JACKIE: Fast enumeration of single- and multi-copy CRISPR binding sites

Jacqueline Jufen Zhu[1], Albert Wu Cheng[1,2,3,4,5]

1. The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA
2. The Jackson Laboratory Cancer Center, Bar Harbor, ME 04609, USA
3. Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06030, USA
4. Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT 06030, USA
5. Correspondence: albert.cheng@jax.org

ABSTRACT

Summary: CRISPR-based methods for genome, epigenome editing and imaging have provided powerful tools to interrogate the functions of the genome. The design of guide RNA (gRNA) is a vital step of CRISPR experiments. We report here the implementation of JACKIE (Jackie and Albert's CRISPR K-mer Instances Enumerator), a pipeline for enumerating all potential single- and multi-copy CRISPR sites in the genome. We demonstrate the application of JACKIE to identify locus-specific repetitive sequences for CRISPR/Casilio-based genomic labeling.

Availability: Source codes and CRISPR site databases (JACKIEdb) for hg38 and mm10 are available for download at http://crispr.software/JACKIE

1. INTRODUCTION

CRISPR/Cas technologies have revolutionized genome research by providing versatile tools for "writing" the genome, epigenome and transcriptome as well as "reading" the dynamics of the genome through live cell genomic imaging techniques (Sander and Joung, 2014). Genomic

imaging by CRISPR/Cas-based technologies require clustered repetitive sequences (Cheng, et al., 2016; Ma, et al., 2018; Maass, et al., 2018; Qin, et al., 2017). dCas9-based gene activation and epigenome editing also benefited from synergistic binding events (Cheng, et al., 2016; Cheng, et al., 2013; Maeder, et al., 2013; Mali, et al., 2013; Taghbalout, et al., 2019). CRISPR-mediated gene deletion may be induced using a pair of gRNAs flanking the region of interest (Canver, et al., 2017) while other CRISPR applications may benefit from a database of unique CRISPR binding sites.

For fast enumeration of all potential CRISPR binding sites and their locations of occurrences in the genome, we have implemented JACKIE (Jackie and Albert's CRISPR K-mer Instances Enumerator) pipeline that is compatible with high performance computing (HPC) clusters. We further provide helper scripts for extracting binding sites within regions of interest, as well as a wrapper script to run Cas-OFFinder (Bae, et al., 2014) to identify off-target profiles of selected CRISPR binding sites.

2. IMPLEMENTATION

JACKIE first scans the genome and enumerates any potential SpCas9 binding sites with the PAM constraint (nGG). To allow memory and computation efficiency, sequences are encoded into an unsigned 64-bit integer (NucKey) by encoding each nucleotide in 3 bits (Fig 1). The chromosome and location for each instance are also encoded as integer values. The NucKey, chromosome and location (KeyedPosition) are written into files prefixed by a predefined prefix length (default to 6). The second step: Each bin file is loaded into a list in memory, and sorted by NucKey, thus placing records of CRISPR sites with identical sequences adjacent to each other in the list. The sequence-sorted list is then traversed to output CRISPR sites in bed interval files. The extensive divide-and-conquer design allows JACKIE to be highly parallelized on an HPC cluster (Fig 1).

Helper scripts were written to take the bed file outputs from JACKIE to further collapse same-chromosome CRISPR binding sites into an extended bed interval file format with each record containing the CRISPR binding sites encoded as blocks. Such extended bed files can then be loaded onto genome browsers for visualization. Other helper scripts were written for downstream processing, such as for the fast extraction of CRISPR sites overlapping defined set of intervals and automated queries into an offline version of Cas-OFFinder (Bae, et al., 2014) to compute off-target profiles for selected sgRNA sequences. sgRNA sequences, number of occurrences, and span sizes (the distance between the 5' most site to 3' most site for CRISPR sites on the same chromosome) are encoded in the output files so that CRISPR sites can be filtered by simple text processing software (e.g., awk) to identify sites for specific purposes. We used JACKIE to generate CRISPR binding site databases (JACKIEdb) for human genome version hg38 and mouse genome mm10, which took ~35min and ~22min, respectively, on an HPC cluster with 2,568 Intel Xeon compute cores and >38TB of system memory.

