Published in final edited form as:

Curr Genet. 2018 December ; 64(6): 1229–1238. doi:10.1007/s00294-018-0850-8.

Scarless genome editing: progress towards understanding genotype-phenotype relationships

Gregory L. Elison^{1,2} and Murat Acar^{1,2,3,4,*}

¹Department of Molecular Cellular and Developmental Biology, Yale University, 219 Prospect Street, New Haven, CT 06511

²Systems Biology Institute, Yale University, 850 West Campus Drive, West Haven, CT 06516

³Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, 300 George Street, Suite 501, New Haven, CT 06511

⁴Department of Physics, Yale University, 217 Prospect Street, New Haven, CT 06511

Abstract

The ability to predict phenotype from genotype has been an elusive goal for the biological sciences for several decades. Progress decoding genotype-phenotype relationships has been hampered by the challenge of introducing precise genetic changes to specific genomic locations. Here we provide a comparative review of the major techniques that have been historically used to make genetic changes in cells as well as the development of the CRISPR technology which enabled the ability to make marker-free disruptions in endogenous genomic locations. We also discuss how the achievement of truly scarless genome editing has required further adjustments of the original CRISPR method. We conclude by examining recently developed genome editing methods which are not reliant on the induction of a DNA double strand break and discuss the future of both genome engineering and the study of genotype-phenotype relationships.

INTRODUCTION

The increasing promise of biological technology has fostered considerable enthusiasm that the eradication of genetic disease and ability to improve crops and other organisms through synthetic biology will become realities in the near future^{1,2}. While these are all distinctly different goals, all share a common requirement in order to come to fruition: the ability to accurately predict the phenotypic consequences of genetic change² (Fig.1). Without this, most forms of human-engineered biology will be reliant on laborious and inefficient design processes to move forward³. This is consistent across all regions of the genome, including protein coding content, regulatory content, structural RNA content, and other areas, including those which have not yet had their functions elucidated⁴. Unfortunately, to say that being able to make such prediction is difficult is an immense understatement; the function of the vast majority of base pairs in even the most well-studied organisms is a mystery³. Even within coding regions, the degree to which individual base pair content matters to protein

*To whom correspondence should be addressed: murat.acar@yale.edu.

function is largely unknown, especially for those coding for amino acids far from any active sites⁵. While the generation of such comprehensive knowledge will likely take decades, studies on genotype-phenotype relationships is vitally important to build the solid foundations needed for the future. In this review, we present the history of attempts to edit and engineer the genome in order to test the influence of genotype on phenotype and discuss new methods which will carry this field of study into the future.

The power of phenotype prediction from genotype can already be seen, as information about certain specific mutations and their linkage to genetic disease can now be used in medical research⁶. Attempts have been made for several years to either select healthy fertilized eggs before artificial implantation or, more ambitiously, to correct enough cells in adult humans to mitigate the disease phenotype⁷. Additionally, clinical trials for treatment of genetic disease in adults are already underway⁸. However, while this is extraordinarily important work from a medical perspective, it is limited to cases in which a clear phenotype is observed and then linked with a causative genotype⁶. This linking is only available due to the vast numbers of healthy human genomes which can be compared to the genome in question in order to identify causative mutations, a situation that is widely inapplicable to other organisms or to non-disease phenotypes.

As befits a field of considerable interest, a large variety of approaches are currently being used to further scientific knowledge of genotype-phenotype relationships^{9,10}. The most developed are those using computational approaches in an attempt to predict the changes in protein structure and binding properties resulting from critical base pair changes in and around the active sites of enzymes⁵. Unfortunately, even these studies are unable to make consistent or accurate predictions of protein function based on genetic changes, and require the predicted changes to actually be made *in vivo* and extensively tested in order to verify the predictions⁵.

In this review, we describe the history of attempts to link genotype to phenotype focusing on the technological developments which enabled new generations of research. We begin with the first disruptive and limited editing capabilities and continue through to the recent development of genome editing utilizing clustered regularly interspaced short palindromic repeats (CRISPR). We then detail the recent modification of CRISPR, its 2-step application, and explain how it alleviates the problems inherent to the original method. We also comment on additional modern adaptations of CRISPR and their utility for the future of genome editing.

Impediments to the prediction of phenotype from genotype

One of the greatest obstacles to making phenotypic predictions from genotypic data has been the lack of ability to generate and test large numbers of edits *in vivo* in various organisms in order to assess the effects of mutations¹¹. This is necessary in order to easily form and test hypotheses regarding the impact of desired mutations on phenotype. Additionally, over the course of progress during the past several decades, it has become increasingly obvious that the genome is generally less tolerant to change than was previously imagined, implying that accurate information can only be obtained from methods that do not scar the genome¹². Another general issue is the problem of comparability between species¹³. Many of the

species that would be useful to be able to predict phenotypes for are not model organisms which have the existing tools for prolonged study¹⁴.

Many genome-editing techniques make unwanted changes to the genome which hinders any attempt to definitively link any intended genetic changes with phenotypic consequences¹². Some, during the course of editing, scar the genome such that unwanted changes are made to the region being investigated and often analyzing the effects of the scar are practically impossible¹². Other techniques avoid scarring by severely reducing the number of edits; as a result, cycles of hypothesis testing become too laborious for practical study¹⁵. Yet another common workaround is to make large numbers of edits in synthetic systems, for example on non-integrating plasmids¹⁶. This exposes experiments to inevitable plasmid copy-number fluctuations from cell to cell, and deprives studies of the ability to examine the effects of chromatin on the regions in question¹⁶. To be able to address all of these concerns, an ideal method must be able to create large numbers of scarless edits in endogenous genomic locations^{12,17}.

