CleanFinder: Browser-Native Analysis of Editing Outcomes and Allelic States

Spiessbach H.[1], Ramachandran H.[2], and Rossi A.[1,2,#]

[1] Environmental Adaptation and Cellular Resilience Laboratory, IUF-Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany
[2] Genome Engineering and Model Development Laboratory, IUF-Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany

[#] Correspondence at: andrea.rossi@IUF-duesseldorf.de

Precise validation of genome editing by targeted sequencing is a critical, multi-step process. Existing tools often separate amplicon definition from data analysis, creating fragmentation and added complexity. We developed CleanFinder, a browser-native application that unifies these steps into a single, streamlined workflow.
Based on user-provided sgRNAs or primer pairs, CleanFinder retrieves the corresponding genomic context and automatically defines the amplicon region, allowing users to copy or visualize the targeted sequence. When sgRNAs are supplied, they also establish sequence anchors that guide the alignment of sequencing reads for accurate quantification of editing outcomes. These anchors are fully adjustable, enabling users to extend or contract upstream and downstream regions as needed. The analytical engine then classifies sequencing reads into key functional categories, while a dedicated allelic dropout module identifies heterozygous single-nucleotide polymorphisms (SNPs) to support direct assessment of allelic dropout events.
To provide essential biological context, CleanFinder includes an interactive gene viewer that maps sgRNA targets and visualizes transcript-specific coding sequences, protein translations, and overall gene structure. Importantly, CleanFinder operates entirely client-side, ensuring complete data privacy since genomic information is never uploaded and no installation is required. By integrating advanced analytical and visualization capabilities into a secure, all-in-one solution, CleanFinder makes robust genome editing analysis accessible to any researcher, regardless of bioinformatics expertise.

Introduction

Genome engineering has profoundly reshaped experimental biology, enabling precise genetic modifications that are now routine across basic research and translational applications (1). Amplicon sequencing is the workhorse for validating genome editing outcomes, yet analytical workflows remain fragmented (2). Command-line suites provide analytical depth and batch scalability but require bioinformatics expertise and controlled compute environments (3–7). Conversely, browser-based tools improve

accessibility and often safeguard privacy by running locally, but many rely on pre-defined or preloaded amplicons and are largely confined to intragenic regions, offering limited support for deep intronic or intergenic targets and minimal linkage to genomic context (8–10). As a result, many laboratories must still choose between analytical power and usability.

CleanFinder bridges this gap as a fully browser-native, client-side application that unifies amplicon definition and sequencing-read analysis within a single, integrated workflow. From user-supplied single guide RNAs (sgRNAs) or primers, it retrieves the genomic locus, defines the amplicon, and establishes robust anchor sequences for alignment. The tool automatically classifies reads into key functional categories, wild-type, Homology directed repair (HDR), single-nucleotide polymorphisms (SNP)/substitution, and in-/out-of-frame indels. CleanFinder also identifies heterozygous single-nucleotide polymorphisms (SNPs) to quantify allelic dropout, providing a complementary quality-control layer for genome-edited clones. A built-in schematic gene viewer visualizes sgRNA positions, transcript-specific coding regions, and amino-acid translations, thereby contextualizing edits across both intragenic and intergenic regions. Coupled with interactive mutation profiling, and requiring no installation or data upload, CleanFinder provides a practical, privacy-preserving bridge from experimental design to validated genome-editing outcomes.

Results

CleanFinder is implemented as an interactive single-page application in JavaScript and operates entirely client-side to ensure data privacy. The software is organized into multiple interconnected modules, the Amplicon Finder, Transcript Finder, Genome PCR, FASTQ Analyzer, Allelic Dropout Analyzer, and Genome Viewer, that together provide a continuous workflow from experimental design to validated outcome analysis.

Amplicon sequences are retrieved dynamically from Ensembl reference genomes (11) rather than relying on preloaded or static user inputs. When a user provides PCR primers or an sgRNA, CleanFinder performs an in silico search against the reference genome to identify the precise locus and define a robust amplicon anchored in genomic context. Primer binding sites, PAM sequences, genomic coordinates, and nearby genes are displayed interactively, with direct links to UCSC (12) and Ensembl genome browsers. By integrating primer design, amplicon definition, and sequencing read analysis within the same environment, CleanFinder eliminates common sources of error such as mismatched reference builds or miscopied sequences and provides a reproducible, privacy-preserving analysis pipeline. The CleanFinder workflow begins by defining a reference amplicon through two distinct modes: paired-primer mode or sgRNA mode (Fig.1). In paired-primer mode, the software locates the forward and reverse primer binding sites within the reference genome and defines the amplicon as the intervening region, optionally extended by user-specified flanking bases. In

single/sgRNA mode, a user-provided guide sequence is searched within a defined genomic context; once a unique hit is identified, CleanFinder defines the amplicon boundaries by adding upstream and downstream padding (default: 100 bp). Strand orientation is inferred automatically, if the guide is detected as its reverse complement, the negative strand is assigned for all subsequent analyses.