3. USE CASES

Some potential applications of JACKIE pipeline are illustrated in Fig 2. For example, highly repetitive sequences (>30 copies) and low repetitive arrays (~6 copies) can be used for imaging applications (Fig 2a) (Cheng, et al., 2016; Ma, et al., 2018; Maass, et al., 2018; Qin, et al., 2017). These sites can be identified from JACKIE output by filtering for the desired range of repeat number and span size. To demonstrate such use case, we have selected three CRISPR binding sequences with >30 copies clustered specifically within 10kb region. We then applied Casilio imaging to visualize the sequence clusters via fluorescence microscopy (Fig 2b), producing 3 fluorescence foci in each case, consistent with the ploidy of the HEK293T cells used in the experiments. Similarly, promoter-specific repeats might be identified (e.g., >3 copies, within 1kb) overlapping gene promoters for synergistic gene activation or repression by CRISPRa or CRISPRi technologies (Fig 2c) (Cheng, et al., 2016; Cheng, et al., 2013; Maeder, et al., 2013;

77 Mali, et al., 2013; Taghbalout, et al., 2019). Two or more copy sites flanking or spanning a
78 region or interest might be identified for inducing deletion or complex chromosome
79 rearrangements (Fig 2d). (Canver, et al., 2017; Choi and Meyerson, 2014; Maddalo, et al., 2014;
80 Maresch, et al., 2016).

81 Although currently designed for SpCas9 binding site enumeration, JACKIE can be
82 modified to identify binding sites for other CRISPR/Cas systems by changing the code segment
83 written for the PAM requirement filtering. Alternatively, PAM requirement might be removed from
84 the codes to enable identification of all k-mer sites and their occurrences in the genome. The
85 use of one uint64 for encoding NucKey limits the enumeration of sequences up to 20mer. By
86 using multiple NucKey, or a variable-size vector of NucKey, longer k-mer sequences can be
87 enumerated.

88

89 **Experimental Material & Methods**

90 sgRNA spacer sequences (as listed in Fig 2b) were cloned into sgRNA-15xPBSc expression
91 vector via an oligo-annealing protocol as previously described (Cheng, et al., 2016). HEK293T
92 cells were cultivated in Dulbecco's modified Eagle's medium (DMEM) (Sigma) with 10% fetal
93 bovine serum (FBS)(Lonza), 4% Glutamax (Gibco), 1% Sodium Pyruvate (Gibco) and penicillin-
94 streptomycin (Gibco). Incubator conditions were 37 °C and 5% CO2. Cells were seeded into 24-
95 well plates the day before being transfected with 50ng pAC1445-dCas9 plasmid (Addgene
96 #73169), 25ng pAC1447-Clover-PUFa plasmid (Addgene #73689) and 250ng sgRNA-15xPBSc
97 plasmid. Cells were fixed, mounted on slides, and imaged 48 hours after transfection on a Leica
98 SP8 confocal microscope.

99

Acknowledgement

This work has been supported by the Jackson Laboratory internal grants (to A.C.) and National Human Genome Research Institute 1R01HG009900 (to A.C).

