Synchronization and Error Correction Using Optical and Gene Tags

Linking a database of barcode sequences with cytometric data is prone to errors. In IBC-generating chips, spots for oligonucleotide synthesis could be empty or occupied by multiple cores, or cores may stick elsewhere. In cell/IBC pairing droplets, the one-to-one ratio could be compromised. In a cytometer, multiple cells could be detected as a single event, or cells may stick or switch elution order. These errors would manifest themselves in mismatching indexes in the barcode database and cytometric data, and they would have to be corrected. Two possible correction approaches could be employed.

For presequencing correction, fluorescently marked reference cores could be added into IBCs identified by serial number (but not IBCs with intrinsic IDs) and to analyzed cells. Images of the IBC-producing chip and the cell accumulator (Figure 1A,D) would reveal the elution order of these reference IBCs among regular cells and IBCs (Figure 1A,D; red and violet dots, respectively). These data will be compared with data from on-flow detectors (Figure 1A,B,E,F).

For postsequencing correction, known barcode sequences of paired reference IBCs (one from the cell sample and another from the IBC supply) will be compared with pairing data (Figure 1F) and order of elution expected from optical analysis (Figure 1A,B,E).

Applied in parallel, these methods would reveal if IBCs and cells switched positions in elution order or were lost, and they would reveal junk particle reads and pairing mismatches. Compromised data will be repaired for confirmed cases of switched positions. Alternatively, if the correction data is insufficient for repair, the compromised data for cells between reference IBCs (Figure 1A,B; blue dots between red dots) will be eliminated.

Concluding Remarks

We anticipate that the recent single cell sequencing technologies based on in-drop barcoding are ready to be augmented with the full power of on-flow detection methods, such as cytometry, enabling a transformative analytical tool for comprehensive assessment of heterogeneity in genetic makeup and phenotypic properties of cells, organelles, exosomes, and other biological entities. Here we have suggested pathways to developing such an analytical tool using the concept of IBCs.

Acknowledgments

The project was supported by the Center for Translational Pediatric Research (CPTR) NIH

Center of Biomedical Research Excellence award P20GM121293 (to Alan Tackett) and in part by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000.

[1]X-BIO Institute, University of Tyumen, Tyumen, Russia, 625003

[2]Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, AR 72205-7199, USA

*Correspondence:
dan14444@yahoo.com (D.S.A.),
BLZybaylov@uams.edu (B.L.Z.)

https://doi.org/10.1016/j.tibtech.2019.09.002

References

1. Leites, E.P. and Morais, V.A. (2018) Mitochondrial quality control pathways: PINK1 acts as a gatekeeper. *Biochem. Biophys. Res. Commun.* 500, 45–50
2. Lan, F. *et al.* (2017) Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* 35, 640–646
3. Stoeckius, M. *et al.* (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods.* 14, 865–868
4. Peterson, V.M. *et al.* (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* 35, 936–939
5. Andreyev, D. and Arriaga, E.A. (2007) Simultaneous laser-induced fluorescence and scattering detection of individual particles separated by capillary electrophoresis. *Anal. Chem.* 79, 5474–5478
6. McGrath, K.E. and Archer, D. (2015) Characterizing cell morphology using imaging flow cytometry. *Science.* 349, 999
7. Andreyev, D.S. and Arriaga, E.A. (2007) Fabrication of perforated sub-micron silica shells. *Scr. Mater.* 57, 957–959
8. Kimmerling, R.J. *et al.* (2016) A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nat. Commun.* 7, 10220
9. Rosenthal, A. and Voldman, J. (2005) Dielectrophoretic traps for single-particle patterning. *Biophys. J.* 88, 2193–2205
10. Nuwaysir, E.F. *et al.* (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* 12, 1749–1755
11. Sukhanova, A. and Nabiev, I. (2008) Fluorescent nanocrystal-encoded microbeads for multiplexed cancer imaging and diagnosis. *Crit. Rev. Oncol. Hematol.* 68, 39–59
12. Wood, D.K. *et al.* (2007) A feasible approach to all-electronic digital labeling and readout for cell identification. *Lab Chip.* 7, 469–474
13. Morgan, E. *et al.* (2004) Cytometric bead array: a multiplexed assay platform with applications in various areas of biology. *Clin. Immunol.* 110, 252–266
14. Battersby, B.J. *et al.* Nanomics Biosystems Pty Ltd. Device and methods for directed synthesis of chemical libraries, 1448789A2