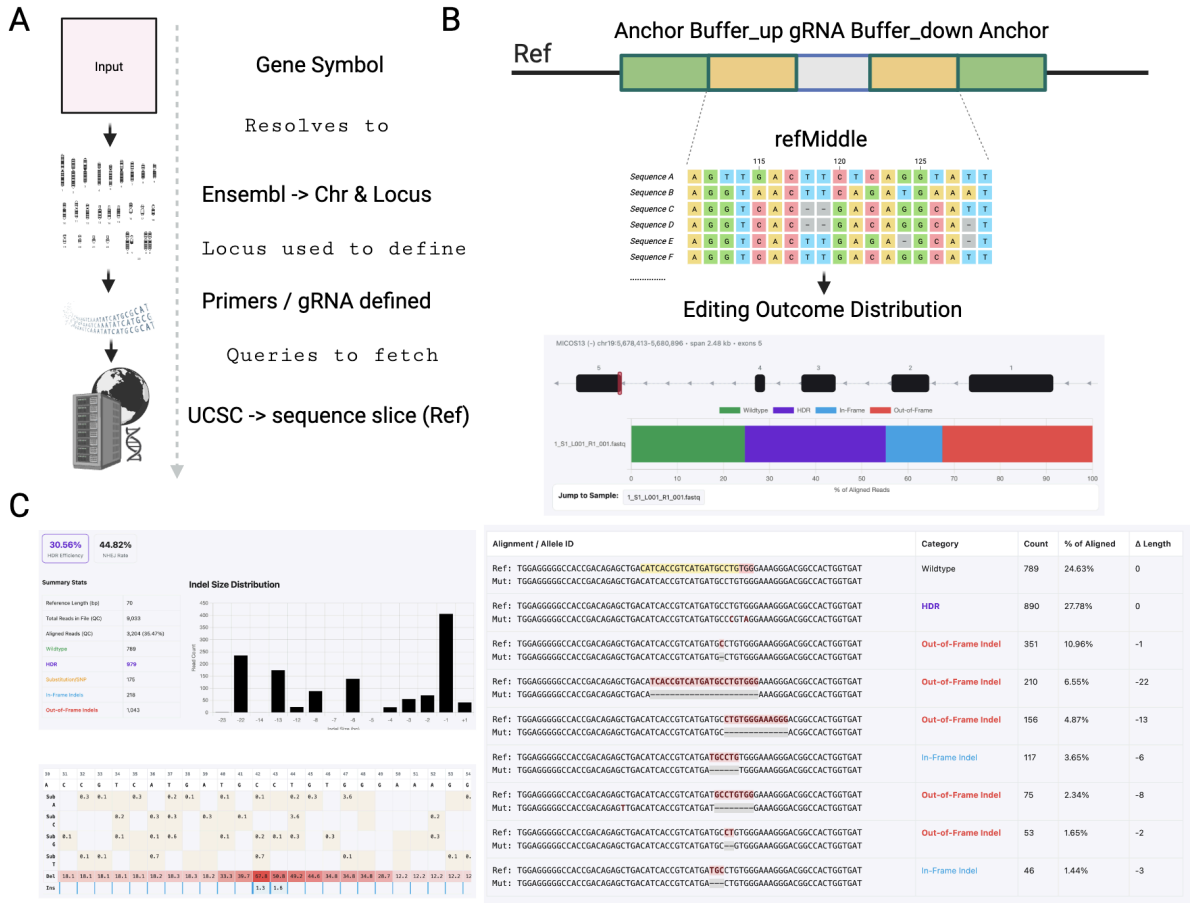


Figure 1. **Overview of the CleanFinder analysis workflow and example output for the MIC13 locus.** (**A**) (A) CleanFinder automatically retrieves the target genomic sequence from public databases based on a user-defined gene symbol. The symbol is first resolved via Ensembl, which provides the corresponding chromosome and locus. This information is then used to define the primer or sgRNA position and to generate sequence queries that fetch the corresponding genomic region from the UCSC Genome Browser, producing the reference amplicon (Ref). (B) Within the reference, CleanFinder defines a structured analytical window composed of two outer anchors and an internal refMiddle segment containing Buffer_up, sgRNA, and Buffer_down regions. Reads containing both 5′ and 3′ anchors are extracted and globally aligned to the reference using a Needleman–Wunsch algorithm (ref) with affine gap penalties. Each aligned read is classified according to its editing outcome, and results are summarized as an Editing Outcome Distribution, showing the relative proportions of Wildtype,

HDR, In-frame, and Out-of-frame alleles. (C) Example analysis of a *MIC13* (*MICOS13*) editing experiment. The summary dashboard displays the proportion of aligned reads, the indel size distribution, and the positional heat map of insertions and deletions. Individual alleles are listed with their aligned sequences, category, indel length (Δ), and frequency among all reads.

To improve efficiency and prevent time-consuming full-genome scans, users specify a chromosome and optionally a gene ID or symbol, which CleanFinder uses to restrict the search space to the relevant genomic region. Defining the chromosome is essential to enable a direct, API-based retrieval of the target locus (Fig. 1A); without it, the software would need to iteratively query the entire genome, resulting in prohibitive processing times. This targeted approach ensures that sequence extraction remains both fast and reproducible, even when operating on large reference assemblies.

Thus, when a gene is provided, the application automatically retrieves its chromosomal coordinates from *Ensembl* and focuses the search on that region with adjustable padding. This design choice enables precise and rapid amplicon definition within the browser while maintaining compatibility with large reference genomes.

The resulting amplicon is visualized together with annotated primer or guide sites, PAM motifs, and overlapping gene structures, providing immediate confirmation of genomic context. Multiple hits are reported and flagged for review to ensure specificity, and all amplicon coordinates, strand assignments, and reference build information are recorded internally for reproducibility. Once the amplicon is defined by either method, the workflow transitions seamlessly to the anchor-based algorithm for FASTQ read classification and mutation profiling.