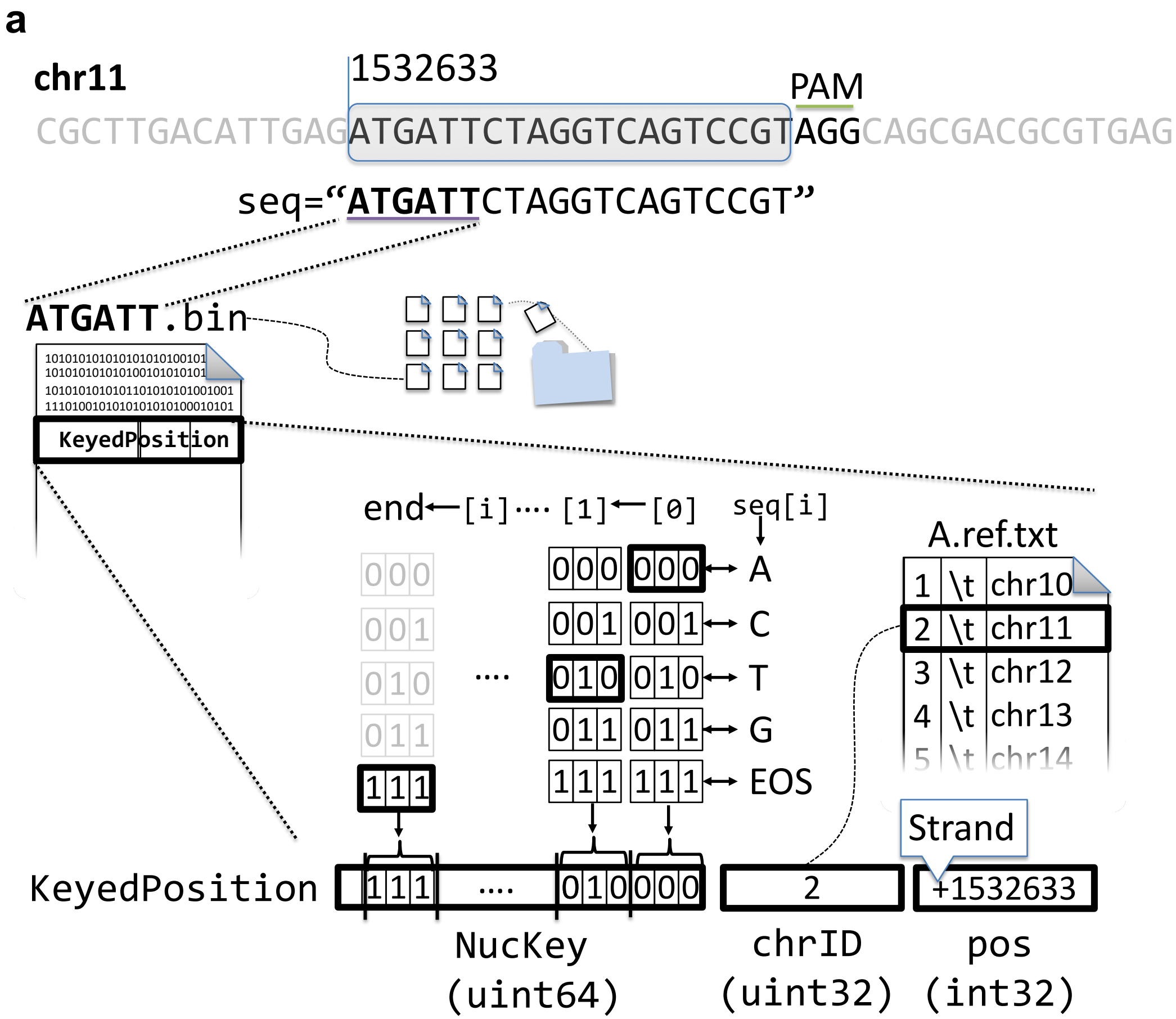
Figure Legends

Fig 1. Implementation of JACKIE. (a) JACKIE first scans the genome for potential SpCas9 binding sites, and encode each occurrence in a data structure called KeyedPosition, consisting of NucKey (uint64), chrID (uint32), and position (int32). NucKey records a binary representation of the 20-mer spacer sequence using 3 bits per nucleotide. chrID records an integer representing a chromosome. A chrID-chromosome mapping file is generated automatically. Position records the location of the binding site, with negative and positive integers representing the minus strand and positive strand, respectively. JACKIE outputs KeyedPosition as it scans the genome. For efficient sorting in the next step, KeyedPosition records are appended to a file named by the 6-mer prefix (<6merPrefix>.bin). To parallelize this step, four processes are started as separate cluster jobs, each focusing on the A, C, T, or G as the first nucleotide of binding sites. **(b)** In the second step, JACKIE loads each <6merPrefix>.bin file and performs a sort on the KeyedPosition records on the NucKey variable using std_sort routine and then traverse the sorted list of KeyedPosition to output a bed interval file containing the coordinate position of each CRISPR sites with item name encoded by <NucKey>.<copyNumber>/<sequence>, and score field recording the copy number. Parallelization is achieved by either starting separate job on individual <6merPrefix>.bin files, or by starting batch jobs focusing on 2mer prefices (e.g., AA AT AC AG operating on AA*.bin, AT*.bin, AC*.bin, AG*.bin, etc). The outputs (<6merPrefix>.bed) are then merged into a combined bed file using cat unix command. Helper scripts are provided for downstream processing of the bed file into collapsed bed file with each record encoding all binding sites of

