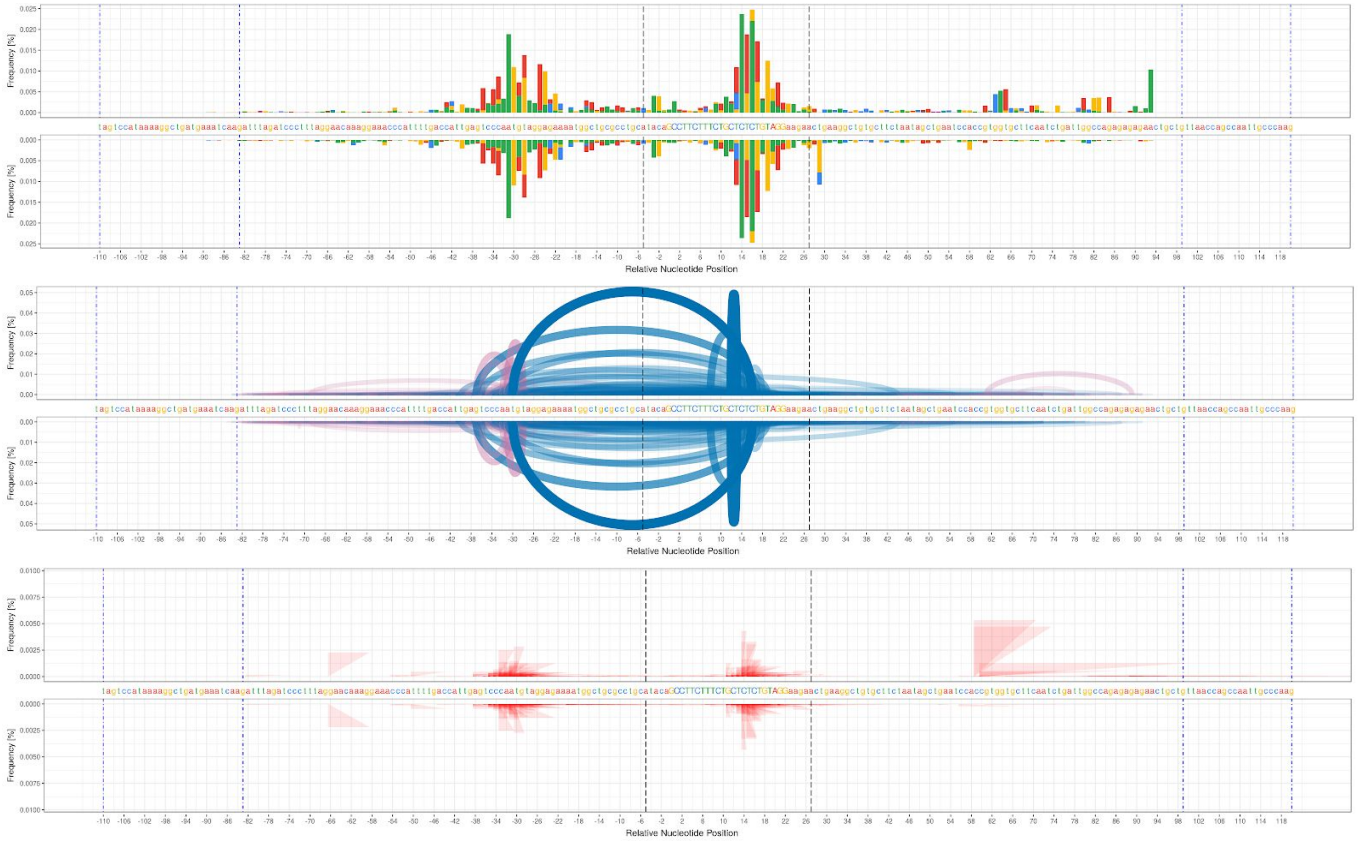
Supplemental Fig S20. The paired-end read consensus rules for ampliCan. Events marked in green are considered real cut sites while those in gray are not. When reads from forward (purple) and reverse (blue) reads are in agreement there is a consensus (top row). When two reads overlap, but disagrees the event from the strand with a higher alignment score is used (row 2). In situations where an event is only covered by one read, that read is preferred (row 3). In rare cases where there are events on one strand and the other has continuous alignment ampliCan allows users to define the behaviour (default is promiscuous rule enabled).



Supplemental Fig S21. Percentage of reads with ambiguous indels caused by disagreement of forward and reverse reads. ampliCan consensus rules help to mitigate mis-estimation that could arise from these events.

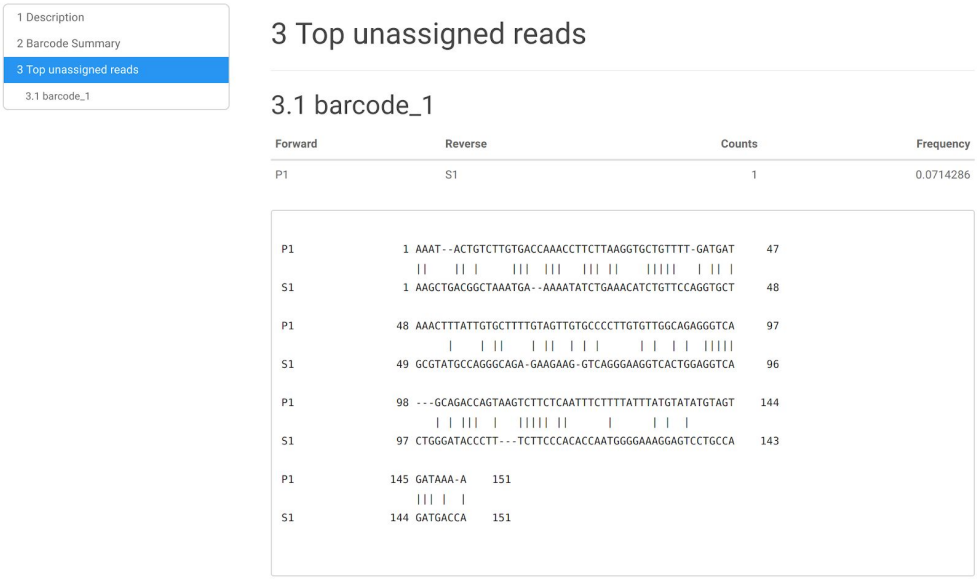


Supplemental Fig S22. ampliCan evaluates forward and reverse reads separately which allows users to spot potential problems. Here, purple colour indicates cuts outside of the gRNA expected cut site, a large majority of those reads comes from problematic alignments, where forward read and reverse read are in disagreement.