Read Classification and Alignment Algorithm

CleanFinder employs a robust, buffer-based anchoring algorithm to efficiently identify on-target reads while remaining resilient to sequencing artifacts. For a given reference amplicon, the system defines two anchor sequences based on three user-configurable parameters: anchor length, upstream buffer, and downstream buffer. Unlike conventional methods that use primers as direct anchors, CleanFinder defines anchors in flanking regions outside the primer binding sites (Fig.1B). During analysis, each FASTQ read is scanned for the presence of both a 5′ and a 3′ anchor, allowing up to two mismatches by default. Reads containing both anchors are considered on-target. The sequence between the inner edges of these anchors, the analytical payload or refMiddle, is extracted and aligned globally to the reference amplicon using a modified Needleman–Wunsch algorithm (13) optimized for speed and sensitivity around the expected cleavage site.

The choice of a Needleman–Wunsch alignment framework was deliberate to ensure compatibility across sequencing platforms with different error profiles. Unlike heuristic aligners (e.g., Minimap2, Bowtie2, BLAST) optimized for short reads (14), global dynamic programming maintains base-level precision across the entire amplicon, which

is critical for quantifying insertions and deletions near the CRISPR cut site. The implementation was further optimized to tolerate the higher indel and substitution rates typical of Oxford Nanopore reads, while preserving the accuracy required for Illumina data. In addition, the classification logic adapts to the expected editing footprints of different nucleases: small (1–10 bp) indels for SpCas9 and broader (up to ~30 bp) deletions for Cas12a are both recognized and accurately quantified within the same analytical framework (15). This design enables CleanFinder to analyze both short- and long-read datasets without requiring parameter adjustments, ensuring consistent classification across technologies.

This two-step approach, rapid anchor-based filtering followed by focused dynamic alignment, achieves both accuracy and performance within a browser-native environment. The buffer-based architecture enhances robustness by decoupling read detection from the exact primer sequences, ensuring that on-target reads are recovered even when SNPs or sequencing errors occur within primer binding sites. In sgRNA-based analyses, the buffer coordinates also provide a stable spatial frame of reference: the protospacer start position within the aligned allele directly corresponds to the upstream buffer length.

After alignment, reads are automatically classified into intuitive categories: wild-type (perfect match to the reference), HDR (perfect match to a user-supplied donor template), known allele (match to a user-defined variant), substitution (one or more base mismatches), in-frame indel (indel size divisible by three), and out-of-frame indel (indel size not divisible by three) (Fig.1C).

Interactive dashboards summarize allele frequencies, indel spectra, and positional mutation heatmaps, while a lightweight genome viewer overlays exons and coding sequences to facilitate functional interpretation. CleanFinder efficiently processes tens of thousands of reads interactively without server infrastructure. Its architecture is compatible with both Illumina short reads and Oxford Nanopore long reads, without requiring parameter adjustments between platforms. As an example, Figure 1 shows the analysis of *MICOS13* bulk-editing experiments in human iPSCs. The user entered the gene name, which CleanFinder resolved to the corresponding genomic locus, and supplied the sgRNA sequence. Anchors were defined automatically, establishing the reference frame for read alignment and classification. The analysis revealed a spectrum of editing outcomes, including wild-type reads, HDR events, substitutions, and both in-frame and out-of-frame indels. The positional mutation profile highlighted recurrent deletion hotspots, while the allele breakdown table enabled direct inspection of the aligned sequences. In addition to these editing statistics, CleanFinder provides an interactive mutational heatmap and an integrated gene viewer of the genomic context of the gRNA target site. These visualization tools allow users to directly relate editing

outcomes to their underlying genomic position, facilitating rapid interpretation of on-target editing patterns and locus-specific effects.

Allelic Dropout Detection via Heterozygous SNP Analysis

On-target large deletions are a well-documented and undesirable outcome of genome editing (16, 17). In addition to quantifying editing outcomes, CleanFinder incorporates a specialized module to address a common challenge in CRISPR experiments: allelic dropout, where one allele is preferentially lost due to large deletions or primer-site mutations (Fig.2). To diagnose this, the tool leverages the deep sequencing data to identify naturally occurring heterozygous Single Nucleotide Polymorphisms (SNPs) within the amplicon, which act as endogenous genomic barcodes. When a heterozygous SNP (a biallelic state) is present, it provides strong evidence that both parental alleles were successfully amplified. In contrast, the absence of such a SNP in a known heterozygous sample suggests a monoallelic state, indicative of allelic dropout.

Figure 2. Allelic dropout and heterozygous SNP tracking logic implemented in CleanFinder. (**A**) Conceptual representation of two parental haplotypes carrying a heterozygous SNP (A/C) and a shared 5-bp deletion introduced by CRISPR editing. Both haplotypes are detected within the sequencing reads, confirming the presence of two distinct alleles in the sample. (**B**) Schematic of the global alignment of long-read sequences (ONT) to the reference. Reads are aligned using a Needleman–Wunsch algorithm tolerant to insertions and deletions. Gaps (–) correspond to indels typical of ONT data but do not affect SNP counting. The SNP locus (A/C) is highlighted at the aligned position where

base frequencies are calculated across all reads. (**C**) Example output from the Allelic Dropout module. The table summarizes the highest-confidence heterozygous SNP identified within the amplicon, including its position relative to the anchor, detected bases, allele frequencies, read coverage, and reference sequence context. The near-equal frequencies (A = 0.513, C = 0.487) confirm a biallelic configuration, indicating no allelic dropout.