One laudable attempt to avoid the difficulty in making large numbers of genetic edits has been the development of directed evolution¹⁸. Directed evolution begins with the creation of a library of mutagenized DNA. The DNA region to be mutated usually corresponds to either an open reading frame or a specific region of a coding sequence. This library is then subjected to screening using a specific activity assay in order to identify mutations which can achieve the desired improvement in protein activity^{19–21}. Achieving the desired outcome typically requires multiple rounds of the mutagenize-and-screen approach. Depending on the assay to measure the effects of the introduced mutations on protein activity, directed evolution can be performed *in vitro* or *in vivo*. It is important to highlight that research using the directed evolution approach often focuses on understanding how mutations affect the structure and biochemical function of a specific protein. The vast majority of directed evolution studies do not attempt to predict the effect of mutations on the cellular phenotype. Using directed evolution, bacteria or yeast are mainly used for the expression of a large number of mutants for examining their biochemical catalytic/binding activity. Among the exceptions to this norm is a previous work published by Fridman *et al.*²² in which the authors generated PCNA mutants with increased affinity to different partners and then integrated these mutants into yeast to examine their effect on the cellular phenotype of DNA replication.

While the directed evolution approach has succeeded in creating novel genomic sequences, which can then be linked to phenotypic effects, the lack of its ability to “rationally” investigate the impact of individual mutations, either alone or in combination, limits the ability of this technique to answer questions which are required for accurate prediction of phenotype from genotype. In many ways, it may be said that the pursuit of the ultimate goal of predicting phenotype from genotype has been the development of new methods coming close and closer to the goal of large scale, scarless genome editing. As is often the case for scientific advances, many of these methods were of vital importance, but also uncovered more problems to be overcome.

Disruptive methods offered the first glimpses of function on a base-pair scale

The goal of successfully linking genotype to phenotype has existed in some form ever since DNA was established as the carrier of cellular information. As such, there have been a wide variety of attempts to identify the genetic regions most important for phenotypic expression. Many of these early attempts have obvious flaws by modern standards, but still greatly expanded scientific knowledge at the time.

One of the first, and crudest, methods of identifying DNA domains crucial for phenotypic expression was to make (or find by chance) a series of random mutations in a gene or promoter of interest and then determine which mutations had an effect on the output of the gene in question^{23,24}. This was used to great effect by Johnston and Davis in 1984²⁵ as they studied the yeast *GAL1/GAL10* bidirectional promoter. In order to determine which base pairs were actually needed for expression, they systematically made deletion mutants of the promoter on a plasmid and identified which deletions eliminated expression and which were still permissive. By doing so, they were able to correctly identify the region surrounding the first three activator sites within the promoter. While they were unable to determine exactly what about that region was crucial to expression, they were able to locate the region in question with high accuracy. Future groups would go on to identify the activator binding sites in question²⁶, and thanks to the earlier characterizations, they knew exactly where to look.

As technology progressed and identification of relevant regions of DNA could be determined with more precision, the influence of specific base pair changes could be understood. This step was crucial for the ability to definitively link base pair content to phenotypic output, but techniques of the time could not yet produce DNA with desired changes to the same extent that we can today. As such, the technique called saturation mutagenesis was developed, which would mutagenize each individual base pair within a region of interest separately²⁷. The changes could then be examined to see exactly which base pairs were important for the phenotype under study. This was demonstrated wonderfully by Myers *et al.* when studying the promoter region of the beta-globin gene in mouse cells²⁸. By obtaining promoters with virtually all of the 130 bp upstream of the transcription start site (TSS) and placing these in front of an amino acid marker on a plasmid, they were able to identify deleterious mutations, neutral mutations, and even two enhancing mutations. They found that the majority of these mutations were neutral, but that base pairs within three previously identified regulatory areas were critical for gene expression, allowing the identification of those binding sites. While the identification of crucial base pairs allowed a much finer level of genotypic detail to be achieved, the lack of customization created problems regarding the actual development and testing of hypotheses regarding those bases.

The most recent of the attempts to understand genotypic effects on phenotype without consideration of the endogenous genomic loci have utilized centromeric plasmid systems²⁹. These low-copy plasmids are retained in a stable manner in the cell, at least in yeast, and allow expression of genes without disrupting the organism's genome in any way³⁰. One recent exemplary work illustrating this technique was done by Sharon *et al.* in 2014¹⁶. The researchers were able to make thousands of different edits integrated into the yeast *GAL1/*

GAL10 promoter in low copy number plasmids and investigate the consequences of these mutations in cells possessing one of these. This allowed them to determine the impact of binding site number, and to a limited effect binding site position, on expression levels and noise from the GAL1 promoter. The main drawback to this approach, and to those which preceded it, was the unknown nature of the potential differences between expression from a plasmid vs. expression from the genome. In recent years, it has become increasingly clear that the impact of native chromatin environments on gene expression is critical to understanding the normal functioning of genes and promoters³¹.