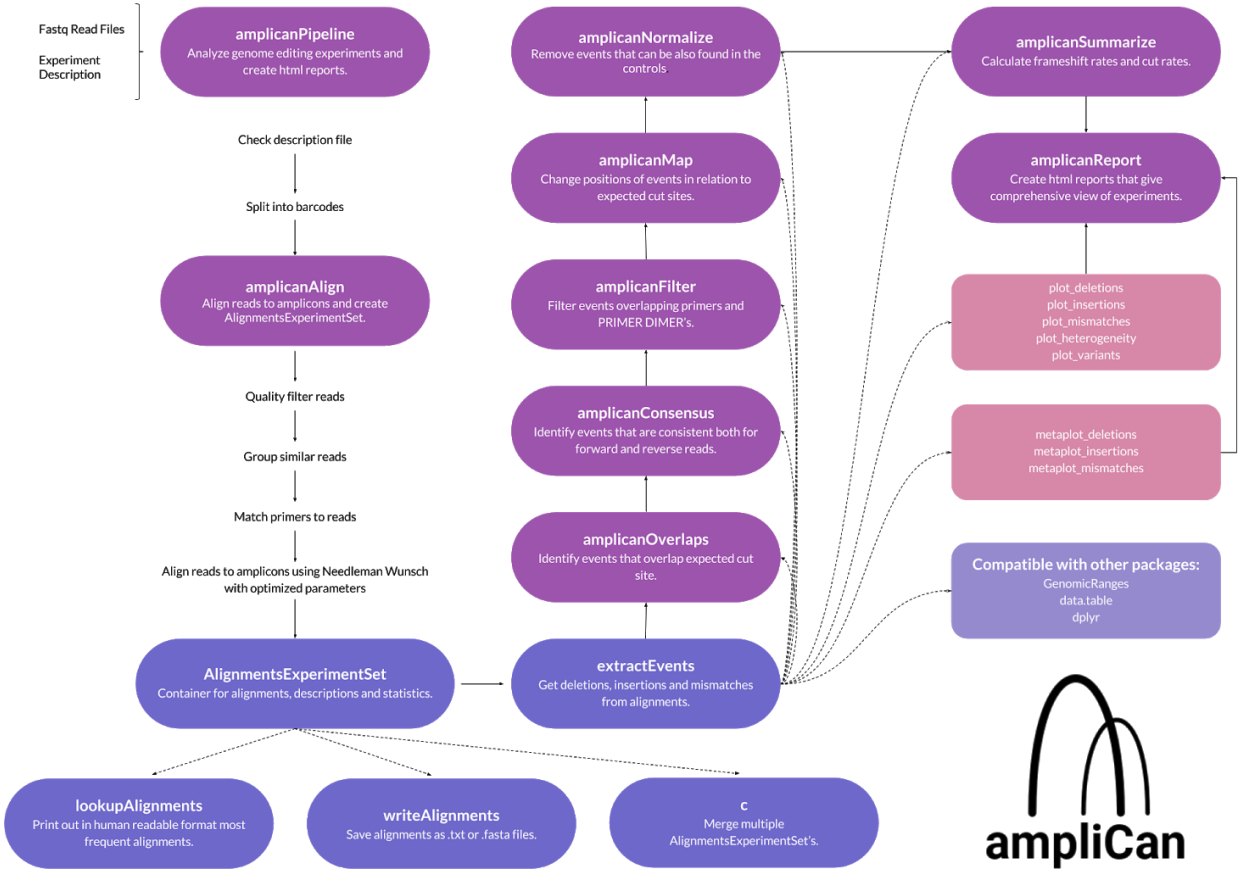


Supplemental Fig S23. Deletion plot (**top**), insertion plot (**middle**), mismatch plot (**bottom**). This presents a case where the user might have mischaracterized their own experiment which can be spotted with these plots. Here nickases were used, but only one of the gRNAs was

specified in the configuration file. A bimodal distribution of mismatch events and insertions is the result. A quick examination allows the user to realize the expected cut site should be extended to include the second gRNA, enabling more precise mutation efficiency estimation.



Supplemental Fig S24. Screenshot of the example barcode report, top unassigned read section. Human readable alignment of forward and reverse reads of the top most frequent unassigned reads is presented. Huge fragmentation and poor alignment suggest contamination, while low frequency indicate proper specification of the primers in the config file.



Supplemental Fig S25. Overview of the ampliCan pipeline.

Supplemental Tables

One edited read with variable number of total reads.		Estimated editing efficiency.		
Number of reads	True editing efficiency [%]	Edited [%] with normalization threshold 1% (default)	Edited [%] with Normalization threshold 0%	Control (noise, not edited) [%]
1	100	100	100	0
10	10	10	10	0
100	1	1.001001001	1.001001001	0
1000	0.1	0.1102093979	0.1001903617	0.0100190362
10000	0.01	0.0260724815	0.0100278775	0.016044604
100000	0.001	0.0187490099	0.0010026209	0.0177463891

Supplemental Table S1. Table displaying precision of the ampliCan when detecting editing events with variable normalization threshold. ampliCan can be used to identify extremely low frequency (0.001% and potentially lower) editing, but requires use of no normalization threshold (0%, below Illumina sequencing noise and alignment imperfections). Default normalization is set to 1% (above standard Illumina noise) and is recommended for standard applications as this setting negates the chance of removal of real editing events due to random sequencing and alignment noise and underestimating true editing efficiency at cost of small errors in precision for extremely deep sequencing.

Type	Source	Analysis	Figures
Real Dataset 1, Zebrafish 263 loci with controls	Gagnon et al. 2014 and newly generated	Comparison of discrepancies between tools.	Fig 1. and Supplemental Figs S1, S3, S4, S9, S13, S15, S18, S21, S22, S23, S24.
Real Dataset 2, Human, ~1,400 genomic loci	Chari et al. 2015	Example of large insertions in real data.	Supplemental Fig S10.
Synthetic Dataset 2, 20 loci x 4 efficiency rates x 3 off-target rates = 240 experiments	Lindsay et al. 2016	Reproducing previous benchmark.	Supplemental Fig S7.
Synthetic Dataset 3, As in Lindsay et al. 2016 x 3 mismatch rates = 720 experiments	Simulated based on Lindsay et al. 2016 strategy with variable parameters.	More comprehensive analysis of contaminant reads.	Fig 2. and Supplemental Figs S8, S19.
Synthetic Dataset 4, 20 loci x 4 efficiency rates x 4 types of indels = 320 experiments		Analysis of indel size influence on the efficiency estimates.	Fig 2. And Supplemental Fig S11.
Synthetic Dataset 5, 20 loci x 3 NHEJ rates x 3 HDR rates x 3 donor types = 540 experiments		Homology Directed Repair	Supplemental Fig S12.
Synthetic Dataset 1, 1 loci x 6 efficiency rates x 3 normalization thresholds = 18 experiments		Precision of ampliCan	Supplemental Tab S8.

Supplemental Table S2. Table displays all datasets used and their origin.

Tool	Aligner	Notes
ampliCan	Needleman- Wunsch from Biostrings (Pages et al.) as local alignment	
CrispRVariant (Lindsay et al. 2016)	BWA-MEM(Li 2013) as global alignment	Filters primer dimers by restricting start/end of indel, but may result in missing larger deletions. Larger insertions can be missed through mapping.
ampliconDIVider (Varshney et al. 2015) (ampliconDIV_minimal.sh)	BWA-MEM as global alignment	Sometimes returns estimates above 100%, these values were filtered. We run only the variant counting step of the pipeline. The full pipeline requires the commercial novoAlign.
CRISPResso & CRISPRessoPooled (Pinello et al. 2016)	BWA-MEM + FLASH (in Pooled) as local alignment	Poor performance due to unknown issues. It was installed and run in the same fashion as in Lindsay et al. 2016 analysis. CRISPRessoPooled uses local-global alignment strategy.