125 the same sequence in the same chromosome, for extracting binding sites within defined
126 regions, or for running the Cas-OFFinder program to identify off-target profiles of selected
127 sequences.

128 **Fig 2. Potential applications of JACKIE pipeline. (a)** JACKIE can be used to identify
129 clustered CRISPR binding sites with high copy number or low copy number for CRISPR-based
130 genomic imaging experiments. **(b)** To demonstate genomic imaging use case, we filtered
131 sgRNA sequences with >30 copies clustered within specific <10kb regions and performed
132 Casilio imaging experiment using designed sgRNAs. UCSC genome browser screenshots are
133 shown on the left with the line-bar annotating the CRISPR sites, and fluorescence microscopy
134 images shown on the right. Microscopy images were derived by merging green (Casilio-
135 labeling) and blue (DAPI: nucleus stain). Arrows point to fluorescent loci of interest. Scale bars
136 indicate 5μm. **(c)** JACKIE can be used to identify clustered CRISPR binding sites for synergistic
137 activation or repression of target cis-regulatory elements (e.g., promoters) using CRISPR-based
138 epigenetic editing approaches (e.g., CRISPRa or CRISPRi). **(d)** JACKIE can identify unique
139 CRISPR binding sites in the genome for precise indel induction, dual CRISPR sites for targeted
140 deletion, or clustered sites spanning a genomic region to induce complex rearrangement
141 events.

Figure 1


a
chr11
1532633
PAM
CGCTTGACATTGAGATGATTCTAGGTCAGTCCGTAGGCAGCGACGCGTGAG
seq="ATGATTCTAGGTCAGTCCGT"
ATGATT.bin
KeyedPosition
end←[i]···· [1]←[0]
seq[i]
000
001
010
011
111
000 000 A
001 001 C
010 010 T
011 011 G
111 111 EOS
A.ref.txt
1 \t chr10
2 \t chr11
3 \t chr12
4 \t chr13
5 \t chr14
Strand
KeyedPosition
111 ···· 010000
2
+1532633
NucKey
(uint64)
chrID
(uint32)
pos
(int32)