Early attempts at endogenous genome editing enabled crucial, but limited, advances

The optimal solution to the problems mentioned in the previous section would be to edit the DNA at the endogenous genomic locus³². This would allow the structure of the endogenous chromatin structure to be taken into account, while also removing any other unwanted changes to the genome. In addition, copy number would no longer be an issue, and questions regarding expression from non-genomic locations would be rendered irrelevant. While several techniques were developed to fulfil these conditions^{33,34}, they all ultimately suffered from the same fundamental problem: the inefficiency of generating the large numbers of edits needed for functional hypothesis testing.

The first technique developed for the manipulation of endogenous genomic DNA was based on the principle of URA-FOA counterselection using 5-fluoroorotic acid (5-FOA)¹⁵. In the presence of the URA3 gene product, 5-FOA is converted to a toxic intermediate killing the URA3-expressing cells. This was developed and demonstrated by Boeke *et al.* who were able to replace endogenous genomic regions with synthetic constructs *in vivo*¹⁵. To do this, a URA3 gene cassette is integrated into a genomic region to edit, which enables a yeast cell to survive on media without uracil. After selection on –URA media, the cells are transformed again together with the edited content in growth media containing both uracil and 5-FOA so that only cells that replace the URA3 cassette with the edited region survive. Theoretically, any cells that survive in 5-FOA media are good candidates for successful editing, but in reality the number of false positives severely hinders the usefulness of this technique. The development of direct genome editing techniques led to a sharp decline in the use of this technique.

Genome editing using zinc finger nucleases (ZFNs) was another promising technique around the turn of the millennium³⁵. The idea of cutting DNA near a region of interest, providing the cell with donor DNA containing the desired changes, and allowing the cell to repair the break with homologous recombination using the donor has offered a tantalizing possibility for direct genomic editing. Ever since DNA synthesis became cost-efficient, the remaining difficulty in developing and utilizing this method was actually causing a double strand break in the correct location in the genome. This problem was overcome by using zinc finger nucleases in 2000³⁶. ZFNs are a class of artificial proteins containing multiple zinc finger motifs bound to a nuclease. Each motif recognizes and binds to a nine base pair DNA sequence, and the nuclease then cleaves the DNA at that location. In 2003, Bibikova *et al.* were able to modify the DNA recognition region of a ZFN to target a sequence in the drosophila genome and cleave it *in vivo*³³. For the first time, they also demonstrated that

addition of DNA containing homology to the cut region could be used for repair by the cell, and showed that this could lead to genome editing with considerable efficiency. This allowed scarless editing of *in vivo* locations via hijacking of the cell's DNA repair pathways. Though ZFNs had their moment in the sun, and were still being developed prior to the introduction of CRISPR³⁷, it required a large amount of time and effort to create new ZFNs (usually via directed evolution) and they were not as versatile as they were needed to be.

In the first decade of the 21st century, the main competitor to ZFNs was the development of transcription activator-like effector nucleases (TALENs)³⁸. As with ZFNs, the goal was to be able to program a protein in order to target a specific DNA sequence for cleaving. TALENs were based on a similar idea to ZFNs, but instead of using the bulky zinc finger groups, they used smaller DNA recognition proteins (TALEs) which recognized 3-bp sequences³⁴. Christian et al. created the first of these TALEs fused to a nuclease in order to make the first TALEN which would both bind and cut DNA. At the same time, they demonstrated the ability to create novel TALEs (which were then made into TALENs), and opened the door for the production of molecules which could cut DNA in any desired location. While these showed promise, they were still unwieldy to construct and they ultimately met the same fate as the ZFNs as CRISPR was introduced.

CRISPR has revolutionized genome editing at the cost of the reintroduction of an old flaw

The problem of unwieldy double strand break induction was finally solved, for the most part, in 2012 with the development of clustered regularly interspaced short palindromic repeats (CRISPR), and it seemed as though mapping genotype to phenotype would finally become a reality. While CRISPR eliminated almost all of the remaining problems associated with genome editing, it also reintroduced a few older ones, which would prove difficult to eliminate.

The process of CRISPR development began in 2007 with the elucidation of an interesting feature of the immune system for a wide variety of bacterial species^{39,40}. In brief, the cells were shown to have incorporated short (20bp) sequences in the aftermath of viral invasions in long regions of DNA which could be expressed as RNA, cleaved, assembled into a final product with a second RNA, and bound to a protein called Cas^{9,40,41}. Cas9 is a nuclease and uses the RNA fragments to locate targets based on the RNA sequence^{40,41}. The identified DNA sequence (the complement of the initial RNA fragment) is cleaved and degraded by the cell. Initially, this was regarded as an interesting feature of a few bacterial families^{39,41}, but it was quickly realized by multiple groups that such a system could be imported into other cell types and used as a method to create double strand breaks^{11,42,43}. Because the targeting of Cas9 to the DNA is directed by an RNA strand instead of unwieldy protein complexes, the CRISPR system showed immediate promise to overcome the challenges associated with the previous genome editing methods which relied on hard-to-design protein complexes.