Supplemental Table S3. Table of compared tools, with align strategy.

```
ID: ID_1 read_id: 1 Count: 3
AAGCTGACGGCTAAATGAAAAATGTCAAACGTCTGTTCCAG------------------------------------------------------------------------AAAAAAAAAAAAAAAAAAAATTCCCACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCAGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT
AAGCTGACGGCTAAATGAAAAATGTCAAACATCTGTTCCAGGTGCTGCGTATGCCAGGGCAGAGGAGGTGGTCAGGGAACTGGTGGAGGTCACTGGGATACCCTTTC-------------------TTCCCACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCTGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT

AAGCTGACGGCTAAATGAAAAATGTCAAACGTCTGTTCCAG------------------------------------------------------------------------AAAAAAAAAAAAAAAAAAAATTCCCACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCAGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT
AAGCTGACGGCTAAATGAAAAATGTCAAACATCTGTTCCAGGTGCTGCGTATGCCAGGGCAGAGGAGGTGGTCAGGGAACTGGTGGAGGTCACTGGGATACCCTTTC-------------------TTCCCACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCTGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT

ID: ID_1 read_id: 2 Count: 2
AAGCTGACGGCTAAATGAAAAATGTCAAACGTCTGTT----------------------------------------------------------------------------AAAAAAAAAAAAAAAAAAAACACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCAGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT
AAGCTGACGGCTAAATGAAAAATGTCAAACATCTGTTCCAGGTGCTGCGTATGCCAGGGCAGAGGAGGTGGTCAGGGAACTGGTGGAGGTCACTGGGATACCCTTTCTTCC-------------------CACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCTGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT

AAGCTGACGGCTAAATGAAAAATGTCAAACGTCTGTT----------------------------------------------------------------------------AAAAAAAAAAAAAAAAAAAACACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCAGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT
AAGCTGACGGCTAAATGAAAAATGTCAAACATCTGTTCCAGGTGCTGCGTATGCCAGGGCAGAGGAGGTGGTCAGGGAACTGGTGGAGGTCACTGGGATACCCTTTCTTCC-------------------CACACCAATGGGGAAAGGAGTCCTGCCAGATGACCATCC
CAACTGTGTTGCTGCAGCCAGATCCAGGTGTGTTTGCGCTTGTGTAATT
```

Supplemental Table S4. Example of human readable output. Aligned reads are assigned to the experiment (ID, read_id) and sorted based on count (Count). For each pair alignment is presented with top part representing forward read aligned to amplicon and bottom presenting reverse read aligned to amplicon.

seqnames	start	end	width	strand	originally	replacement	type	read_id	score	counts
ID_1	108	127	20	+		AAAAAAAAAAAAAAAAAAAA	insertion	1	597	3
ID_1	112	131	20	+		AAAAAAAAAAAAAAAAAAAA	insertion	2	557	2
ID_1	42	107	66	+			deletion	1	597	3
ID_1	38	111	74	+			deletion	2	557	2
ID_1	24	173	150	+			deletion	3	193	1
ID_1	34	117	84	+			deletion	4	532	1
ID_1	31	31	1	+	A	G	mismatch	1	597	3
ID_1	163	163	1	+	T	A	mismatch	1	597	3
ID_1	31	31	1	+	A	G	mismatch	2	557	2
ID_1	163	163	1	+	T	A	mismatch	2	557	2
ID_1	23	23	1	+	T	G	mismatch	3	193	1
ID_1	176	176	1	+	A	T	mismatch	3	193	1
ID_1	177	177	1	+	G	C	mismatch	3	193	1
ID_1	31	31	1	+	A	G	mismatch	4	532	1
ID_1	171	171	1	+	G	A	mismatch	4	532	1
ID_1	108	127	20	-		AAAAAAAAAAAAAAAAAAAA	insertion	1	597	3
ID_1	112	131	20	-		AAAAAAAAAAAAAAAAAAAA	insertion	2	557	2
ID_1	42	107	66	-			deletion	1	597	3
ID_1	38	111	74	-			deletion	2	557	2
ID_1	24	173	150	-			deletion	3	193	1
ID_1	34	117	84	-			deletion	4	532	1
ID_1	31	31	1	-	A	G	mismatch	1	597	3
ID_1	163	163	1	-	T	A	mismatch	1	597	3
ID_1	31	31	1	-	A	G	mismatch	2	557	2
ID_1	163	163	1	-	T	A	mismatch	2	557	2
ID_1	23	23	1	-	T	G	mismatch	3	193	1
ID_1	176	176	1	-	A	T	mismatch	3	193	1
ID_1	177	177	1	-	G	C	mismatch	3	193	1
ID_1	31	31	1	-	A	G	mismatch	4	532	1
ID_1	171	171	1	-	G	A	mismatch	4	532	1
ID_2	115	127	13	-			deletion	1	845	3
ID_2	106	114	9	-			deletion	2	865	2
ID_2	101	105	5	-			deletion	3	885	1
ID_2	115	127	13	+			deletion	1	845	3
ID_2	106	114	9	+			deletion	2	865	2
ID_2	101	105	5	+			deletion	3	885	1
ID_3	171	171	1	+		T	insertion	1	819	5
ID_3	74	89	16	+		CCCCCCCCCCCCCCCC	insertion	3	860	1
ID_3	66	83	18	+			deletion	1	819	5

Supplemental Table S5. Example GenomicRanges table output with additional meta-columns. This representation of alignments allows for efficient manipulation and processing of the data.

ID	Barcode	Forward_Reads	Reverse_Reads	Group	Control	guideRNA	Forward_Primer	Reverse_Primer	Direction	Amplicon
ID_1	barcode_1	R1_001.fastq	R2_001.fastq	Betty	0	AGGTGGTCAGGGAACTGG	AAGCTGACGGCTAAATGA	AATTACACAAGCGCAAACACAC	0	...cagaggAGGTGGTCAGGGAACTGGtgga...
ID_2	barcode_1	R1_001.fastq	R2_001.fastq	Tom	0	TGACCCTCTGCCAACACAAGGGG	TGACCAAACCTTCTTAAGGTGC	CTCTGCTGCAAAATGCAAGG	1	...agtgtgCCCCTTGTGTTGGCAGAGGGTCAg...
ID_3	barcode_2	R1_002.fastq	R2_002.fastq	Tom	0	AGGTGGTCAGGGAACTGG	AAGCTGACGGCTAAATGA	AATTACACAAGCGCAAACACAC	0	...ggAGGTGGTCAGGGAACTGGtgga...
ID_4	barcode_2	R1_002.fastq	R2_002.fastq	Betty	0	GTCCCTGCAACATTAAAGGCCGG	GCTGGCAACATTCCTACCAGT	GAGCGCTGAGGCAGGATTAT	0	...ccaGTCCCTGCAACATTAAAGGCCGGaag...

Supplemental Table S6. Example ampliCan config file. ampliCan requires this file as minimal input, together with relevant fastq files. More precise, up to date description of the file can be found in the ampliCan vignettes.

References

Chari R, Mali P, Moosburner M, Church GM. 2015. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* **12**: 823–826.

Gagnon JA, Valen E, Thyme SB, Huang P, Akhmetova L, Ahkmetova L, Pauli A, Montague TG, Zimmerman S, Richter C, et al. 2014. Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* **9**: e98186.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**: 593–594.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bioGN]*. http://arxiv.org/abs/1303.3997.

Lindsay H, Burger A, Biyong B, Felker A, Hess C, Zaugg J, Chiavacci E, Anders C, Jinek M, Mosimann C, et al. 2016. CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nat Biotechnol* **34**: 701–702.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. 2009. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**: 2607–2608.