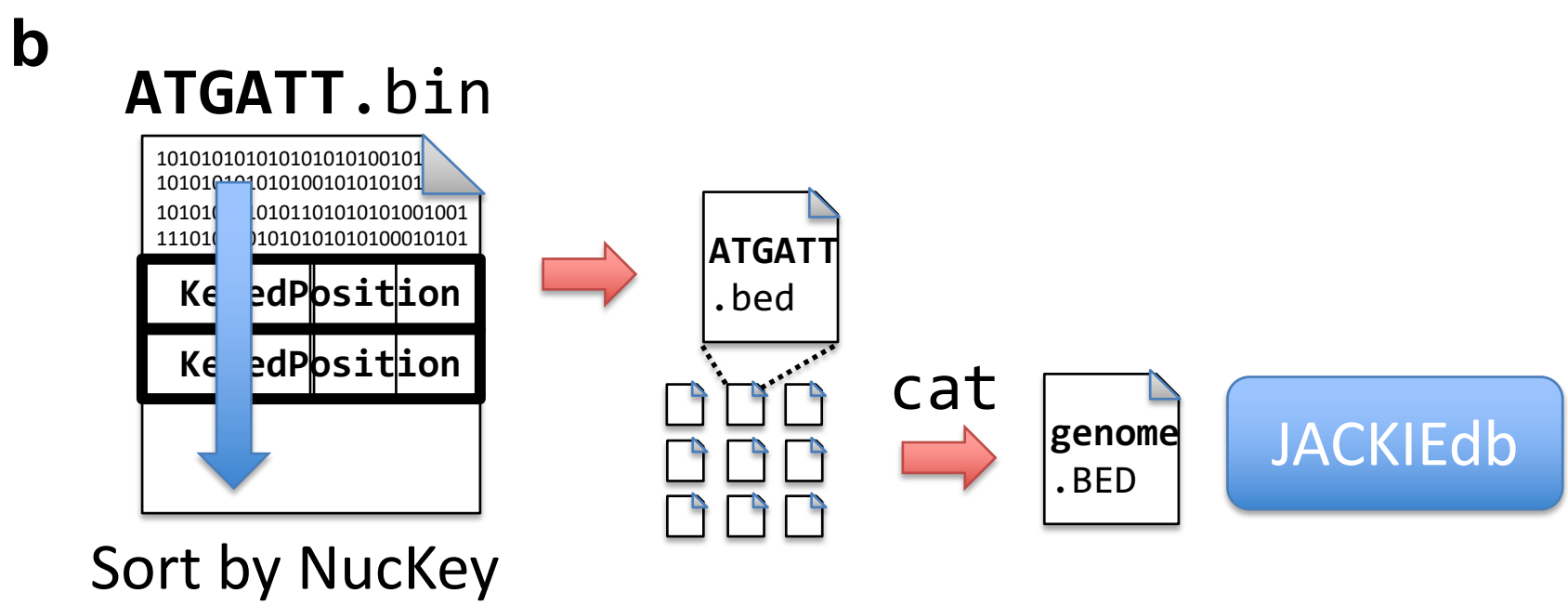
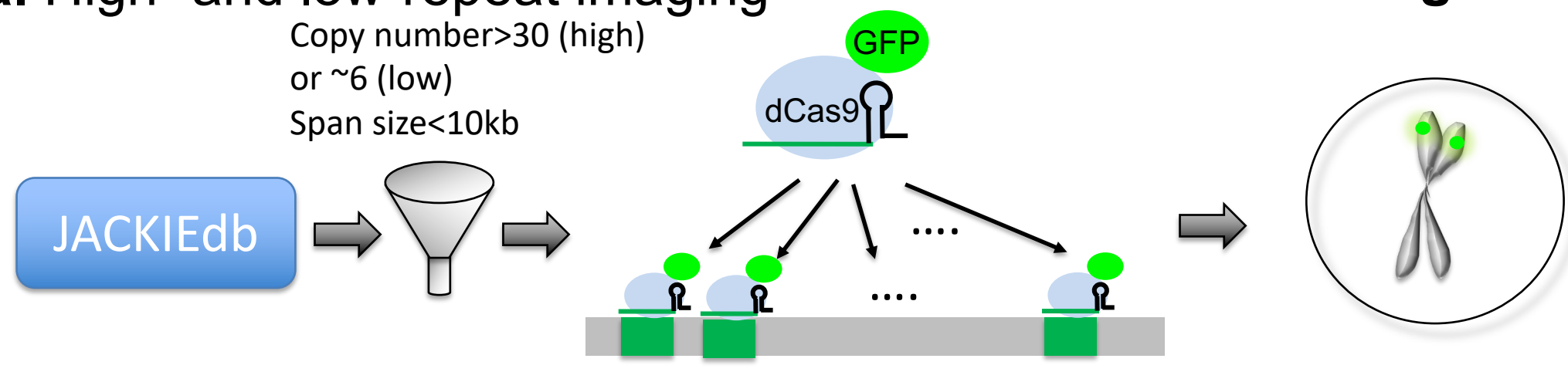

b
ATGATT.bin
KeyedPosition
KeyedPosition
Sort by NucKey
ATGATT
.bed
cat
genome
.BED
JACKIEdb