The use of CRISPR as a genome editing technique was first demonstrated in bacterial systems¹¹, followed by a variety of studies showing its efficacy in various cell types^{44–47}, most prominently in mammalian cells^{43,47}. Multiple publications quickly demonstrated an ability to fuse the two RNA molecules used for natural CRISPR function into a single guide

RNA (gRNA)^{42,43} targeting a specific DNA site. This breakthrough reduced the number of components needed for CRISPR genome editing to three: the Cas9 nuclease, a gRNA, and a donor oligonucleotide. Almost immediately after this, CRISPR using gRNA targeting was shown to work extremely well in *Saccharomyces Cerevisiae* by DiCarlo *et al.*⁴⁴. The editing efficiency of the technique was shown to be ~75% or more, even when demonstrating multiplexable systems^{12,48,49}. Although such early papers primarily caused deletions or other disruptions of genes rather than true editing via the addition of a donor repair template, true editing was quickly demonstrated as well⁴³. This technology has also been combined with other well-studied systems, including an attempt to control transcriptional activity via transposons⁵⁰.

Despite the enormous successes of CRISPR, and its rapid dominance of the genome editing field, it is not without flaws. These are twofold, and are consequences of the way the Cas enzymes work: any edits must disrupt the CRISPR cut or protospacer adjacent motif (PAM) site being used, and edits cannot be made more than 50–100 bp away from any cut^{12,17,48,49}. The first of these is due to the structure of the CRISPR system itself. Because the Cas9 enzyme will cut any DNA with the correct sequence, if a donor is provided bearing the gRNA targeting sequence, it will either be cut and degraded prior to repair, or the repaired chromosome will be recut until it repairs incorrectly (Fig. 2A). This issue was recognized almost immediately after the initial discovery of CRISPR, but it was practically ignored as most applications of CRISPR are still limited to the disruption of genes and for editing of protein coding regions^{12,49} where there is room for imperfect editing due to codon degeneracies. For such applications, the disruptions are either irrelevant, or can be changed to create mutations which result in identical amino acid codons while still disrupting further CRISPR action¹². For other applications, however, the use of CRISPR results in a genomic scar due to the disruption of the targeted cut site. The second flaw of the CRISPR technique was more complex and varied between organisms: the tendency of the cell to use an imperfect repair mechanism (i.e. non-homologous end joining) if the desired edit is too far from the cut site⁴⁹. This has not traditionally been a problem except for the case in which an appropriate cut site cannot be found close to a region of interest. However, if a variety of edits are desired across a relatively large area (100s-1000s of bps), its completion requires multiple rounds of cutting and editing⁴⁹.

2-step CRISPR as a precise and efficient genome editing method to edit large DNA regions

The recently developed modification of CRISPR, known as 2-step CRISPR⁵¹ solves the above-mentioned problems at the cost of an additional editing step. Rather than giving the cut DNA strand a donor template containing the desired edits, two gRNAs are designed to have Cas9 cut in locations flanking a region of interest which can be multiple kb long. A donor is then introduced to replace the region of interest with a small (30–50 bp) sequence containing a novel CRISPR cut site and PAM sequence which is unique in the genome (Fig 2B). Software developed to identify ideal CRISPR targets can be used to generate these novel sequences⁵². After the cell successfully repairs itself using the donor template, the gRNAs are naturally degraded over a short period of time. At this point, any DNA sequence can be inserted into the original region of interest during a second round of editing as long as the cell can tolerate the genomic change introduced at the end of the first round. A single

gRNA targeting the introduced cut site allows for repair using any template desired, with the end result being the replacement of the original region with the desired sequence. By temporally separating the first gRNAs from the desired product, the final edit is no longer targeted by Cas9 and genomic scarring is prevented.

The major benefit of this technique as opposed to other forms of CRISPR is that it allows totally scarless genome editing on a fast timescale across multiple kilo bases of genomic DNA. By preventing the initial gRNAs and the final edited donor from being in the cell at the same time, the problem of the gRNA cutting the donor can be completely avoided. This means that a donor can be provided for the final editing step which would have otherwise been unable to be used with a single-step editing process, thus ensuring that the final edit does not need to have its gRNA binding site or PAM sequence modified. Thus, the final edited version of the genome contains only the desired edits. As a useful side benefit of this technique, the region to be edited is not limited to any significant degree by proximity to a CRISPR cut site. In addition, edits that are many hundreds of base pairs apart from each other may be introduced in a single round of editing. Taken together, this technique allows researchers to achieve scarless genome editing to investigate genotype-phenotype relationships on a base pair level.

The only undesirable aspect of the 2-step CRISPR method is that it requires an intermediate step between the initial CRISPR cuts and the final repair. In single celled organisms or cultured cells, this is only a problem if the region being worked on is lethal if it is knocked out. Otherwise, it is straightforward to keep strains alive with the intermediate stage intact in the genome. Even if the user wishes to make edits throughout the genome, the time for successful editing is merely doubled for each new location and this is generally not excessively time-limiting. For editing needs that must be achieved in a single step, the technique as it currently exists would have significant difficulty in keeping the two steps spatially separated. There is also the possibility of loss of epigenetic information during editing, but it is not yet known if this will be a problem.

An interesting variant of this approach was recently published in which the authors applied the same two-step strategy in a slightly different manner. In the first step of the technique, instead of replacing the region of interest with a novel CRISPR cut site, Soreanu *et al.*⁵³ replaced it with traditional antibiotic markers. The second step of the technique was the same as in the original 2-step technique: using gRNAs to target the introduced marker, the authors cut out the marker and integrated the desired DNA content to the region of interest. The end result was the same as the one achieved with the original 2-step CRISPR method. The use of a marker rather than a novel CRISPR cut site may allow for easier selection after the first step of the method.