The analytical logic for this module is tailored for high-confidence SNP discovery from long-read data (Fig.2A). Unlike the editing analysis, which is centered on a known cut site, this workflow begins with a user-defined stable anchor sequence, a short, unique region expected to be conserved across all reads. The process unfolds in three stages:
1. Read Filtering: The entire FASTQ dataset is first scanned to isolate reads containing the stable anchor sequence (or its reverse complement). This pre-filtering step enriches for on-target reads suitable for analysis; 2. Matrix Construction: For each filtered read, a local alignment is performed against the reference amplicon using a Needleman-Wunsch dynamic programming algorithm. From these alignments, CleanFinder constructs a positional base-frequency matrix. This matrix tabulates the counts of A, C, G, and T observed at every position along the reference amplicon across thousands of reads (Fig.2B). 3. SNP Calling: The algorithm then interrogates this matrix to find high-confidence heterozygous SNP candidates. A position is called as a candidate SNP only if it meets a stringent set of user-configurable criteria: sufficient read coverage (default: >500x), the presence of exactly two dominant bases, and allele frequencies falling within a heterozygous range (default: 30-70%). If multiple candidates are found, they are ranked by a quality score that prioritizes high coverage and allele frequencies closest to the ideal 50:50 ratio (Fig. 2C).
The final output is a definitive call of "BIALLELIC" or "MONOALLELIC," accompanied by a table detailing the position, base change, allele frequencies, and coverage of the highest-quality SNP. This provides researchers with a clear, data-driven assessment of whether both alleles are present in their post-editing sequencing pool. To demonstrate the utility of this module, long-read ONT data from the XPA locus were analyzed, where CleanFinder's SNP-calling function detected loss of heterozygosity consistent with monoallelic editing (Fig. 2C) (18). This example highlights the ability of the allelic dropout workflow to reveal subtle genomic outcomes that would otherwise remain undetected in conventional CRISPR read analyses.

Beyond amplicon definition and read classification, CleanFinder includes auxiliary modules that extend its functionality across the experimental workflow. The Transcript Finder allows users to explore transcript isoforms and coding sequence structures surrounding the target site, while the Genome PCR module performs in silico amplification to validate primer specificity and detect potential off-target loci. The integrated Genome Viewer provides an interactive schematic of gene organization, highlighting exons, CDS regions, and gRNA or primer positions within their genomic

context. Together, these features transform CleanFinder from a specialized analysis utility into a comprehensive browser-native environment for designing, validating, and interpreting genome-editing experiments.

Discussion

CleanFinder was benchmarked across a range of genome-editing datasets derived from distinct target genes, cell types, and sequencing platforms. The pipeline performed consistently across both short-read (Illumina) and long-read (Oxford Nanopore) data without requiring parameter adjustments, accurately reproducing known editing outcomes and allelic configurations. This cross-platform compatibility is enabled by the use of a modified Needleman–Wunsch alignment algorithm, which maintains high sensitivity to insertions and deletions near the cut site while accommodating the higher error rates typical of long-read sequencing. Together, these features highlight the robustness and versatility of the underlying architecture and support its application to diverse genome-engineering experiments.

A central innovation of CleanFinder is its buffer-based anchoring algorithm, which ensures accurate read identification even when sequence variations occur near primer or sgRNA sites. Unlike precompiled CRISPR analysis platforms, CleanFinder does not rely on preloaded gRNA libraries or exon annotations, nor does it require the user to manually provide a reference sequence. Instead, it dynamically retrieves and aligns user-supplied gRNA sequences within the specified genomic region. This design allows accurate targeting and analysis of both exonic and deep intronic loci, ensuring that even custom or non-coding targets are correctly recognized and aligned. By anchoring alignment in stable flanking regions rather than within primer sequences, the method remains resilient to SNPs or local sequencing errors that can lead to false read rejection in conventional pipelines (3,6). This strategy captures a more complete and unbiased spectrum of on-target reads, reducing systematic underestimation of editing efficiency and providing a truer representation of repair diversity.

CleanFinder operates entirely client-side (8-10), performing all computations locally within the user's web browser. This design removes dependence on external servers and ensures that sequencing data remain on the user's device, an increasingly common approach to facilitate privacy and reproducibility. By eliminating installation and compatibility barriers, CleanFinder provides a lightweight and accessible environment for data analysis, allowing experimentalists to explore results interactively while maintaining full control over both input and output data.
Beyond post hoc data analysis, CleanFinder also serves as a practical tool for experimental design. Its ability to retrieve and visualize genomic sequences surrounding a target region supports the design of PCR and qPCR primers, including those for allelic

dropout validation and long-range knock-in assays. The integrated Genome Viewer and Transcript Finder modules allow users to examine which transcript isoforms are amplified by a given primer pair and to explore corresponding coding sequences and predicted protein translations. Graphical visualization of CRISPR target sites within annotated gene structures provides immediate insight into whether an edit affects coding, untranslated, or intergenic regions, extending CleanFinder's utility beyond data analysis to experimental planning and interpretation within the same browser environment.