Forum

Predicting CRISPR/Cas9-Induced Mutations for Precise Genome Editing

Kutubuddin A. Molla[1,2,3,@] and Yinong Yang[1,2,*]

SpCas9 creates blunt end cuts in the genome and generates random and unpredictable mutations through error-prone repair systems. However, a

growing body of recent evidence points instead to Cas9-induced staggered end generation, nonrandomness of mutations, and the predictability of editing outcomes using machine learning models.

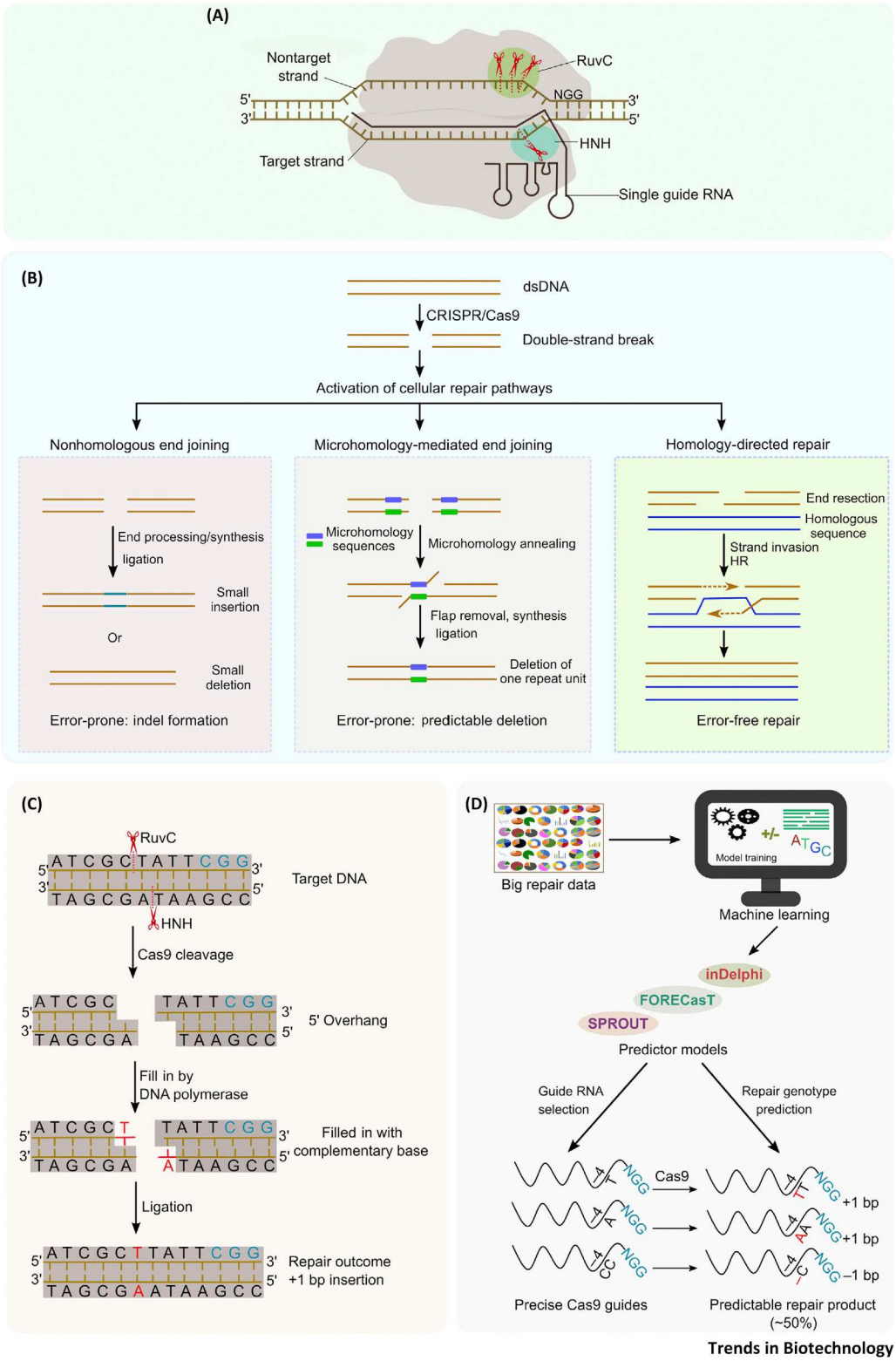
Cas9 Cleavage Creates Both Blunt and Staggered Ends