Editing without double strand breaks

While 2-step CRISPR successfully avoids the scarring problem inherent to native CRISPR, recent studies have attempted to overcome this problem in another manner: making edits without a double strand break at all. While these may supplant editing via CRISPR-induced double strand breaks in the future, for the moment they come with their own set of challenges.

The first of these utilizes dCas9, a version of Cas9 which is able to bind DNA as usual but which has lost its nuclease activity⁵⁴. dCas9 may be fused or tethered to a vast number of other proteins in order to bring these to a specific genomic location. This protein has been revolutionary for a number of different fields much as the nuclease-active Cas9 has been for genome editing. In 2016, Komor *et al.* tethered a cytidine deaminase to dCas9 *in vivo*⁵⁵. Cytidine deaminases act on cytosine bases and convert them to uracil (and eventually thymine) bases. By bringing these deaminases to a specific genomic locus, the group was able to convert nearby C-G base pairs to T-A base pairs without cutting the DNA in a process they called 'base editing'. They later found that by using a Cas9 nickase rather than dCas9 they could trick the cell into believing that the G was the incorrect base rather than the T in the mismatch, greatly increasing the efficiency of the technique.

The great benefit of this approach is that scarring due to DNA repair is avoided as the DNA is changed without a double strand break; however, the technique has a long way to go before it can be as versatile as other forms of CRISPR editing. The first problem is that the chemistry that converts C to T is relatively easy to perform and enzymes exist which catalyze the reaction. To make different types of edits, enzymes need to be found, or in the worst case, designed and expressed for this approach to work. In addition, only a single edit may be made at a time, meaning that if multiple base pair mutations are desired, an equivalent number of locations must be targeted. The authors also encountered off-target mutations of cysteines in close proximity to the target. This poses obvious problems if attempting to edit a G-C rich region. Despite these drawbacks, this method is superior to traditional CRISPR in a certain limited set of circumstances, and improvements to the versatility of the technique could make it a serious alternative to both traditional CRISPR and its 2-step application.

Another recent attempt to make genomic edits without double strand breaks relies on the development of multiplex automated genome engineering (MAGE) in eukaryotes (eMAGE)⁵⁶. Although the concept has existed for several years in *E. coli*⁵⁷, it was only recently demonstrated by Barbieri *et al.*⁵⁶. Briefly, the technique utilizes small synthetic DNA sequences which act as Okazaki fragments during DNA replication and which are able to impart their edits to the newly formed DNA strands. Several rounds of editing may introduce a large number of mutations to a region with high efficiency. In this way, a region may have a large number of edits introduced without cutting of the DNA and without the presence of substantial off-target effects.

While its ability to generate large-scale diversity of sequence is impressive, there are still some drawbacks associated with this technique, as for most of the new techniques. For example, unless large numbers of editing rounds are performed, the cells which are obtained at the end of the experiment will contain only some of the desired edits and separating the populations containing identical edits from each other is very challenging unless each edit results in a clearly selectable phenotype. However, it should be noted that precise genome editing may not end up being the most common application area of this technique. Instead, eMAGE is capable of generating exceptionally large amounts of genetic diversity over one or a few regions in a way that no other method can currently match. While CRISPR variants are superior for specific editing, eMAGE is able to make up for the major deficiencies of

CRISPR when it comes to generating diversity. The technique is extremely well-tailored for a variety of purposes, and using it together with the 2-step CRISPR can have a much greater effect than the use of either technique separately.

CONCLUSION

After decades of work attempting to link genotype to phenotype, the research community finally has the means to better understand genotype-phenotype relationship by making large-scale scarless genome editing and studying their phenotypic consequences. We anticipate that the remaining challenges, such as the ones associated with editing essential genes and working with difficult-to-edit organisms, will be overcome in the near future. For example, in the context of improving the current state of the 2-step CRISPR technique, essential genes could be temporarily replaced during the editing process through the addition (e.g. via plasmids) of the genes together with their own promoters, but containing a single base pair mutation to prevent their being targeted by the gRNA in use. These could be introduced prior to the first cutting step and removed after the second editing step when the cell/organism should be able to survive on its own. It may also be possible to put the CRISPR components associated with both steps of the editing process into a cell and separate them temporally instead of spatially. We envision a system in which the reactions of the first editing step to be activated first, while the other is protected, then have the second step's reactions proceed afterwards.

The application areas for the scarless genome editing techniques are boundless. For example, there is currently a great need for better genetic engineering of crop plants. From curing of disease at a genetic level to pest resistance, there is a whole host of possibilities when it comes to making crop plants better at what they do and making what they do more useful to us in the process^{58,59}. Genetic diseases are, of course, not limited to plants, and a wide variety of them plague the human population today, some of which have known causes and many of which do not⁶⁰. While several with known causes are actively being targeted for treatment with existing technologies, the ability to understand exactly how phenotypic consequences result from genomic mutations would be groundbreaking for the medical community.

After a long history of methods development which provided more efficient and less disrupting genome editing, the refinement of the CRISPR technique create a newfound promise going forward. This should, for the first time, enable the kind of large-scale testing on endogenous genomic locations needed to begin to investigate the relationship between genotype and phenotype at a base-pair resolution.

