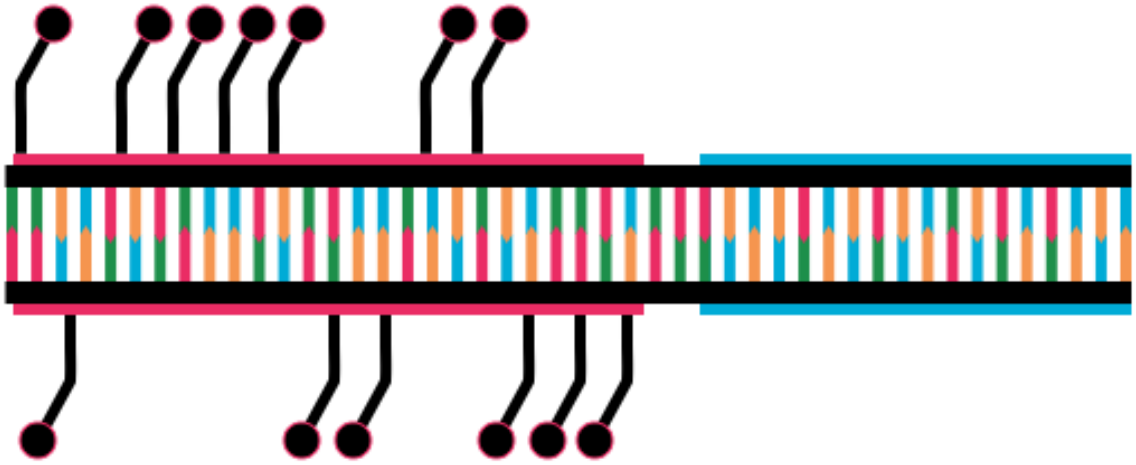
Functional Validation of Noncoding DNA with CRISPR Tiling

Søren Hough
November 15, 2016



CRISPR is quickly becoming the go-to tool for multiplex functional mapping of the genome. In many cases, this application has largely focused on better understanding protein-coding DNA — yet the majority of DNA, sometimes called “the dark genome,” is never translated. As with functional domain mapping in proteins, researchers are now turning to CRISPR to explore the dark genome, too.

Untapped Potential of the Dark Genome

Noncoding DNA makes up 98% of the human genome (Elgar & Vavouri 2008) and is poorly conserved across species, even within common taxa (vertebrates, placental mammals) (Sanjana et al. 2016). This leaves the function of the noncoding genome function largely a mystery. In recent years, researchers have probed this DNA and concluded that the noncoding genome plays an indirect role on gene expression via epigenetic state, chromatin accessibility, transcription factor binding, evolutionary conservation, 3D structure and more (Sanjana et al. 2016).

However, functional genomic studies to date have largely focused on the better conserved protein coding genome. Understanding the regulatory function of noncoding DNA offers an unprecedented perspective on the differences between cell lines and even individual organisms in a population. The more we learn about patient-to-patient variation from national genomics initiatives, the clearer it has become that we need to better understand the whole genome — not just the 2% that codes for protein.

The Biological Limitations of Reporter Assays

Studies around the noncoding genome have mostly relied on reporter assays. In these assays, putative enhancers are removed from the model organism, inserted in front of a reporter (e.g. luciferase) and function is based on a fluorescence-based readout (Melnikov et al. 2012). These have been instructive in illuminating the purpose and function of enhancers and other noncoding elements affecting gene regulation. Unfortunately, these studies also come with caveats.

Parallel reporter assays remove enhancers from their endogenous contexts in cell lines and organisms. This is a key point when discussing the noncoding genome which is inherently intertwined with the 3D structure of DNA. This may explain why some validated enhancers lose their function when removed from their endogenous context (Korkmaz et al. 2016). Further, it is estimated that more than 500,000 enhancers exist in the human genome. Although efforts have been made to expand the capacity of reporter assays (Melnikov et al. 2012), lack of biological context remains a barrier to accurately mapping these regions.