Cas9 nuclease was thought to make a blunt-ended cut, specifically 3 base pairs (bp) upstream (i.e., the –3 position) of the NGG protospacer adjacent motif (PAM) [1]. Active Cas9 contains two nuclease domains, HNH and RuvC, which are responsible for cleaving the target and nontarget DNA strand, respectively [1]. However, other studies suggested that Cas9 can also produce 1 nucleotide (nt) 5′ staggered ends [1,2]. Towards resolving the controversy, Shou and colleagues found that along with blunt ends, Cas9 also generates nonblunt ends with 1–3 nt 5′ overhangs [3]. Their study showed that HNH accurately cuts at the –3 position upstream of the PAM sequence, whereas RuvC flexibly cuts at either –3, –4, –5, or even further upstream (Figure 1A) [3]. In contrast to the earlier observation that the nontarget strand can be chewed up by the 3′–5′ exonuclease activity of RuvC after the initial cut at –3 base, they found the cleavage is indeed endonucleolytic at –3 or further upstream [3]. A single nucleotide insertion identical to the nucleotide at –4 position was observed as a common repair outcome in another study, further indicating asymmetric DNA cleavage by Cas9 [4]. Similar observations of templated origin of 1 bp insertions in yeast, mouse embryonic stem cells (mESCs), and mammalian cells strongly implies the generation of Cas9-induced 1 bp 5′ staggered DNA ends and subsequent filling in by a DNA polymerase [5–7]. Interestingly, *in vitro* studies revealed that D10A Cas9 nickase (RuvC mutated) cleaves exactly at the –3 position of the target strand, and H840A Cas9 nickase (HNH mutated) makes a flexible cut at –3, –4, and –5 of the nontarget strand [3]. Accurate cleavage at –3 of the target strand by HNH is likely due to the restriction imposed by target DNA–sgRNA (single guide RNA) hybrid formation, whereas the availability of displaced flexible single-strand nontarget DNA may result in the plasticity of RuvC cleavage [8]. Therefore, Cas9 most likely produces both blunt and staggered ends. Upon repair, a blunt end may give rise to random deletion, template-independent insertion, or wild type [7]. By contrast, the generation of an overhang and nonhomologous end joining (NHEJ)-mediated repair can result in predictable templated insertion, making it more desirable for precise genome engineering.

Cas9-Induced Double-Strand Break Repair Outcome is Nonrandom

Convincing results from recent studies suggest that the time has come to rethink the notion of unpredictable and random repair from the Cas9-induced double-strand breaks (DSBs) [3,4,9–11]. They also inspire a path towards template-free precise editing in the genome independently of homology-directed repair (HDR) (Figure 1B). The first indication of nonrandom repair of Cas9-created DSBs came from the study of Li and colleagues [2]. Analysis of 223 DSB repair outcomes in the human genome further showed that the mutation is not arbitrary [12]. One of the major factors determining the outcome is the nature of protospacer sequences, rather than the larger genomic context, cell lines, or delivery of editing reagents [12]. For instance, a single protospacer targeting seven distinct genomic sites yielded similar repair events [12]. Increasing evidence from large-scale recent studies in human cell lines further suggest that the cut site adjacent sequence is an important determining factor for Cas9-induced mutations [3,4,9]. The most consistently predictable class of editing outcome is the single nucleotide insertion [4,9–11]; the inserted nucleotide was found to be identical to the nucleotide at –4 from the PAM sequence [4,9]. If a protospacer contains T at –4, insertion of another T is the most predictable outcome among all mutations. Recently, independent studies in budding yeast, human, and mouse cell lines confirmed this observation [5–7,11]. Consistent findings across the cell lines and even different organisms indicate that indeed the nucleotide at the –4 position is the most influential in determining repair outcome, at least for the predictable insertion. However, the predictability gradually decreases in the order T>A>C>G at the –4 nucleotide position [4,9,11].