GLOSSARY TERMS

Directed Evolution

A process in which a specific DNA element (often an open reading frame or enzymatic binding site on a protein) is mutagenized in order to create a large library. This library is then put through a specific *in vitro* screening process, generally for multiple rounds, in order to identify mutations which give rise to a desired effect on the assayed activity or function.

DNA Binding Motif

A region of a protein which is capable of binding to a specific DNA sequence

Donor Template

An (often) short piece of DNA containing desired DNA edits flanked by regions homologous to the region to be edited. Cells may use it as a repair template during homologous-recombination based DNA-repair mechanism

Double Strand Break

A type of DNA damage in which both strands have been severed in close proximity to each other. This results in the creation of two distinct strands from the original

Endogenous DNA

The native DNA of a cell

Genome Editing

The process of intentionally altering (adding, removing, or changing) the genome of a cell

Genome Scarring

The introduction of unwanted or unintended edits to cellular DNA during genome editing

Genotype

The DNA composition of a genomic region or of the entire genome

Homologous recombination based DNA repair

A type of DNA damage repair mechanism in which the cell uses a homologous region to the one which has been cut as a template to repair the cut strand. During natural repair, this template is generally the sister chromosome of the one being cut, but during genome editing this template is generally synthetic DNA introduced to the cell for this purpose

Mutagenesis

The process of inducing random mutations to a piece of DNA or to an entire genome. Historically this has been done using radiation or mutagenic chemicals. A mutagenized stretch of DNA will contain a number of random mutations of all types

Nuclease

A protein which is capable of inducing a DNA double strand break

Phenotype

A behavior of a cell which is observable to a researcher

Plasmid-based expression system

A means of gene expression in which the gene of interest is placed into a non-integrating plasmid and transformed into a cell

REFERENCES

1. Mukherji S & van Oudenaarden A Synthetic biology: understanding biological design from synthetic circuits. Nature reviews. Genetics 10, 859–871, doi:10.1038/nrg2697 (2009).

2. Purnick PE & Weiss R The second wave of synthetic biology: from modules to systems. *Nature reviews. Molecular cell biology* 10, 410–422, doi:10.1038/nrm2698 (2009). [PubMed: 19461664]
3. Andrianantoandro E, Basu S, Karig DK & Weiss R Synthetic biology: new engineering rules for an emerging discipline. *Molecular systems biology* 2, 2006 0028, doi:10.1038/msb4100073 (2006). [PubMed: 16738572]
4. Taft RJ, Pheasant M & Mattick JS The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays : news and reviews in molecular, cellular and developmental biology* 29, 288–299, doi:10.1002/bies.20544 (2007).
5. Dill KA & MacCallum JL The Protein-Folding Problem, 50 Years On. *Science* 338, 1042–1046 (2012). [PubMed: 23180855]
6. Hirschhorn JN, Lohmueller K, Byrne E & Hirschhorn K A comprehensive review of genetic association studies. *Genetics in Medicine* 4, 45–61 (2002). [PubMed: 11882781]
7. Miller HI Germline gene therapy: We're ready. *Science* 348, 1325, doi:10.1126/science (2015).
8. Naldini L Gene therapy returns to centre stage. *Nature* 526, 351–360, doi:10.1038/nature15818 (2015). [PubMed: 26469046]
9. Ritchie MD, Holzinger ER, Li R, Pendergrass SA & Kim D Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics* 16, 85–97, doi:10.1038/nrg3868 (2015).
10. Bush WS, Oetjens MT & Crawford DC Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature reviews. Genetics* 17, 129–145, doi:10.1038/nrg.2015.36 (2016).
11. Jiang W, Bikard D, Cox D, Zhang F & Marraffini LA RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* 31, 233–239, doi:10.1038/nbt.2508 (2013).
12. Mans R et al. CRISPR/Cas9: a molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in *Saccharomyces cerevisiae*. *FEMS yeast research* 15, 1–15, doi:10.1093/femsyr/fov004 (2015).
13. Sittig LJ et al. Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. *Neuron* 91, 1253–1259, doi:10.1016/j.neuron.2016.08.013 (2016). [PubMed: 27618673]
14. Weinhandl K, Winkler M, Glieder A & Camattari A Carbon source dependent promoters in yeasts. *Microbial Cell Factories* 13, 1–17 (2014). [PubMed: 24387764]
15. Boeke JD, Trueheart J, Natsoulis G & Fink GR . 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods in Enzymology* 154, 164–175 (1987). [PubMed: 3323810]
16. Sharon E et al. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome research* 24, 1698–1706, doi:10.1101/gr.168773.113 (2014). [PubMed: 25030889]
17. Ryan OW, Poddar S & Cate JH CRISPR-Cas9 Genome Engineering in *Saccharomyces cerevisiae* Cells. *Cold Spring Harbor protocols* 2016, 525–533, doi:10.1101/pdb.prot086827 (2016).
18. Arnold FH Engineering proteins for nonnatural environments. *The FASEB Journal* 7, 744–749 (1993). [PubMed: 8330682]
19. Shao Z & Arnold FH Engineering new functions and altering existing functions. *Current Opinion in Structural Biology* 6, 513–518 (1996). [PubMed: 8794158]
20. Packer MS & Liu DR Methods for the directed evolution of proteins. *Nature reviews. Genetics* 16, 379–394, doi:10.1038/nrg3927 (2015).
21. Renata H, Wang ZJ & Arnold FH Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution. *Angewandte Chemie* 54, 3351–3367, doi:10.1002/anie.201409470 (2015). [PubMed: 25649694]
22. Fridman Y et al. Subtle alterations in PCNA-partner interactions severely impair DNA replication and repair. *PLoS biology* 8, e1000507, doi:10.1371/journal.pbio.1000507 (2010). [PubMed: 20967232]
23. Benoist C & Chambon P In vivo sequence requirements of the SV40 early promoter region. *Nature* 290, 304–310 (1981). [PubMed: 6259538]
24. DOUGLAS HC & CONDIE F THE GENETIC CONTROL OF GALACTOSE UTILIZATION IN *SACCHAROMYCES*. *Journal of Bacteriology* 68, 662–670 (1954). [PubMed: 13221541]