Pages H, Gentleman R, Aboyoun P, DebRoy S. Biostrings: String objects representing biological sequences, and matching algorithms, 2008. *R package version* **2**: 160.

Pinello L, Canver MC, Hoban MD, Orkin SH, Kohn DB, Bauer DE, Yuan G-C. 2016. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat Biotechnol* **34**: 695–697.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biol* **14**: R51.

Shah AN, Davey CF, Whitebirch AC, Miller AC, Moens CB. 2016. Rapid Reverse Genetic

Screening Using CRISPR in Zebrafish. *Zebrafish* **13**: 152–153.

Varshney GK, Pei W, LaFave MC, Idol J, Xu L, Gallardo V, Carrington B, Bishop K, Jones M, Li M, et al. 2015. High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res* **25**: 1030–1042.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Minimizing Index Hopping.
https://www.illumina.com/science/education/minimizing-index-hopping.html (Accessed January 22, 2019).

7. Conclusions and future perspectives

Scientific software should be intuitive to use by a group of researchers that do not need to understand all of the details, yet at the same time provide opportunities for optimization by seasoned experts. Scientific tools should also be frequently updated to match the speed of the developing field of genome editing. I believe I adhered to this goal with both CHOPCHOP and ampliCan, which are designed to contain current advances in the field. At the same time, it cannot be stressed enough that the usefulness of even the most wonderful tool can be hindered by proprietary rights, therefore all programs presented in this thesis are completely free to use by academic and non-profit users.

In the era of a reproducibility crisis in science, state-of-the-art bioinformatics tools can gain quality with thorough benchmarks and continuous updates. Currently, the field of designing CRISPR gRNAs is quickly developing. Novel methods are published as feature-incomplete tools. Meanwhile, CHOPCHOP tries to include recent advances with continuous updates, not satisfied with just maintenance.

Ideally, CHOPCHOP would be reimplemented in the future with better memory usage and multi-core support. CHOPCHOP could be further improved with integration of precomputed databases of tested guideRNAs e.g. Brunello (Sanson et al. 2018). Definitive improvements to the off-target search could also be made, for instance, integration of SNP databases for better awareness of genome heterogeneity (Haeussler et al. 2016). Allowing for more than 3 mismatch off-target searches, as well as searches with bulges requires a specialized aligner (Bae, Park, and Kim 2014), and together with the latest off-target scoring algorithms would elucidate more comprehensive off-target searches than currently offered. In the future, more robust models for gRNA efficiency and repair profile predictions will become available. It will become easier to integrate a new machine learning models as the field of machine learning becomes more standardized and models become more transferable between languages and package versions.

Analysis of the CRISPR edited experiments, especially from amplicon sequencing data, has proven to be more challenging than initially expected. ampliCan was created to remedy those issues and ensure precision of estimated editing efficiencies. ampliCan itself is advocating for event level resolution that builds a framework which could be potentially used in other scientific fields. For instance, extraction of events per read is sought for in nanopore sequencing to detect modified bases, but currently only mismatch pileups on the genomic/transcriptomic level are available (Rand et al. 2017; Simpson et al. 2017). The ampliCan methodology could potentially be used to cover alignments to the genome, not only to the amplicons, further expanding the range of possible applications.

In the future, when computational power is less of an issue, multiple sequence alignments (amplicon sequences aligned with forward and reverse reads together) might be an improvement to the current alignment strategy of ampliCan. Multiple sequence alignment has high potential to solve issues related to mapping, as even with optimized alignments, ampliCan can still struggle with some reads with longer indels. When this happens or the field develops further, we can also extend the benchmarks to compare additional tools such as CRISPR-DAV (Xuning Wang et al. 2017), CRISPR-GA (Güell, Yang, and Church 2014), BATCH-GE (Boel et al. 2016) or CRIS.py (Connelly and Pruett-Miller 2019).

In summary, bioinformatics tools need to accompany scientific progress. Along these lines, my scientific contribution has been the creation of tools to facilitate researchers' use of CRISPR. I hope that the software that I have developed will serve scientists across the world, and indirectly contribute to our understanding of biological life.

References

Aarts, Marieke, and Hein te Riele. 2010. “Subtle Gene Modification in Mouse ES Cells: Evidence for Incorporation of Unmodified Oligonucleotides without Induction of DNA Damage.” *Nucleic Acids Research* 38 (20): 6956–67.

Abadi, Shiran, Winston X. Yan, David Amar, and Itay Mayrose. 2017. “A Machine Learning Approach for Predicting CRISPR-Cas9 Cleavage Efficiencies and Patterns Underlying Its Mechanism of Action.” *PLoS Computational Biology* 13 (10): e1005807.

Abudayyeh, Omar O., Jonathan S. Gootenberg, Patrick Essletzbichler, Shuo Han, Julia Joung, Joseph J. Belanto, Vanessa Verdine, et al. 2017. “RNA Targeting with CRISPR-Cas13.” *Nature*, October. https://doi.org/10.1038/nature24049.

Abudayyeh, Omar O., Jonathan S. Gootenberg, Silvana Konermann, Julia Joung, Ian M. Slaymaker, David B. T. Cox, Sergey Shmakov, et al. 2016. “C2c2 Is a Single-Component Programmable RNA-Guided RNA-Targeting CRISPR Effector.” *Science* 353 (6299): aaf5573.

Bae, Sangsu, Jeongbin Park, and Jin-Soo Kim. 2014. “Cas-OFFinder: A Fast and Versatile Algorithm That Searches for Potential off-Target Sites of Cas9 RNA-Guided Endonucleases.” *Bioinformatics* 30 (10): 1473–75.

Boel, Annekatrien, Woutert Steyaert, Nina De Rocker, Björn Menten, Bert Callewaert, Anne De Paepe, Paul Coucke, and Andy Willaert. 2016. “BATCH-GE: Batch Analysis of Next-Generation Sequencing Data for Genome Editing Assessment.” *Scientific Reports* 6 (July): 30330.

Bolotin, Alexander, Benoit Quinquis, Alexei Sorokin, and S. Dusko Ehrlich. 2005. “Clustered Regularly Interspaced Short Palindrome Repeats (CRISPRs) Have Spacers of Extrachromosomal Origin.” *Microbiology* 151 (Pt 8): 2551–61.

Bradford, J., and D. Perrin. 2018. “A Benchmark of Computational CRISPR-Cas9 Guide Design Methods.” *BioRxiv*. https://www.biorxiv.org/content/10.1101/498782v1.abstract.

Brouns, Stan J. J., Matthijs M. Jore, Magnus Lundgren, Edze R. Westra, Rik J. H. Slijkhuis, Ambrosius P. L. Snijders, Mark J. Dickman, Kira S. Makarova, Eugene V. Koonin, and John van der Oost. 2008. “Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes.” *Science* 321 (5891): 960–64.

Cameron, Peter, Chris K. Fuller, Paul D. Donohoue, Brittnee N. Jones, Matthew S. Thompson, Matthew M. Carter, Scott Gradia, et al. 2017. “Mapping the Genomic Landscape of CRISPR-Cas9 Cleavage.” *Nature Methods* 14 (6): 600–606.