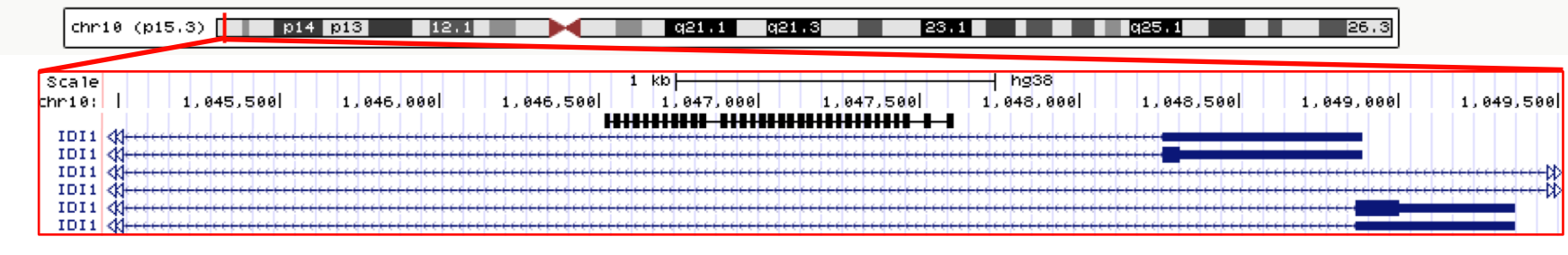
Figure 2

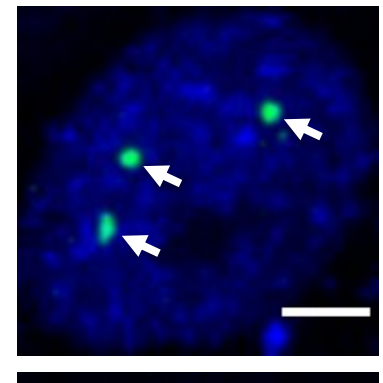
a. High- and low-repeat imaging



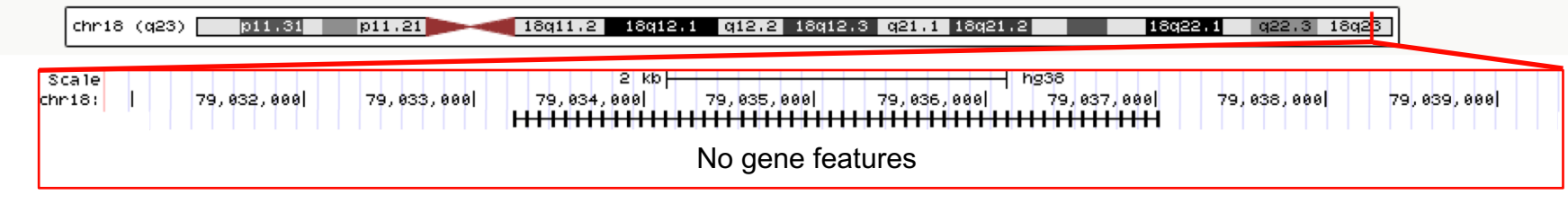
b. Examples sgRNA design for imaging (hg38, >30 copies, <10kb)

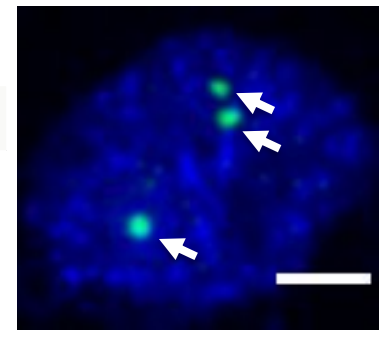
i. GGTGTAGACAGTGGAGCAGC, chr10:1046518-1047610, 33 copies