25. Johnston M & Davis RW Sequences That Regulate the Divergent GALJ-GALIO Promoter in *Saccharomyces cerevisiae*. *Molecular and cellular biology* 4, 1440–1448 (1984). [PubMed: 6092912]
26. BRAM RJ & KORNBERG RD Specific protein binding to far upstream activating sequences in polymerase II promoters. *Proceedings of the National Academy of Sciences of the United States of America* 82, 43–47 (1985). [PubMed: 3881758]
27. Myers RM, Lerman LS & Maniatis T A General Method for Saturation Mutagenesis of Cloned DNA Fragments. *Science* 229, 242–247 (1985). [PubMed: 2990046]
28. MYERS RM, TILLY K & MANIATIS T Fine Structure Genetic Analysis of a β -Globin Promoter. *Science* 232, 613–618 (1986). [PubMed: 3457470]
29. Sharon E et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* 30, 521–530, doi:10.1038/nbt.2205 (2012).
30. Elledge SJ & Davis RW A family of versatile centromeric vectors designed for use in the sectoring-shuffle mutagenesis assay in *Saccharomyces cerevisiae*. *Gene* 70, 303–312 (1988). [PubMed: 3063604]
31. Carey LB, van Dijk D, Sloot PM, Kaandorp JA & Segal E Promoter sequence determines the relationship between expression level and noise. *PLoS biology* 11, e1001528, doi:10.1371/journal.pbio.1001528 (2013). [PubMed: 23565060]
32. SCHERER S & DAVIS RW Replacement of chromosome segments with altered DNA sequences constructed in vitro. *Proceedings of the National Academy of Sciences of the United States of America* 76, 4951–4955 (1979). [PubMed: 388424]
33. Bibikova M, Beumer K, Trautman JK & Carroll D Enhancing gene targeting with designed zinc finger nucleases. *Science* 300, 764, doi:10.1126/science.1079512 (2003). [PubMed: 12730594]
34. Christian M et al. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* 186, 757–761, doi:10.1534/genetics.110.120717 (2010). [PubMed: 20660643]
35. Urnov FD, Rebar EJ, Holmes MC, Zhang HS & Gregory PD Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics* 11, 636–646, doi:10.1038/nrg2842 (2010).
36. Smith J et al. Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. *Nucleic acids research* 28, 3361–3369 (2000). [PubMed: 10954606]
37. Miller JC et al. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* 25, 778–785, doi:10.1038/nbt1319 (2007).
38. Joung JK & Sander JD TALENs: a widely applicable technology for targeted genome editing. *Nature reviews. Molecular cell biology* 14, 49–55, doi:10.1038/nrm3486 (2013). [PubMed: 23169466]
39. Barrangou R et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712, doi:10.1126/science.1138140 (2007). [PubMed: 17379808]
40. Gasiunas G, Barrangou R, Horvath P & Siksnys V 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, E2579–E2586, doi:10.1073/pnas.1208507109 (2012). [PubMed: 22949671]
41. Brouns SJ et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321, 960–964, doi:10.1126/science.1159689 (2008). [PubMed: 18703739]
42. Jinek M et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821, doi:10.1126/science.1225829 (2012). [PubMed: 22745249]
43. Mali P et al. RNA-guided human genome engineering via Cas9. *Science* 339, 823–826, doi:10.1126/science.1232033 (2013). [PubMed: 23287722]
44. DiCarlo JE et al. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic acids research* 41, 4336–4343, doi:10.1093/nar/gkt135 (2013). [PubMed: 23460208]
45. Hruscha A et al. Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development* 140, 4982–4987, doi:10.1242/dev.099085 (2013). [PubMed: 24257628]