Chari, Raj, Prashant Mali, Mark Moosburner, and George M. Church. 2015. “Unraveling CRISPR-Cas9 Genome Engineering Parameters via a Library-on-Library Approach.” *Nature Methods* 12 (9): 823–26.

Chen, Yue, Bruce A. McClane, Derek J. Fisher, Julian I. Rood, and Phalguni Gupta. 2005. “Construction of an Alpha Toxin Gene Knockout Mutant of Clostridium Perfringens Type A by Use of a Mobile Group II Intron.” *Applied and Environmental Microbiology* 71 (11): 7542–47.

Clement, Kendell, Holly Rees, Matthew C. Canver, Jason M. Gehrke, Rick Farouni, Jonathan Y. Hsu, Mitchel A. Cole, et al. 2019. “CRISPResso2 Provides Accurate

and Rapid Genome Editing Sequence Analysis." *Nature Biotechnology* 37 (3): 224–26.

Cong, Le, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, et al. 2013. "Multiplex Genome Engineering Using CRISPR/Cas Systems." *Science* 339 (6121): 819–23.

Connelly, Jon P., and Shondra M. Pruett-Miller. 2019. "CRIS.py: A Versatile and High-Throughput Analysis Program for CRISPR-Based Genome Editing." *Scientific Reports* 9 (1): 4194.

Cui, Yingbo, Jiaming Xu, Minxia Cheng, Xiangke Liao, and Shaoliang Peng. 2018. "Review of CRISPR/Cas9 sgRNA Design Tools." *Interdisciplinary Sciences, Computational Life Sciences*, April. https://doi.org/10.1007/s12539-018-0298-z.

Deltcheva, Elitza, Krzysztof Chylinski, Cynthia M. Sharma, Karine Gonzales, Yanjie Chao, Zaid A. Pirzada, Maria R. Eckert, Jörg Vogel, and Emmanuelle Charpentier. 2011. "CRISPR RNA Maturation by Trans-Encoded Small RNA and Host Factor RNase III." *Nature* 471 (7340): 602–7.

Doench, John G., Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F. Donovan, Ian Smith, et al. 2016. "Optimized sgRNA Design to Maximize Activity and Minimize off-Target Effects of CRISPR-Cas9." *Nature Biotechnology* 34 (2): 184–91.

Doench, John G., Ella Hartenian, Daniel B. Graham, Zuzana Tothova, Mudra Hegde, Ian Smith, Meagan Sullender, Benjamin L. Ebert, Ramnik J. Xavier, and David E. Root. 2014. "Rational Design of Highly Active sgRNAs for CRISPR-Cas9-Mediated Gene Inactivation." *Nature Biotechnology* 32 (12): 1262–67.

Donoho, G., M. Jasin, and P. Berg. 1998. "Analysis of Gene Targeting and Intrachromosomal Homologous Recombination Stimulated by Genomic Double-Strand Breaks in Mouse Embryonic Stem Cells." *Molecular and Cellular Biology* 18 (7): 4070–78.

Friedland, Ari E., Reshica Baral, Pankhuri Singhal, Katherine Loveluck, Shen Shen, Minerva Sanchez, Eugenio Marco, et al. 2015. "Characterization of Staphylococcus Aureus Cas9: A Smaller Cas9 for All-in-One Adeno-Associated Virus Delivery and Paired Nickase Applications." *Genome Biology* 16 (November): 257.

Gabrieli, Tslil, Hila Sharim, Dena Fridman, Nissim Arbib, Yael Michaeli, and Yuval Ebenstein. 2018. "Selective Nanopore Sequencing of Human BRCA1 by Cas9-Assisted Targeting of Chromosome Segments (CATCH)." *Nucleic Acids Research* 46 (14): e87.

Garneau, Josiane E., Marie-Ève Dupuis, Manuela Villion, Dennis A. Romero, Rodolphe Barrangou, Patrick Boyaval, Christophe Fremaux, Philippe Horvath, Alfonso H. Magadán, and Sylvain Moineau. 2010. "The CRISPR/Cas Bacterial Immune System Cleaves Bacteriophage and Plasmid DNA." *Nature* 468 (7320): 67–71.

Gasiunas, Giedrius, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. 2012. "Cas9-crRNA Ribonucleoprotein Complex Mediates Specific DNA Cleavage for Adaptive Immunity in Bacteria." *Proceedings of the National Academy of Sciences of the United States of America* 109 (39): E2579–86.

Gaudelli, Nicole M., Alexis C. Komor, Holly A. Rees, Michael S. Packer, Ahmed H.

Badran, David I. Bryson, and David R. Liu. 2017. “Programmable Base Editing of A•T to G•C in Genomic DNA without DNA Cleavage.” *Nature* 551 (October): 464.

Goff, Stephen P., and Paul Berg. 1976. “Construction of Hybrid Viruses Containing SV40 and λ Phage DNA Segments and Their Propagation in Cultured Monkey Cells.” *Cell* 9 (4, Part 2): 695–705.

Gootenberg, Jonathan S., Omar O. Abudayyeh, Max J. Kellner, Julia Joung, James J. Collins, and Feng Zhang. 2018. “Multiplexed and Portable Nucleic Acid Detection Platform with Cas13, Cas12a, and Csm6.” *Science* 360 (6387): 439–44.

Gootenberg, Jonathan S., Omar O. Abudayyeh, Jeong Wook Lee, Patrick Essletzbichler, Aaron J. Dy, Julia Joung, Vanessa Verdine, et al. 2017. “Nucleic Acid Detection with CRISPR-Cas13a/C2c2.” *Science*, April, eaam9321.

Güell, Marc, Luhan Yang, and George M. Church. 2014. “Genome Editing Assessment Using CRISPR Genome Analyzer (CRISPR-GA).” *Bioinformatics* 30 (20): 2968–70.

Haeussler, Maximilian, Kai Schönig, Hélène Eckert, Alexis Eschstruth, Joffrey Mianné, Jean-Baptiste Renaud, Sylvie Schneider-Maunoury, et al. 2016. “Evaluation of off-Target and on-Target Scoring Algorithms and Integration into the Guide RNA Selection Tool CRISPOR.” *Genome Biology* 17 (1): 148.

Heigwer, Florian, Grainne Kerr, and Michael Boutros. 2014. “E-CRISP: Fast CRISPR Target Site Identification.” *Nature Methods* 11 (January): 122.

Hille, Frank, Hagen Richter, Shi Pey Wong, Majda Bratovič, Sarah Ressel, and Emmanuelle Charpentier. 2018. “The Biology of CRISPR-Cas: Backward and Forward.” *Cell* 172 (6): 1239–59.

Horlbeck, Max A., Lea B. Witkowsky, Benjamin Guglielmi, Joseph M. Replogle, Luke A. Gilbert, Jacqueline E. Villalta, Sharon E. Torigoe, Robert Tjian, and Jonathan S. Weissman. 2016. “Nucleosomes Impede Cas9 Access to DNA in Vivo and in Vitro.” *eLife* 5 (March). https://doi.org/10.7554/eLife.12677.

Hsu, Patrick D., Eric S. Lander, and Feng Zhang. 2014. “Development and Applications of CRISPR-Cas9 for Genome Engineering.” *Cell* 157 (6): 1262–78.

Jeltsch, A., C. Wenz, W. Wende, U. Selent, and A. Pingoud. 1996. “Engineering Novel Restriction Endonucleases: Principles and Applications.” *Trends in Biotechnology* 14 (7): 235–38.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.” *Science* 337 (6096): 816–21.