CleanFinder integrates several analytical steps that are often distributed across different tools, such as in silico primer validation, SNP-based allelic-dropout detection, and interactive mutation visualization, within a single, user-friendly interface. This integration streamlines data inspection and improves reproducibility by keeping all major quality-control and visualization functions in one environment. Overall, CleanFinder provides a practical, browser-based solution that facilitates transparent and consistent analysis of genome-editing experiments.
CleanFinder provides a robust analysis tool, and ongoing improvements in sequencing technology will further benefit the types of analyses CleanFinder performs, particularly for challenging low-diversity libraries.
In fact, recent advances in sequencing technology are expected to further improve the quality and accessibility of amplicon-based genome editing analyses. In particular, the new Illumina MiSeq i100 system, equipped with patterned flow cells and XLEAP-SBS chemistry, provides substantially higher signal uniformity and base-calling accuracy for low-diversity libraries such as CRISPR amplicons. These improvements reduce the need for high PhiX spike-in percentages, thereby maximizing sequencing capacity without compromising quality. On the long-read side, continued support for low-cost flow cells such as Oxford Nanopore's Flongle remains important, as these devices enable rapid and cost-efficient sequencing of small targeted panels, an ideal complement for tools like CleanFinder in pilot or validation experiments, although their long-term availability appears uncertain.

Current limitations include performance constraints with very large FASTQ datasets (>200 MB) and the focus on the human reference genome, although expansion to additional species is planned. Future developments will aim to extend support for the analysis of larger genomic rearrangements, such as deletions spanning several kilobases.
In conclusion, CleanFinder helps bridge the gap between wet-lab experimentation and computational analysis. By providing an accessible browser-native platform, it serves as a valuable tool for any lab performing molecular biology and genome-editing validation.

Methods

Implementation

CleanFinder is implemented as a self-contained single-page application (SPA) using standard web technologies (HTML5, CSS3, and JavaScript), without reliance on external front-end frameworks. All data processing, including sequence alignment and analysis, is executed entirely client-side in the user's web browser, ensuring data privacy and eliminating the need for server-side computation or software installation. The application interfaces with public bioinformatics databases through their respective REST APIs; it resolves gene symbols to coordinates using the Ensembl API and retrieves genomic sequences from the UCSC Genome Browser API. Interactive data visualizations are generated using the Chart.js library, while the gene browser renders schematics as Scalable Vector Graphics (SVG). CleanFinder is available at andrearossi.de and can be launched with the prompt *open cleanfinder*.

Genomic Locus Definition and Sequence Retrieval

The initial step in the CleanFinder workflow is the retrieval of a target genomic sequence for analysis. The target locus is defined by the user either by chromosome coordinates or by a gene symbol, which is resolved to its genomic coordinates via the Ensembl REST API. The application then fetches a large genomic segment (user-defined chunk size, default 500kb) centered on this locus from the UCSC Genome Browser API to serve as the search space. Amplicon definition then proceeds via one of two modes. For paired-primer mode, the algorithm scans the retrieved genomic sequence for an exact match of the forward primer. Upon a successful match, it then searches the downstream sequence for the reverse complement of the reverse primer. A valid amplicon is defined only if both primers are found in the correct 5'→3' orientation and the resulting product size does not exceed a user-defined maximum (default 5 kb).
In sgRNA mode, the application searches for the 20 bp protospacer sequence. A valid target site is confirmed by the presence of a canonical Protospacer Adjacent Motif (PAM) immediately adjacent to the protospacer: NGG for S. pyogenes Cas9 or TTTV for Acidaminococcus sp. BV3L6 Cas12a. The search is performed on both DNA strands by also querying for the reverse complement of the protospacer. Once a valid sgRNA target site is located, the reference amplicon is computationally generated by adding user-specified upstream and downstream flanking sequences (default 100 bp).

Upon successful identification by either method, CleanFinder reports the exact amplicon coordinates, overlaps with known gene annotations, including transcript, coding sequence, and protein translation regions, the predicted CRISPR/Cas cut site within the

gene context, and the complete amplicon sequence, which can be also directly visualized on the UCSC Genome Browser for user review.

Allelic Dropout

The Allelic Dropout SNP Analyzer module requires three user-provided inputs to initiate an analysis: a FASTQ file containing long-read sequencing data, the corresponding wild-type reference amplicon sequence, and a stable anchor sequence. The reference amplicon can be directly obtained from the gDNA Finder module. The anchor sequence, typically a conserved region of approximately 20 bp expected to be free of variation, serves as a stable landmark for initial read filtering. Key analysis parameters are user-configurable, including the maximum number of anchor-containing reads to process and the minimum and maximum allele frequency thresholds (e.g., 30-70%) for calling a heterozygous SNP. The computational pipeline begins by scanning the raw FASTQ data to filter for reads that contain the user-defined anchor sequence (or its reverse complement). A subset of these reads, limited by the user's setting, proceeds to the alignment step. The alignments are then processed to construct a positional base-frequency matrix, which tabulates the nucleotide counts (A, C, G, and T) at each position of the reference amplicon. Finally, the algorithm interrogates this matrix to identify high-confidence heterozygous SNP candidates. A position is designated a candidate if it meets the user-defined criteria, which include minimum read coverage and the presence of exactly two dominant alleles with frequencies within the specified heterozygous range. The highest-scoring SNP candidate is then reported, with the results display showing the SNP's position, the reference and variant bases, their respective frequencies, the total coverage at that locus, and a flanking sequence context (10 bp upstream and downstream) for verification.

Alignment algorithm

CleanFinder employs two implementations of a Needleman–Wunsch dynamic programming algorithm for global sequence alignment, each tailored to a specific analytical task. The core algorithm populates a scoring matrix to determine the optimal alignment path between each read and the reference sequence.

For the primary Read Classification in the FASTQ Analyzer, a more sophisticated implementation with affine gap penalties (see below). This model penalizes the creation of a new gap more strongly than the extension of an existing one, which enables accurate reconstruction of the precise length and position of CRISPR-induced indels when aligning the refMiddle payload.