46. Li-En Jao SRW, and Wenbiao Chen Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proceedings of the National Academy of Sciences of the United States of America* 110, 13904–13908 (2013). [PubMed: 23918387]
47. Wang T, Wei JJ, Sabatini DM & Lander ES Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 343, 80–84, doi:10.1126/science.1246981 (2014). [PubMed: 24336569]
48. Horwitz AA et al. Efficient Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas. *Cell systems* 1, 88–96, doi:10.1016/j.cels.2015.02.001 (2015). [PubMed: 27135688]
49. Ryan OW et al. Selection of chromosomal DNA libraries using a multiplex CRISPR system. *eLife* 3, e03703, doi:10.7554/eLife.03703 (2014).
50. Vaschetto LM Modulating signaling networks by CRISPR/Cas9-mediated transposable element insertion. *Current genetics* 64, 405–412, doi:10.1007/s00294-017-0765-9 (2018). [PubMed: 29032444]
51. Elison GL, Song R & Acar M A Precise Genome Editing Method Reveals Insights into the Activity of Eukaryotic Promoters. *Cell reports* 18, 275–286, doi:10.1016/j.celrep.2016.12.014 (2017). [PubMed: 28052256]
52. Stemmer M, Thumberger T, Del Sol Keyer M, Wittbrodt J & Mateo JL CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PloS one* 10, e0124633, doi:10.1371/journal.pone.0124633 (2015). [PubMed: 25909470]
53. Soreanu I, Hendler A, Dahan D, Dovrat D & Aharoni A Marker-free genetic manipulations in yeast using CRISPR/CAS9 system. *Current genetics*, doi:10.1007/s00294-018-0831-y (2018).
54. Qi LS et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152, 1173–1183, doi:10.1016/j.cell.2013.02.022 (2013). [PubMed: 23452860]
55. Komor AC, Kim YB, Packer MS, Zuris JA & Liu DR Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424, doi:10.1038/nature17946 (2016). [PubMed: 27096365]
56. Barbieri EM, Muir P, Akhuetie-Oni BO, Yellman CM & Isaacs FJ Precise Editing at DNA Replication Forks Enables Multiplex Genome Engineering in Eukaryotes. *Cell* 171, 1453–1467 e1413, doi:10.1016/j.cell.2017.10.034 (2017). [PubMed: 29153834]
57. Wang HH et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460, 894–898, doi:10.1038/nature08187 (2009). [PubMed: 19633652]
58. Collinge DB, Lund OS & Thordal-Christensen H What are the prospects for genetically engineered, disease resistant plants? *European Journal of Plant Pathology* 121, 217–231, doi: 10.1007/s10658-007-9229-2 (2008).
59. Fraser PD, Enfissi EM & Bramley PM Genetic engineering of carotenoid formation in tomato fruit and the potential application of systems and synthetic biology approaches. *Archives of biochemistry and biophysics* 483, 196–204, doi:10.1016/j.abb.2008.10.009 (2009). [PubMed: 18948075]
60. Friedmann T & Roblin R Gene Therapy for Human Genetic Disease? *Science* 175, 949–955 (1972). [PubMed: 5061866]

HIGHLIGHTED REFERENCES

- Johnston and Davis (1984) – While not the first to identify important promoter regulatory regions based on mutagenesis assays, this article provided an extremely clear account of the process conducted on the canonical yeast GAL1 promoter.
- Myers et al. (1986) – This paper used saturation mutagenesis to locate the important base pairs involved in the regulation of the mouse β -globulin gene and is an example of the first generation of studies able to discern exactly which base pairs were critical for expression from a promoter.
- Sharon et al. (2014) – This is one of the first papers to make truly large-scale edits to a promoter region in order to associate changes in base pair structure with changes in phenotypic output, thus allowing for useful genotype-phenotype relationships to be discerned.

- Boeke et al. (1987) – This study demonstrated that URA3–5FOA counterselection could be used to scarlessly edit endogenous yeast genomes well before the introduction of editing via the repairing of double strand breaks.
- Smith et al. (2000) – This work created the first zinc finger nucleases by fusing a nuclease to a zinc finger DNA binding protein and represented the start of many attempts to use double strand breaks to edit endogenous DNA regions.
- Christian et al. (2010) – This article demonstrated the viability of attaching a nuclease to TALE proteins as a way to induce double strand breaks using more pliable guides than ZFNs.
- Barrangou et al. (2007) – This was the first work to describe the CRISPR/Cas system as a form of bacterial immune system and led the way for the adaption of the system to be used for genome editing.
- Jinek et al. (2012) – The authors successfully transformed CRISPR into a functional genome editing method and demonstrated the usefulness of the technique in bacteria.
- Mali et al. (2013) – This paper was the first to create a widely useable guide RNA for use in CRISPR editing and was one of the first to successfully edit eukaryotic genomes in vivo.
- Elison et al. (2017) – This paper was able to eliminate the scarring problem associated with traditional CRISPR by introducing a second editing step, and in doing so paved the way for a new generation of techniques able to further the understanding of the interaction between genotype and phenotype

Figure 1. The importance of understanding genotype/phenotype relationships.

Greater understanding of the relationship between genotype and phenotype is critical for a large number of fields. At the same time, advances in these fields contribute to greater knowledge which can be applied elsewhere. This prompts a multi-element feedback loop in which improvements in one area may be applied across a variety of fields.

Figure 2. The advantage of 2-step CRISPR.

A. Traditional CRISPR methods are not scarless. If the donor template provided for genome editing contains the same cut site and PAM sequence which were originally cut, both the donor and the final edit may be cut again. This will continue until the donor has been completely degraded and/or the double strand break repairs incorrectly. In situations where it is undesirable to alter the cut site, this style of CRISPR editing fails. **B.** 2-Step CRISPR is able to provide totally scarless genome editing. 2-Step CRISPR works in two steps in order to temporally separate the initial cutting of the genome with the repair using a donor template. The genome is cut in two locations just outside of the region to be edited and is replaced using a donor template containing only a novel CRISPR cut site. This strain is isolated and then undergoes a second round of editing. In this step the novel cut site is targeted and cut and the edited region of interest is added as the donor. Because the original cut sites are no longer being targeted they may be maintained in the donor without edit.