Kato, T., R. H. Rothman, and A. J. Clark. 1977. “Analysis of the Role of Recombination and Repair in Mutagenesis of Escherichia Coli by UV Irradiation.” *Genetics*. https://www.genetics.org/content/87/1/1.short.

Kim, Hui Kwon, Seonwoo Min, Myungjae Song, Soobin Jung, Jae Woo Choi, Younggwang Kim, Sangeun Lee, Sungroh Yoon, and Hyongbum (henry) Kim. 2018. “Deep Learning Improves Prediction of CRISPR–Cpf1 Guide RNA Activity.” *Nature Biotechnology* 36 (January): 239.

Kim, Y. G., J. Cha, and S. Chandrasegaran. 1996. “Hybrid Restriction Enzymes: Zinc Finger Fusions to Fok I Cleavage Domain.” *Proceedings of the National*

Academy of Sciences of the United States of America 93 (3): 1156–60.
Komor, Alexis C., Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. 2016. "Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage." *Nature* 533 (7603): 420–24.
Kulcsár, Péter István, András Tálas, Krisztina Huszár, Zoltán Ligeti, Eszter Tóth, Nóra Weinhardt, Elfrieda Fodor, and Ervin Welker. 2017. "Crossing Enhanced and High Fidelity SpCas9 Nucleases to Optimize Specificity and Cleavage." *Genome Biology* 18 (1): 190.
Kyrou, Kyros, Andrew M. Hammond, Roberto Galizi, Nace Kranjc, Austin Burt, Andrea K. Beaghton, Tony Nolan, and Andrea Crisanti. 2018. "A CRISPR–Cas9 Gene Drive Targeting Doublesex Causes Complete Population Suppression in Caged Anopheles Gambiae Mosquitoes." *Nature Biotechnology* 36 (September): 1062.
Labun, Kornel, Xiaoge Guo, Alejandro Chavez, George Church, James A. Gagnon, and Eivind Valen. 2019. "Accurate Analysis of Genuine CRISPR Editing Events with ampliCan." *Genome Research*, March. https://doi.org/10.1101/gr.244293.118.
Labun, Kornel, Tessa G. Montague, James A. Gagnon, Summer B. Thyme, and Eivind Valen. 2016. "CHOPCHOP v2: A Web Tool for the next Generation of CRISPR Genome Engineering." *Nucleic Acids Research* 44 (W1): W272–76.
Labun, Kornel, Tessa G. Montague, Maximilian Krause, Yamila N. Torres Cleuren, Håkon Tjeldnes, and Eivind Valen. 2019. "CHOPCHOP v3: Expanding the CRISPR Web Toolbox beyond Genome Editing." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz365.
Lackner, Daniel H., Alexia Carré, Paloma M. Guzzardo, Carina Banning, Ramu Mangena, Tom Henley, Sarah Oberndorfer, et al. 2015. "A Generic Strategy for CRISPR-Cas9-Mediated Gene Tagging." *Nature Communications* 6 (December): 10237.
Lander, Eric S. 2016. "The Heroes of CRISPR." *Cell* 164 (1-2): 18–28.
Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25.
Lei, Y., L. Lu, H. Y. Liu, S. Li, F. Xing, and L. L. Chen. 2014. "CRISPR-P: A Web Tool for Synthetic Single-Guide RNA Design of CRISPR-System in Plants." *Molecular Plant*. https://www.cell.com/molecular-plant/pdf/S1674-2052(14)60952-7.pdf.
Lindsay, Helen, Alexa Burger, Berthin Biyong, Anastasia Felker, Christopher Hess, Jonas Zaugg, Elena Chiavacci, et al. 2016. "CrispRVariants Charts the Mutation Spectrum of Genome Engineering Experiments." *Nature Biotechnology* 34 (7): 701–2.
Listgarten, Jennifer, Michael Weinstein, Benjamin P. Kleinstiver, Alexander A. Sousa, J. Keith Joung, Jake Crawford, Kevin Gao, et al. 2018. "Prediction of off-Target Activities for the End-to-End Design of CRISPR Guide RNAs." *Nature Biomedical Engineering* 2 (1): 38–47.
Liu, Hao, Yuduan Ding, Yanqing Zhou, Wenqi Jin, Kabin Xie, and Ling-Ling Chen. 2017. "CRISPR-P 2.0: An Improved CRISPR-Cas9 Tool for Genome Editing in

Plants." *Molecular Plant* 10 (3): 530–32.
Liu, Jun-Jie, Natalia Orlova, Benjamin L. Oakes, Enbo Ma, Hannah B. Spinner, Katherine L. M. Baney, Jonathan Chuck, et al. 2019. "CasX Enzymes Comprise a Distinct Family of RNA-Guided Genome Editors." *Nature* 566 (7743): 218–23.
Makarova, Kira S., and Eugene V. Koonin. 2013. "Evolution and Classification of CRISPR-Cas Systems and Cas Protein Families." *CRISPR-Cas Systems*. https://doi.org/10.1007/978-3-662-45794-8_3.
Makarova, K. S., Y. I. Wolf, and E. V. Koonin. 2018. "Classification and Nomenclature of CRISPR-Cas Systems: Where from Here?" *The CRISPR Journal*. https://www.liebertpub.com/doi/abs/10.1089/crispr.2018.0033.
Mali, Prashant, John Aach, P. Benjamin Stranges, Kevin M. Esvelt, Mark Moosburner, Sriram Kosuri, Luhan Yang, and George M. Church. 2013. "CAS9 Transcriptional Activators for Target Specificity Screening and Paired Nickases for Cooperative Genome Engineering." *Nature Biotechnology* 31 (9): 833–38.
Mali, Prashant, Luhan Yang, Kevin M. Esvelt, John Aach, Marc Guell, James E. DiCarlo, Julie E. Norville, and George M. Church. 2013. "RNA-Guided Human Genome Engineering via Cas9." *Science* 339 (6121): 823–26.
Miller, Jeffrey C., Siyuan Tan, Guijuan Qiao, Kyle A. Barlow, Jianbin Wang, Danny F. Xia, Xiangdong Meng, et al. 2011. "A TALE Nuclease Architecture for Efficient Genome Editing." *Nature Biotechnology* 29 (2): 143–48.
Mojica, F. J., G. Juez, and F. Rodríguez-Valera. 1993. "Transcription at Different Salinities of Haloferax Mediterranei Sequences Adjacent to Partially Modified PstI Sites." *Molecular Microbiology* 9 (3): 613–21.
Mojica, Francisco J. M., César Díez-Villaseñor, Jesús García-Martínez, and Elena Soria. 2005. "Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements." *Journal of Molecular Evolution* 60 (2): 174–82.
Montague, Tessa G., José M. Cruz, James A. Gagnon, George M. Church, and Eivind Valen. 2014. "CHOPCHOP: A CRISPR/Cas9 and TALEN Web Tool for Genome Editing." *Nucleic Acids Research* 42 (Web Server issue): W401–7.
Moon, Su Bin, Jeong Mi Lee, Jeong Gu Kang, Nan-Ee Lee, Dae-In Ha, Kim Do Yon, Sun Hee Kim, et al. 2018. "Highly Efficient Genome Editing by CRISPR-Cpf1 Using CRISPR RNA with a Uridinylate-Rich 3′-Overhang." *Nature Communications* 9 (1): 3651.
Moreno-Mateos, Miguel A., Charles E. Vejnar, Jean-Denis Beaudoin, Juan P. Fernandez, Emily K. Mis, Mustafa K. Khokha, and Antonio J. Giraldez. 2015. "CRISPRscan: Designing Highly Efficient sgRNAs for CRISPR-Cas9 Targeting in Vivo." *Nature Methods* 12 (10): 982–88.
Nami, Fatemeharefeh, Mohsen Basiri, Leila Satarian, Cameron Curtiss, Hossein Baharvand, and Catherine Verfaillie. 2018. "Strategies for In Vivo Genome Editing in Nondividing Cells." *Trends in Biotechnology* 36 (8): 770–86.
Nishida, Keiji, Takayuki Arazoe, Nozomu Yachie, Satomi Banno, Mika Kakimoto, Mayura Tabata, Masao Mochizuki, et al. 2016. "Targeted Nucleotide Editing Using Hybrid Prokaryotic and Vertebrate Adaptive Immune Systems." *Science* 353 (6305). https://doi.org/10.1126/science.aaf8729.
Nishimasu, Hiroshi, Le Cong, Winston X. Yan, F. Ann Ran, Bernd Zetsche, Yinqing