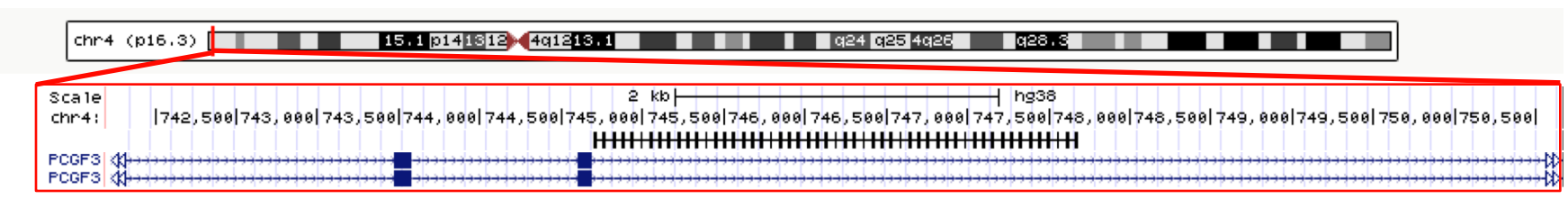


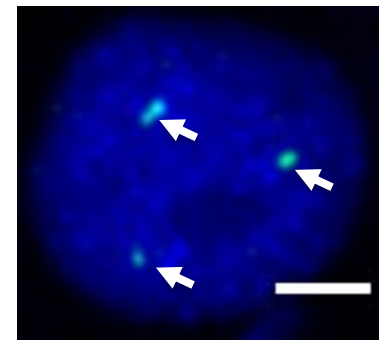
ii. TGAGGACTGAATCTAGAACC, chr18:79033234-79037027, 52 copies