While the same scoring scheme is applied to both SpCas9 and Cas12a datasets, the subsequent classification logic accounts for the distinct indel characteristics of each enzyme: small (1–10 bp) insertions or deletions are most frequent for Cas9, whereas

Cas12a edits typically produce broader deletions extending up to 30 bp or more (15). Accordingly, CleanFinder applies enzyme-specific size thresholds during read classification to correctly interpret large deletions without over-penalizing them.
Dynamic Scoring Matrix: The scoring is dynamically adjusted based on the selected enzyme for more biologically accurate results: SpCas9: { match: +5, mismatch: -4, gap open: -10, gap extend: -1 }; Cas12a: { match: +4, mismatch: -3, gap open: -8, gap extend: -1 }; Default: { match: +2, mismatch: -1, gap open: -5, gap extend: -1 }
For the Allelic Dropout SNP Analyzer, a standard Needleman–Wunsch algorithm with a linear gap penalty is sufficient (scoring: match = +2, mismatch = −1, gap = −2). Since the goal of this module is not to characterize indels but to construct a positional base-frequency matrix, this implementation is optimized for robustness against the high indel error rates characteristic of long-read (ONT) sequencing. During matrix construction, only nucleotide matches and mismatches contribute to the base counts at each position; gaps in the read alignment, representing insertions or deletions, are ignored. This substitution-focused approach enables reliable detection of heterozygous SNPs even within noisy long-read data.

Classification and visualization

Reads are categorized as wild-type, HDR, known allele, substitution, in-frame indel, or out-of-frame indel. Allele counts and frequencies are summarized, and results are displayed as interactive plots (allele frequency bar charts, indel histograms, substitution heat maps). Results can be exported for reporting.

Genome Engineering

Guide RNA sequences were designed using CHOPCHOP (https://chopchop.cbu.uib.no/) (19). The corresponding sgRNAs were cloned into pSpCas9(BB)-2A-GFP (PX458, Addgene #48138). Human iPSCs, KOLF2.1J and iPS12, were maintained under feeder-free conditions in mTeSR™ Plus medium (STEMCELL Technologies) as previously described (20). For transfection, 3 x 10^5 cells were seeded using Accutase Cell Detachment Solution (Pan Biotech) on day 0 in 6-wells pre-coated with geltrex matrix. On day 1, cells were transfected using Lipofectamine™ Stem Reagent (Invitrogen) following the manufacturer's instructions. For knocking out, 2 µg of plasmid were used with 8 µL of Lipofectamine™ Stem Reagent. For knocking in, 2 µg of plasmid and 1 µg of HDR single-stranded Oligodonor (ssODN) were used with 12 µL of Lipofectamine™ Stem Reagent. On day 3 EGFP-positive cells were sorted on a BD FACS Aria III and processed as described previously (20). All sgRNA sequences and primers used are available at https://data.mendeley.com/datasets/54m5svs2rc/1

Deep Sequencing

DNA was extracted using Proteinase K as previously described (9), and amplification was carried out as follows: first-round PCRs were performed using 1 µl of cell lysate as template in a 25 µl reaction with 1 µl of Blitz DNA Polymerase 2 U/µl (KleverLab), mixing 10 µl 2.5xBuffer with 1 µl dNTP mix (10 mM) and 1 µl primer mix (10 µM). Amplifying conditions were optimized: Initial Denaturation 98 °C 2 min; 25 cycles: Denaturation 98 °C 30 sec, Annealing 60 °C 20 sec, Elongation 72 °C 1 min; Final Elongation 72 °C 10 min. Amplifying conditions for *KCNQ2*-primers were further optimized with an Annealing temperature of 65 °C. Custom primers were designed with adaptor overhangs compatible with the second-round PCR. Subsequently, 2 µl of the first-round product was used as template for the second-round PCR, which employed identical cycling conditions (Annealing temperature 60°C) and clone-specific barcoded primers. PCR products were size-separated on a 1% agarose gel and purified with the FastPure Gel DNA Extraction Mini Kit (NeoBiotech). After using a Qubit fluorometer (Thermo Fisher Scientific) for quantifying, amplicons were pooled into an equal representation library of 30 ng/µl, which was further quantified in a size-considering quantification step with. Agilent TAPE station 4150 with TAPE station D5000 High-Sensitivity Kit was used to control the amplicon size. The final library was quantified using NEBNext Library Quant Kit for Illumina (NEB) and loaded at a concentration of 40pM, targeting approximately 20% PhiX. Sequencing was performed on an Illumina MiSeq i100 Plus system using a 5M flow cell in single-end mode (251 cycles) (21) at a depth of around 150-250K reads per sample.

To identify heterozygous SNPs near the target loci, PCR primers were designed to amplify ~3.5 kb of the surrounding regions. Amplification was performed using Q5 High-Fidelity DNA Polymerase (NEB), and products were assessed on a 1% agarose gel. Bands of the expected size were excised and purified using the FastPure Gel DNA Extraction Mini Kit (Vazyme). Library preparation for Oxford Nanopore sequencing followed the Native Barcoding Kit 24V14 protocol (SQK-NBD114.24). Sequencing was carried out on a MinION with a Flongle flow cell and adapter. Raw reads were basecalled using Dorado (v1.0.2) with the super-high-accuracy model, and analyzed with Allelic Dropout SNP Analyzer.