Li, Arisa Kurabayashi, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki. 2015. “Crystal Structure of Staphylococcus Aureus Cas9.” *Cell* 162 (5): 1113–26.

Pinello, Luca, Matthew C. Canver, Megan D. Hoban, Stuart H. Orkin, Donald B. Kohn, Daniel E. Bauer, and Guo-Cheng Yuan. 2016. “Analyzing CRISPR Genome-Editing Experiments with CRISPResso.” *Nature Biotechnology* 34 (7): 695–97.

Pourcel, C., G. Salvignol, and G. Vergnaud. 2005. “CRISPR Elements in Yersinia Pestis Acquire New Repeats by Preferential Uptake of Bacteriophage DNA, and Provide Additional Tools for Evolutionary Studies.” *Microbiology* 151 (Pt 3): 653–63.

Prykhozhij, Sergey V., Vinothkumar Rajan, and Jason N. Berman. 2016. “A Guide to Computational Tools and Design Strategies for Genome Editing Experiments in Zebrafish Using CRISPR/Cas9.” *Zebrafish*. https://doi.org/10.1089/zeb.2015.1158.

Pulecio, Julian, Nipun Verma, Eva Mejía-Ramírez, Danwei Huangfu, and Angel Raya. 2017. “CRISPR/Cas9-Based Engineering of the Epigenome.” *Cell Stem Cell* 21 (4): 431–47.

Qi, Lei S., Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, and Wendell A. Lim. 2013. “Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.” *Cell* 152 (5): 1173–83.

Rakocevic, G., V. Semenyuk, W. P. Lee, and J. Spencer. 2019. “Fast and Accurate Genomic Analyses Using Genome Graphs.” https://www.nature.com/articles/s41588-018-0316-4.

Rand, Arthur C., Miten Jain, Jordan M. Eizenga, Audrey Musselman-Brown, Hugh E. Olsen, Mark Akeson, and Benedict Paten. 2017. “Mapping DNA Methylation with High-Throughput Nanopore Sequencing.” *Nature Methods* 14 (4): 411–13.

Ren, Xingjie, Zhihao Yang, Jiang Xu, Jin Sun, Decai Mao, Yanhui Hu, Su-Juan Yang, et al. 2014. “Enhanced Specificity and Efficiency of the CRISPR/Cas9 System with Optimized sgRNA Parameters in Drosophila.” *Cell Reports* 9 (3): 1151–62.

Rios, Xavier, Adrian W. Briggs, Danos Christodoulou, Josh M. Gorham, Jonathan G. Seidman, and George M. Church. 2012. “Stable Gene Targeting in Human Cells Using Single-Strand Oligonucleotides with Modified Bases.” *PloS One* 7 (5): e36697.

Russell, W. L., E. M. Kelly, P. R. Hunsicker, J. W. Bangham, S. C. Maddux, and E. L. Phipps. 1979. “Specific-Locus Test Shows Ethylnitrosourea to Be the Most Potent Mutagen in the Mouse.” *Proceedings of the National Academy of Sciences of the United States of America* 76 (11): 5818–19.

Sanson, Kendall R., Ruth E. Hanna, Mudra Hegde, Katherine F. Donovan, Christine Strand, Meagan E. Sullender, Emma W. Vaimberg, et al. 2018. “Optimized Libraries for CRISPR-Cas9 Genetic Screens with Multiple Modalities.” *Nature Communications* 9 (1): 5416.

Sapranauskas, Rimantas, Giedrius Gasiunas, Christophe Fremaux, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. 2011. “The Streptococcus Thermophilus CRISPR/Cas System Provides Immunity in Escherichia Coli.”

Nucleic Acids Research 39 (21): 9275–82.
Schaefer, Kellie A., Wen-Hsuan Wu, Diana F. Colgan, Stephen H. Tsang, Alexander G. Bassuk, and Vinit B. Mahajan. 2017. "Unexpected Mutations after CRISPR–Cas9 Editing in Vivo." *Nature Methods*. https://doi.org/10.1038/nmeth.4293.
———. 2018. "Correction: Retraction: Unexpected Mutations after CRISPR–Cas9 Editing in Vivo." *Nature Methods*. https://doi.org/10.1038/nmeth0518-394a.
Schell, J., and M. Van Montagu. 1977. "The Ti-Plasmid of Agrobacterium Tumefaciens, A Natural Vector for the Introduction of NIF Genes in Plants?" *Genetic Engineering for Nitrogen Fixation*. https://doi.org/10.1007/978-1-4684-0880-5_12.
Schöttler, S., C. Wenz, T. Lanio, A. Jeltsch, and A. Pingoud. 1998. "Protein Engineering of the Restriction Endonuclease EcoRV--Structure-Guided Design of Enzyme Variants That Recognize the Base Pairs Flanking the Recognition Site." *European Journal of Biochemistry / FEBS* 258 (1): 184–91.
Sentmanat, Monica F., Samuel T. Peters, Colin P. Florian, Jon P. Connelly, and Shondra M. Pruett-Miller. 2018. "A Survey of Validation Strategies for CRISPR-Cas9 Editing." *Scientific Reports* 8 (1): 888.
Shen, Bin, Wensheng Zhang, Jun Zhang, Jiankui Zhou, Jianying Wang, Li Chen, Lu Wang, et al. 2014. "Efficient Genome Modification by CRISPR-Cas9 Nickase with Minimal off-Target Effects." *Nature Methods* 11 (4): 399–402.
Shen, Max W., Mandana Arbab, Jonathan Y. Hsu, Daniel Worstell, Sannie J. Culbertson, Olga Krabbe, Christopher A. Cassa, David R. Liu, David K. Gifford, and Richard I. Sherwood. 2018. "Predictable and Precise Template-Free CRISPR Editing of Pathogenic Variants." *Nature* 563 (7733): 646–51.
Shmakov, Sergey, Omar O. Abudayyeh, Kira S. Makarova, Yuri I. Wolf, Jonathan S. Gootenberg, Ekaterina Semenova, Leonid Minakhin, et al. 2015. "Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems." *Molecular Cell* 60 (3): 385–97.
Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. "Detecting DNA Cytosine Methylation Using Nanopore Sequencing." *Nature Methods* 14 (4): 407–10.
Spanjaard, B., B. Hu, N. Mitic, and P. Olivares-Chauvet. 2018. "Simultaneous Lineage Tracing and Cell-Type Identification Using CRISPR–Cas9-Induced Genetic Scars." *Nature*. https://www.nature.com/articles/nbt.4124.
Stemmer, Manuel, Thomas Thumberger, Maria del Sol Keyer, Joachim Wittbrodt, and Juan L. Mateo. 2017. "Correction: CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool." *PLOS ONE*. https://doi.org/10.1371/journal.pone.0176619.
Swarts, Daan C., Matthijs M. Jore, Edze R. Westra, Yifan Zhu, Jorijn H. Janssen, Ambrosius P. Snijders, Yanli Wang, et al. 2014. "DNA-Guided DNA Interference by a Prokaryotic Argonaute." *Nature* 507 (7491): 258–61.
Tanenbaum, Marvin E., Luke A. Gilbert, Lei S. Qi, Jonathan S. Weissman, and Ronald D. Vale. 2014. "A Protein-Tagging System for Signal Amplification in Gene Expression and Fluorescence Imaging." *Cell* 159 (3): 635–46.
Thyme, Summer B., Laila Akhmetova, Tessa G. Montague, Eivind Valen, and Alexander F. Schier. 2016. "Internal Guide RNA Interactions Interfere with