Interestingly, the presence of two or more repeating nucleotides at the cut site frequently results in the deletion of the repeating unit [4,9,11]. Iteration of C and G represents the most predictable class of repeating unit for 1 bp deletions [4,9,11]. Similarly, loss of a trinucleotide repeat was found to be the most abundant class of deletion in Cas9 treated mESCs [7]. These findings could be attributed to microhomology-mediated end joining (MMEJ) repair (Figure 1B). Another intriguing recent report showed that the creation of a Cas9 DSB near the center of a tandem microduplication can result in the deletion of one of the repeat sequences [13]. Cells treated with an MMEJ inhibitor drug exhibited a sharp decrease in microhomology-based deletion [13]. In mammalian cell lines with impaired MMEJ repair pathways, paired guide RNA (gRNA) yielded precise DNA fragment deletion, and the inserted



(Figure legend at the bottom of the next page.)

nucleotides at the junction of fragment deletion matched perfectly with the –4 to –10 (upstream of PAM) sequence of the nontarget DNA strand [3]. A preliminary study showed that in NHEJ repair (Figure 1B), incompatible ends generated by DSB undergo processing events like gap filling to make the ends compatible [14]. Once the ends become compatible, they are immediately ligated by ligase 4 to minimize error by further end processing [14]. These findings support the hypothesis that Cas9-induced DSB repair involves 5′ overhang generation, filling in by a polymerase, and subsequent ligation (Figure 1C) [2,5,7]. Specific PAM configuration of the paired guide could result in predictable nucleotide insertion at the junction of deleted, duplicated, or inverted DNA fragments [3]. Hence, the mechanism behind the Cas9-induced mutations, especially insertion of a single base and deletion of a repeating nucleotide unit, is becoming increasingly clear. The studies also indicate that at least a portion of NHEJ and MMEJ repair outcomes are defined and predictable.

Prediction of Repair Outcome Using Machine Learning Models

Given the nonrandom nature of Cas9-induced DSB repair, abundant data available in the public domain can be utilized to develop machine learning algorithms to effectively perform predictions of suitable guides with a high probability of predictable repair outcome and what kind of repair they would generate (Figure 1D and Box 1). Shen and colleagues generated a machine learning model, inDelphi, based on their repair product data from DSBs at 1872 target sites of the human genome [10]. inDelphi categorized precise gRNAs that resulted in a single predictable repair outcome in ≥50% of total editing products. The accuracy of inDelphi prediction was demonstrated by achieving precise deletion and 1 bp insertions in two separate experiments [10]. Using gRNAs identified with inDelphi, this study demonstrated template-free correction of *HPS1* (Hermansky–Pudlak syndrome) and *ATP7A* (Menkes disease) gene mutation with 88% and 94% efficiency, respectively, in patient-derived fibroblasts. Similarly, Allen and coworkers created a computational tool, FORECasT, and demonstrated predictions of in-frame mutations with high accuracy [9] (Box 1). Employing a machine learning approach, a study demonstrated that –2, –3, –4, and –5 nucleotides from PAM are critical for determining editing precision of a target site [4]. Another machine learning model, SPROUT, has been developed to foresee the editing outcome in primary T cells [11]. Like inDelphi, SPROUT correctly predicted and ranked the top sgRNAs based on their likelihood to cause insertion for 73% of the tested genes [11]. Interestingly, SPROUT showed superior performance in repair prediction when compared with inDelphi and FORECasT [11]. Only the targeted genome sequence is required to predict repair outcome using the freely available online tools (Box 1). Those user-friendly web tools would certainly facilitate researchers to fine-tune their experimental design and envisage a part of their CRISPR/Cas9-mediated editing outcome.