All sequencing FASTQ data are available at: https://data.mendeley.com/datasets/54m5svs2rc/1

Acknowledgments

We thank Jochen Dobner for feedback on the software, Vanessa Baltruschat and Marie Brauers for technical assistance.
The authors acknowledge the use of Google's Gemini Pro model for assistance in JavaScript code generation, algorithm refinement, and proofreading.

The project was directly supported by the Deutsche Forschungsgemeinschaft (DFG) (RO5380/10-1).
The Environmental Adaptation and Cellular Resilience lab and the GEMD are also funded by Deutsche Forschungsgemeinschaft (DFG) (RO5380/1-1 and RO5380/13-1), VHL (von Hippel-Lindau) Betroffene Familien e.V. through a grant awarded to A.R, the AFM-Téléthon (25179 to A.R.), the European Union (EU) and North Rhine-Westphalia (NRW) start-up grant (IN-SU-3-001 to A.R.), and the Leibniz Competition (SAW) Cooperative Excellence project (K642/2024 to A.R.).
The IUF is funded by the federal and state governments, the Ministry of Culture and Science of North Rhine-Westphalia (MKW), and the Federal Ministry of Education and Research (BMBF). The figures were generated using Canvas X Draw and Biorender.

Ethics

The applicant laboratory has the necessary permission to work with recombinant DNA and genetically modified organisms under S1 and S2 conditions.

Contributions

Conceptualization (A.R.), methodology and source code (A.R.), investigation (H.S. and H.R.), visualization (H.S. and H.R.), writing - original draft (A.R.), manuscript (A.R. with the input of all authors), review and editing (A.R., H.S. and H.R.): supervision (A.R.), project administration, and funding acquisition (A.R.). All authors reviewed and approved the manuscript.

The authors declare no competing interests.

Availability

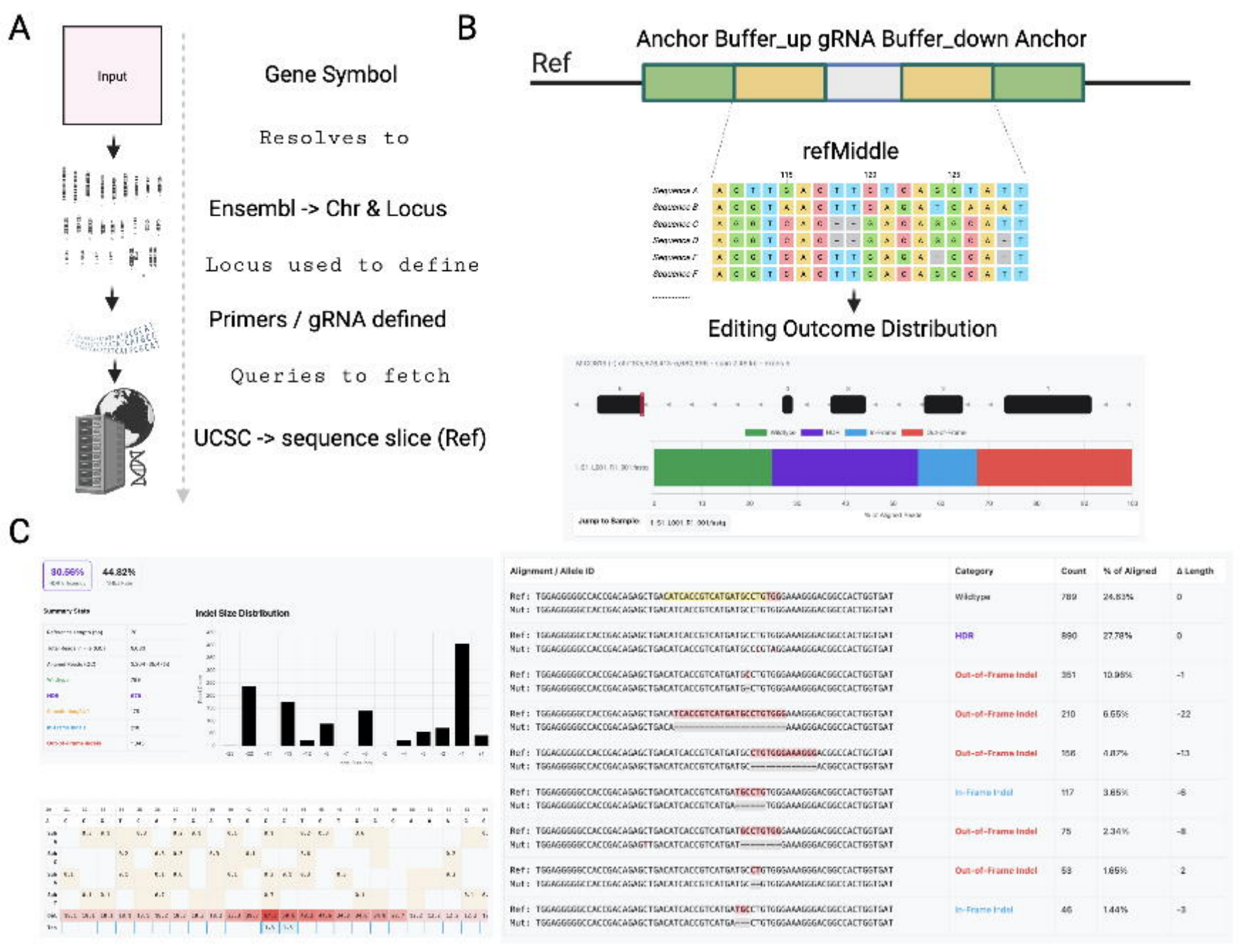
CleanFinder is provided as a browser-native application, accessible directly at www.andrearossi.de (type: open cleanfinder), and runs in any modern browser without installation. The underlying logic, including anchor definition and alignment strategy, is described in the Methods section. Code available at https://github.com/andrearossi-lab/CleanFinder/tree/main Datasets available at: https://data.mendeley.com/datasets/54m5svs2rc/1