Cas9-Mediated Cleavage." *Nature Communications* 7 (June): 11750.
Tsai, Shengdar Q., Zongli Zheng, Nhu T. Nguyen, Matthew Liebers, Ved V. Topkar, Vishal Thapar, Nicolas Wyvekens, et al. 2015. "GUIDE-Seq Enables Genome-Wide Profiling of off-Target Cleavage by CRISPR-Cas Nucleases." *Nature Biotechnology* 33 (2): 187–97.
Uusi-Mäkelä, Meri I. E., Harlan R. Barker, Carina A. Bäuerlein, Tomi Häkkinen, Matti Nykter, and Mika Rämet. 2018. "Chromatin Accessibility Is Associated with CRISPR-Cas9 Efficiency in the Zebrafish (Danio Rerio)." *PloS One* 13 (4): e0196238.
Varshney, Gaurav K., Wuhong Pei, Matthew C. LaFave, Jennifer Idol, Lisha Xu, Viviana Gallardo, Blake Carrington, et al. 2015. "High-Throughput Gene Targeting and Phenotyping in Zebrafish Using CRISPR/Cas9." *Genome Research* 25 (7): 1030–42.
Wang, Harris H., Hwangbeom Kim, Le Cong, Jaehwan Jeong, Duhee Bang, and George M. Church. 2012. "Genome-Scale Promoter Engineering by Coselection MAGE." *Nature Methods*. https://doi.org/10.1038/nmeth.1971.
Wang, Tim, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. 2014. "Genetic Screens in Human Cells Using the CRISPR-Cas9 System." *Science* 343 (6166): 80–84.
Wang, Xiaoling, Yebo Wang, Xiwei Wu, Jinhui Wang, Yingjia Wang, Zhaojun Qiu, Tammy Chang, He Huang, Ren-Jang Lin, and Jiing-Kuan Yee. 2015. "Unbiased Detection of off-Target Cleavage by CRISPR-Cas9 and TALENs Using Integrase-Defective Lentiviral Vectors." *Nature Biotechnology* 33 (2): 175–78.
Wang, Xuning, Charles Tilford, Isaac Neuhaus, Gabe Mintier, Qi Guo, John N. Feder, and Stefan Kirov. 2017. "CRISPR-DAV: CRISPR NGS Data Analysis and Visualization Pipeline." *Bioinformatics* 33 (23): 3811–12.
Wienert, Beeke, Stacia K. Wyman, Christopher D. Richardson, Charles D. Yeh, Pinar Akcakaya, Michelle J. Porritt, Michaela Morlock, et al. 2019. "Unbiased Detection of CRISPR off-Targets in Vivo Using DISCOVER-Seq." *Science* 364 (6437): 286–89.
Wilson, Laurence O. W., Aidan R. O'Brien, and Denis C. Bauer. 2018. "The Current State and Future of CRISPR-Cas9 gRNA Design Tools." *Frontiers in Pharmacology* 9 (July): 749.
Winter, Jan, Marc Schwering, Benedikt Rauscher, Florian Heigwer, and Michael Boutros. 2017. "Abstract A10: CRISPR-AnalyzeR (caR): Web-Based, Interactive and Exploratory Analysis and Documentation of Pooled CRISPR/Cas9 Screens." *Molecular Cancer Therapeutics* 16 (10 Supplement): A10–A10.
Xu, Han, Tengfei Xiao, Chen-Hao Chen, Wei Li, Clifford A. Meyer, Qiu Wu, Di Wu, et al. 2015. "Sequence Determinants of Improved CRISPR sgRNA Design." *Genome Research* 25 (8): 1147–57.
Xu, Ke, A. Francis Stewart, and Andrew C. G. Porter. 2015. "Stimulation of Oligonucleotide-Directed Gene Correction by Redβ Expression and MSH2 Depletion in Human HT1080 Cells." *Molecules and Cells* 38 (1): 33–39.
Xu, Shu, Shasha Cao, Bingjie Zou, Yunyun Yue, Chun Gu, Xin Chen, Pei Wang, et al. 2016. "An Alternative Novel Tool for DNA Editing without Target Sequence Limitation: The Structure-Guided Nuclease." *Genome Biology* 17 (1): 186.

You, Qi, Zhaohui Zhong, Qiurong Ren, Fakhrul Hassan, Yong Zhang, and Tao Zhang. 2018. “CRISPRMatch: An Automatic Calculation and Visualization Tool for High-Throughput CRISPR Genome-Editing Data Analysis.” *International Journal of Biological Sciences* 14 (8): 858–62.

Zetsche, Bernd, Jonathan S. Gootenberg, Omar O. Abudayyeh, Ian M. Slaymaker, Kira S. Makarova, Patrick Essletzbichler, Sara E. Volz, et al. 2015. “Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System.” *Cell* 163 (3): 759–71.

Zhang, Feng, Le Cong, Simona Lodato, Sriram Kosuri, George M. Church, and Paola Arlotta. 2011. “Efficient Construction of Sequence-Specific TAL Effectors for Modulating Mammalian Transcription.” *Nature Biotechnology* 29 (2): 149–53.

Zhang, Jian-Ping, Xiao-Lan Li, Guo-Hua Li, Wanqiu Chen, Cameron Arakaki, Gary D. Botimer, David Baylink, et al. 2017. “Efficient Precise Knockin with a Double Cut HDR Donor after CRISPR/Cas9-Mediated Double-Stranded DNA Cleavage.” *Genome Biology* 18 (1): 35.

Zuo, Erwei, Yidi Sun, Wu Wei, Tanglong Yuan, Wenqin Ying, Hao Sun, Liyun Yuan, Lars M. Steinmetz, Yixue Li, and Hui Yang. 2019. “Cytosine Base Editor Generates Substantial off-Target Single-Nucleotide Variants in Mouse Embryos.” *Science* 364 (6437): 289–92.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230850251 (print)
9788230846667 (PDF)