The Way Forward

Although base editing, a technique that uses a fusion of catalytically impaired nuclease with a nucleotide deaminase to install targeted point mutation, can cause single nucleotide alteration, it cannot generate precise indels [15]. The most recent findings represent a great leap towards template-free precise genome editing, which should facilitate the development of CRISPR-based therapeutics. The evidence suggests that the repair outcome is not always random, and it depends on the type of cut (blunt vs staggered), the cut site's neighboring bases, and the type of repair pathways. A large proportion of Cas9-induced mutations like large deletions, inversion, and translocation are not mechanistically understood. Developing a full mechanistic picture of what percentage of Cas9-mediated DSBs are staggered and how the DSBs are repaired would expedite prediction and in turn precision genome editing. Availability of suitable and specific inhibitors for different DNA repair pathways could facilitate fine-tuning the balance between different pathways for a desirable outcome. Besides, engineering the nuclease domains for increased/decreased plasticity could facilitate diverse applications. Training data comprising of repair genotypes from other nuclease variants could be used

Figure 1. Double-Strand DNA Breaks, Cellular Repair Pathways, and Prediction of CRISPR/Cas9-Induced Mutations.
(A) Staggered cuts by SpCas9. HNH domain cleaves target strand at –3 position, and RuvC domain can make a cut at either –3, –4, –5, or even further. (B) Genomic double-strand break (DSB) generation is followed by different cellular repair pathways. Error-prone nonhomologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ) pathways create the majority of mutations throughout the cell cycle. Homology-directed repair (HDR), active in S/G2 phases of the cell cycle, repairs DSBs without error. (C) Hypothetical model explaining the generation of a 1 base pair (bp) insertion during CRISPR/Cas9-induced DSB repair. (D) Machine learning aids in the prediction of precise guides and their repair outcome. '–4' signifies position of the nucleotide proximate to the 5′ end of protospacer adjacent motif (PAM). Abbreviations: dsDNA, double-strand DNA; FORECasT, favored outcomes of repair events at Cas9 targets; HR, homologous recombination; inDelphi, inDel score (phi); Nucleotide N, A/T/G/C; SPROUT, CRISPR Repair Outcome.

Box 1. How Does Machine Learning Assist Template-Free Precise Editing?

It is a common notion that mutations resulting from Cas9-induced double-strand breaks (DSBs) are randomly generated by error-prone repair systems like canonical nonhomologous end-joining (NHEJ) and microhomology-mediated end joining (MMEJ). Hence, CRISPR/Cas9 editing outcomes have been thought to be highly unpredictable. But recent studies demonstrated that the Cas9-induced mutation is nonrandom and could be predicted based on local sequence properties [3,4,9,12]. Employing Cas9–DSB-generated large mutational datasets, computational predictor models, including inDelphi, FORECasT, and SPROUT, have been developed to predict the editability of gRNAs and their editing outcome (Table I). SPROUT and FORECasT models use gradient boosted tree ensemble and multinomial logistic regression methods, respectively [9,11]. The inDelphi deletion model was developed based on neural network and multitask framework learning, whereas for insertion modeling, a k-nearest neighbor model was used [10]. Although trained on specific cell lines, those models were validated on various cell lines (Table I). In the genome engineering field, the prevailing view is that the precise and error-free modification requires the utilization of homology-directed repair (HDR) pathway, which is highly inefficient in higher organisms and needs the supply of a donor template containing the desired edit. Using inDelphi, however, 183 microduplication and 508 frameshift pathogenic alleles were corrected to the wild type allele in ≥50% of all editing outcomes in mESCs without employing HDR [10].