References

1. Gaj T, et al. ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. 2013.

2. Li C, et al. Computational tools and resources for CRISPR/Cas genome editing. Genomics Proteomics Bioinformatics. 2023.
3. Clement K, et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nat Biotechnol. 2019.
4. Wang X, et al. CRISPR-DAV: CRISPR NGS data analysis and visualization pipeline. Bioinformatics. 2017.
5. Boel A, et al. BATCH-GE: Batch analysis of NGS data for genome editing assessment. Sci Rep. 2016.
6. Labun K, et al. ampliCan: Accurate analysis of genuine CRISPR editing events. Genome Res. 2019.
7. Fu H, et al. CRISPR-GRANT: a cross-platform graphical analysis tool for CRISPR genome editing evaluation. BMC Bioinformatics. 2023.
8. Park J, et al. Cas-Analyzer: an online tool for assessing genome editing results using NGS data. Bioinformatics. 2017.
9. Nguyen T, et al. CRISPRnano: A web-based tool for the integrated analysis of genome editing outcomes. Nucleic Acids Res. 2022;50(W1):W199–W207.
10. Schmid-Burgk JL, et al. OutKnocker: A web tool for rapid and simple genotyping of designer nuclease edited cell lines. Genome Res. 2014;24(10):1719–1723.
11. Dyer SC, *et al.* Ensembl 2025. *Nucleic Acids Res.* 2025;53(D1):D948–D957. doi:10.1093/nar/gkae1071
12. Perez G, et al. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res. 2025;53(D1):D1243–D1249. doi:10.1093/nar/gkae974.
13. Needleman SB, et al. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 Mar 28; 48(3): 443–453.
14. Darby CA, et al. Vargas: heuristic-free alignment for assessing linear and graph read aligners. Bioinformatics. 2020 Jun; 36(12): 3712–3718.
15. Port F, et al. Multiplexed conditional genome editing with Cas12a in Drosophila. Proc Natl Acad Sci U S A. 2020 Aug 25;117(37):22890–22899.
16. Kosicki M, et al. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol. 2018 Jul; 36: 765–771.
17. Simkin D, et al. Homozygous might be hemizygous: CRISPR/Cas9 editing in iPSCs results in detrimental on-target defects that escape standard quality controls. Stem Cell Rep. 2022 Apr 12; 17(4): 993–1008.
18. Papadopoulou M, et al. CRISPR/Cas9-mediated editing of XPA in induced pluripotent stem cells: a model for investigating Xeroderma Pigmentosum and NER dysfunction. Stem Cell Res. 2025; in review. doi: 10.2139/ssrn.5531342.
19. Labun K, et al. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. 2019 Jul 2; 47(W1): W171–W174.

20. Dobner J, et al. Reassessment of marker genes in human induced pluripotent stem cells for enhanced quality control. Nat Commun. 2024 Oct 2; 15: 8547.
21. Binder S, et al. CRISPR/Cas9-mediated editing of ACTB in induced pluripotent stem cells: a model for investigating human ACTB loss-of-function and genetic adaptive responses. Stem Cell Res. 2024 Jun; 77:103395.


A
Input
Gene Symbol
Resolves to
Ensembl -> Chr & Locus
Locus used to define
Primers / gRNA defined
Queries to fetch
UCSC -> sequence slice (Ref)
B
Ref
Anchor Buffer_up gRNA Buffer_down Anchor
refMiddle
Editing Outcome Distribution
C
30.56%
44.82%
Summary Stats
Indel Size Distribution
Alignment / Allele ID
Category
Count
% of Aligned
Δ Length
Wildtype
HDR
Out-of-Frame Indel
In-Frame Indel

A

```
                    SNP(A/C)    Shared 5bp
                                deletion
Haplotype 1 ----------- A ----------- Δ5 -----------
Haplotype 2 ----------- C ----------- Δ5 -----------
                        ↓
              Two Alleles Detected
```

B

```
Ref   : ...C T G A C A G T A G C...Mut of Interest...
Read1: ...C T G - C A G T A G C...Mut of Interest...
Read2: ...C T G A C C G T A G C...Mut of Interest...
Read3: ...C T G A - - G T A G C...Mut of Interest...
Read4: ...C T G A C A G T A G C...Mut of Interest...
Read5: ...C T G - C C G T A G C...Mut of Interest...
Read6: ...C T G A C C G T A G C...Mut of Interest...
Read7: ...C T G A - - G T A G C...Mut of Interest...
Read8: ...C T G A C A G T A G C...Mut of Interest...
Read9: ...C T G - C A G T A G C...Mut of Interest...
ReadX: ...C T G A C C G T A G C...Mut of Interest...
ReadY: ...C T G A - - G T A G C...Mut of Interest...
                    ↓
              SNP Locus A/C
```

C

Highest Quality SNP

Position (rel. to anchor)	Bases	Frequencies	Coverage	Context (Reference)
658	A / C	0.513 / 0.487	2,916	ACCTCCCAAC[A]CATCCCTCAC