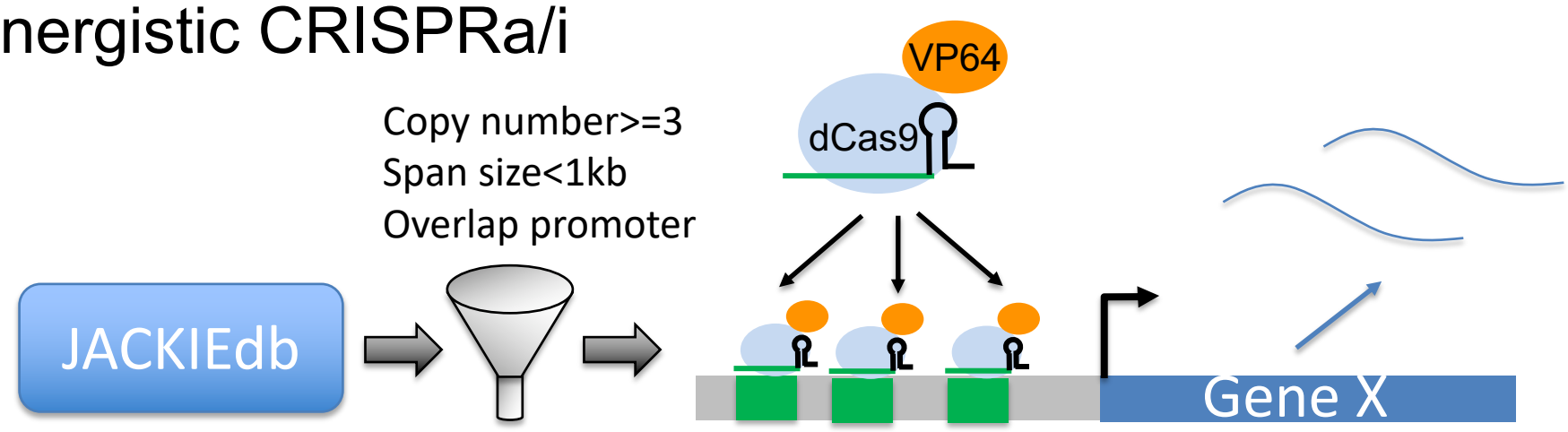


iii. ACTAACACTGCAGCGACCCC, chr4 744703-747683, 53 copies

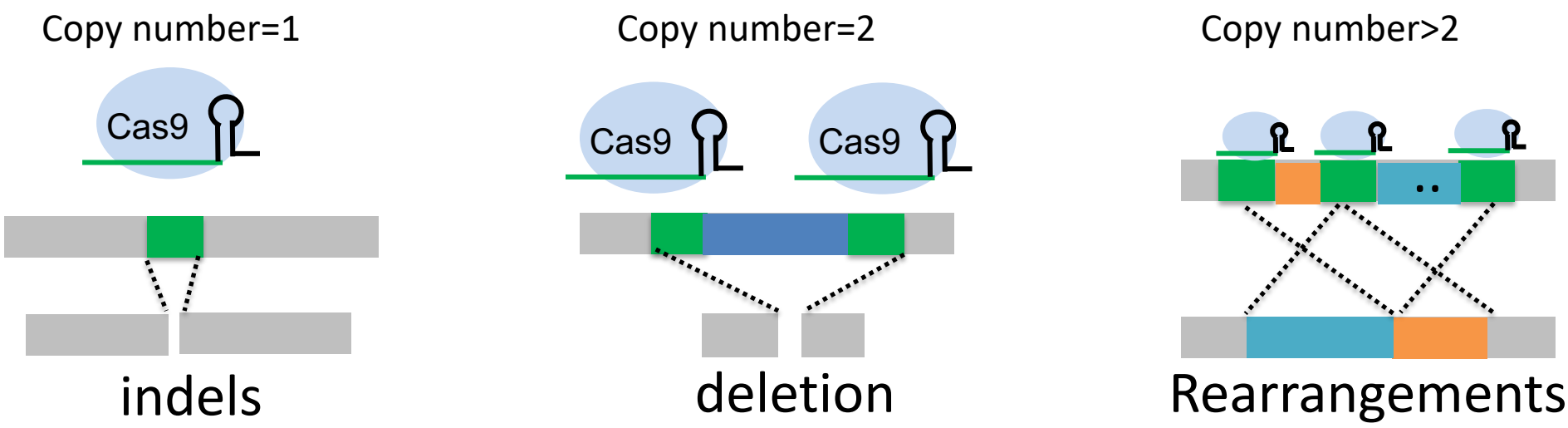




c. Synergistic CRISPRa/i



d. Genome editing applications



142 **References**

143 Bae, S., Park, J. and Kim, J.S. Cas-OFFinder: a fast and versatile algorithm that searches for
144 potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics (Oxford, England)*
145 2014;30(10):1473-1475.

146 Canver, M.C.*, et al.* Characterization of genomic deletion efficiency mediated by clustered
147 regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian
148 cells. *The Journal of biological chemistry* 2017;292(6):2556.

149 Cheng, A.W.*, et al.* Casilio: a versatile CRISPR-Cas9-Pumilio hybrid for gene regulation and
150 genomic labeling. *Cell research* 2016;26(2):254-257.

151 Cheng, A.W.*, et al.* Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided
152 transcriptional activator system. *Cell research* 2013;23(10):1163-1171.

153 Choi, P.S. and Meyerson, M. Targeted genomic rearrangements using CRISPR/Cas
154 technology. *Nature communications* 2014;5:3728.

155 Ma, H.*, et al.* CRISPR-Sirius: RNA scaffolds for signal amplification in genome imaging. *Nature*
156 *methods* 2018;15(11):928-931.

157 Maass, P.G.*, et al.* Spatiotemporal allele organization by allele-specific CRISPR live-cell
158 imaging (SNP-CLING). *Nature structural & molecular biology* 2018;25(2):176-184.

159 Maddalo, D.*, et al.* In vivo engineering of oncogenic chromosomal rearrangements with the
160 CRISPR/Cas9 system. *Nature* 2014;516(7531):423-427.

161 Maeder, M.L.*, et al.* CRISPR RNA-guided activation of endogenous human genes. *Nature*
162 *methods* 2013;10(10):977-979.

163 Mali, P.*, et al.* CAS9 transcriptional activators for target specificity screening and paired
164 nickases for cooperative genome engineering. *Nature biotechnology* 2013;31(9):833-838.

165 Maresch, R.*, et al.* Multiplexed pancreatic genome engineering and cancer induction by
166 transfection-based CRISPR/Cas9 delivery in mice. *Nature communications* 2016;7:10770.

167 Qin, P.*, et al.* Live cell imaging of low- and non-repetitive chromosome loci using CRISPR-Cas9.
168 *Nature communications* 2017;8:14725.
169 Sander, J.D. and Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting
170 genomes. *Nature biotechnology* 2014;32(4):347-355.
171 Taghbalout, A.*, et al.* Enhanced CRISPR-based DNA demethylation by Casilio-ME-mediated
172 RNA-guided coupling of methylcytosine oxidation and DNA repair pathways. *Nature*
173 *communications* 2019;10(1):4296.
174