Table I. Machine Learning Models for Predicting Cas9-Induced DSB Repair Outcome[a]

Model	Repair data used for training	Cell types		Link to access	Accuracy	Refs
		Trained on	Applicable to			
inDelphi	1872 genomic sites	mESCs and human U2OS	mESCs, U2OS, HEK293, HCT116, and K562	https://indelphi.giffordlab.mit.edu	R = 0.88–0.94 (genotype frequency) R = 0.91 (indel length frequency) R = 0.81 (indel frequency)	[10]
FORECasT	5000 synthetic targets	Human K562	K562, CHO, mESC, hiPSC, HAP1, and RPE1	https://partslab.sanger.ac.uk/FORECasT	R = 0.81 (in-frame mutation)	[9]
SPROUT	1656 genomic sites	Human CD4+ T cells	T cells, hiPSC, HEK293, K562, and HCT116	https://github.com/amirmohan/SPROUT	R^2 = 0.59 (indel with insertion) R^2 = 0.56 (fraction of indel)	[11]

[a]Abbreviations: CHO, Chinese hamster ovary; K562, human chronic myelogenous leukemia cell; HAP1, leukemic near-haploid cell; HCT116, human colorectal carcinoma cell; HEK293, human embryonic kidney cells; hiPSC, human-induced pluripotent stem cells; RPE1, human retina epithelial immortalized cells; U2OS, human osteosarcoma cells.

to facilitate prediction of different mutation patterns. Since precise indel generation can introduce single nucleotide polymorphisms and modify alleles to improve agronomic traits, it will also be important to examine DSB repair outcomes in plant systems. This knowledge would facilitate CRISPR/Cas-enabled precision breeding and crop improvement without linkage drag. To optimize the precision of genome editing independent of the low-efficiency HDR, this emerging area of research itself is worthy of attention and rapid investigation.

Acknowledgments

Kutubuddin A. Molla would like to acknowledge the United States-India Educational Foundation (USIEF), New Delhi and US Department of State for the Fulbright-Nehru Postdoctoral Fellowship. This work was supported by National Science Foundation Plant Genome Research Program grant no. 1740874 and the USDA National Institute of Food and Agriculture and Hatch Appropriations under project #PEN04659 and accession #1016432 to Yinong Yang.

[1]Department of Plant Pathology and Environmental Microbiology, The Pennsylvania State University, University Park, PA 16802, USA

[2]Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

[3]ICAR-National Rice Research Institute, Cuttack 753006, India

@Twitter: @Kutub_joy

*Correspondence: yuy3@psu.edu

https://doi.org/10.1016/j.tibtech.2019.08.002

References

1. Jiang, F. and Doudna, J.A. (2017) CRISPR–Cas9 structures and mechanisms. *Annu. Rev. Biophys.* 46, 505–529
2. Li, Y. *et al.* (2015) A versatile reporter system for CRISPR-mediated chromosomal rearrangements. *Genome Biol.* 16, 111
3. Shou, J. *et al.* (2018) Precise and predictable CRISPR chromosomal rearrangements reveal principles of Cas9-mediated nucleotide insertion. *Mol. Cell* 71, 498–509
4. Chakrabarti, A.M. *et al.* (2018) Target-specific precision of CRISPR-mediated genome editing. *Mol. Cell* 73, 699–713
5. Lemos, B.R. *et al.* (2018) CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proc. Natl. Acad. Sci. U. S. A.* 115, E2040–E2047
6. Taheri-Ghahfarokhi, A. *et al.* (2018) Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Res.* 46, 8417–8434

7. Gisler, S. *et al.* (2019) Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. *Nat. Commun.* 10, 1598
8. Raper, A.T. *et al.* (2018) Functional insights revealed by the kinetic mechanism of CRISPR/Cas9. *J. Am. Chem. Soc.* 140, 2971–2984
9. Allen, F. *et al.* (2018) Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* 37, 64–72
10. Shen, M.W. *et al.* (2018) Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* 563, 646–651
11. Leenay, R.T. *et al.* (2019) Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nat. Biotechnol.* 37, 1034–1037
12. van Overbeek, M. *et al.* (2016) DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell* 63, 633–646
13. Iyer, S. *et al.* (2019) Precise therapeutic gene correction by a simple nuclease-induced double-stranded break. *Nature* 568, 561–565
14. Stinson, B.M. *et al.* (2019) A mechanism to minimize errors during non-homologous end joining. *bioRxiv*. Published online February 28, 2019. http:dx.doi.org/10.1101/563197
15. Molla, K.A. and Yang, Y. (2019) CRISPR/Cas-mediated base editing: technical considerations and practical applications. *Trends Biotechnol.* 37, 1121–1142