CRISPR Provides High-Throughput, Contextualized Genomic Interrogation

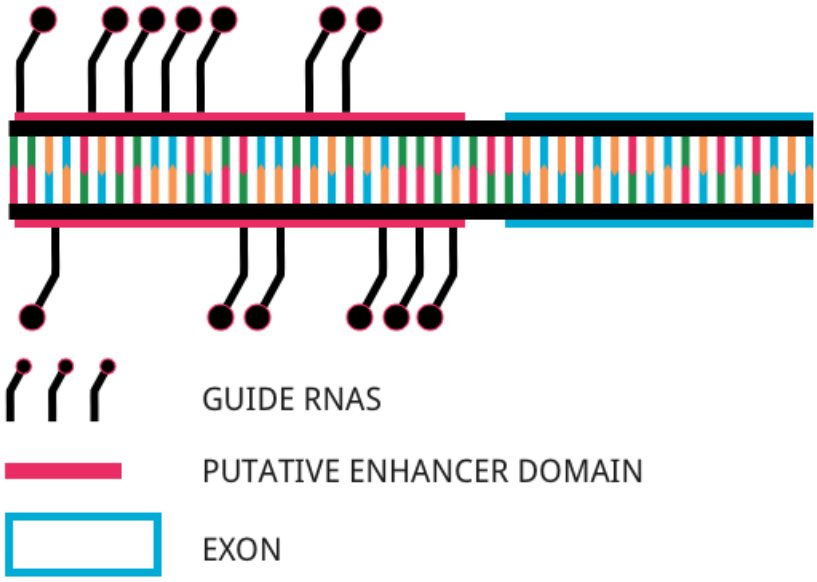


Figure 1. *In a CRISPR tiling experiment, sgRNAs are designed to target (or saturate) as much of the putative regulatory region as possible. This way, multiple guides can be mapped to specific functional loci to elucidate the function of noncoding DNA.*

The advent of CRISPR genome editing presents solutions to both of these problems. Researchers can now use multiplex CRISPR designs (known as CRISPR libraries) that saturate noncoding regions with double-stranded breaks to help elucidate function. This process is known as *in situ* saturating mutagenesis or, more simply, CRISPR tiling (Figure 1).

Further, CRISPR has already been used to validate trends found in genome-wide association studies (GWAS). For example, in 2016, Giani et al. used GWAS data identifying a SNP in the *SH2B3* gene that seemed to be linked to hematopoiesis. The group then showed that *SH2B3* knockdown (shRNA) and knockout (CRISPR) increased red blood cell production. This validated the population variant as a loss-of-function mutation. Expanding the scope of this kind of study to cover a broader range of the genome, particularly regulatory regions, may provide new insight and options for therapeutic development.

Multiple Approaches to CRISPR Tiling

Executing a CRISPR tiling experiment begins by choosing a region of interest. This can be done in several ways depending on the desired outcome of the study. In some cases, this involves searching through the genome to find known regulatory regions: transcription factor binding sites, for example. In others, the investigator may focus on a gene or small set of genes and tile along cis-noncoding regions to determine their effect on expression.

GWAS data and genetic linkage studies can be instructive in helping to narrow the focus of the functional analysis. In a 2015 study, Canver et al. used this approach to find two single nucleotide polymorphisms (SNPs) associated with β-hemoglobin disorder. These SNPs were located in a known intronic *BCL11A* enhancer at putative DNase I hypersensitive sites. This co-localization was instructive; *BCL11A* is known to play a role in the conversion of fetal hemoglobin (HbF) to adult hemoglobin (HbA), a process that goes awry in thalassemia and sickle cell disease (Orkin 2016). A mutation that impacts the expression of *BCL11A* may provide insight into disease pathogenesis and offer a new direction for clinical research.

The team first used CRISPR to delete the ~12 kb enhancer from the host genome and found that this yielded therapeutic effects by increasing the level of HbF in human cells. However, this solution presented several problems. Large deletions can lead to inversions (leaving orientation agnostic enhancers fully functional) and unpredictable indels. Further, the effect of deleting enhancers wholesale on gene expression is drastic; a more ideal solution would be to tweak regulatory regions to modulate phenotype. This could reveal a more precise therapeutic effect.

To pinpoint these regulatory regions, the team put forth the idea of a “composite enhancer” composed of both “essential and dispensable” loci. Canver et al. set about tiling the 12 kb region with a pooled lentiviral CRISPR library in HUDEP-2 cells. They designed 533 sgRNAs for all canonical (NGG) SpCas9 PAM sites in three ~1.2 kb hypersensitive site-adjacent regions in the enhancer. 49 positive control guides were designed to target Exon 2 of *BCL11A* (intended to yield drastic gene knockout) while 120 non-targeting guides served as negative controls.

With this as a background, Canver et al. found that BCL11A protein levels decreased the most in the presence of particular sgRNAs. This helped them map enhancer functions to specific motifs in the enhancer. In particular, they discovered the sgRNAs that targeted the same DNase I hypersensitive sites identified in the GWAS data yielded

significant *BCL11A* depletion. For the sgRNA that led to the most pronounced phenotype, the group verified the effect in a separate human cell culture.

Editing The Noncoding Genome *In Vivo*

Canver et al. then looked to replicate this result *in vivo*. Unfortunately, noncoding DNA tends to vary significantly between distantly related species. Once again, the team generated a lentiviral library for use *in vitro* with mouse cells to identify the differences. Indeed, the team found that out of the three known DNase I hypersensitive sites found in the human (and primate) BCL11A enhancer, only two were conserved in mice.

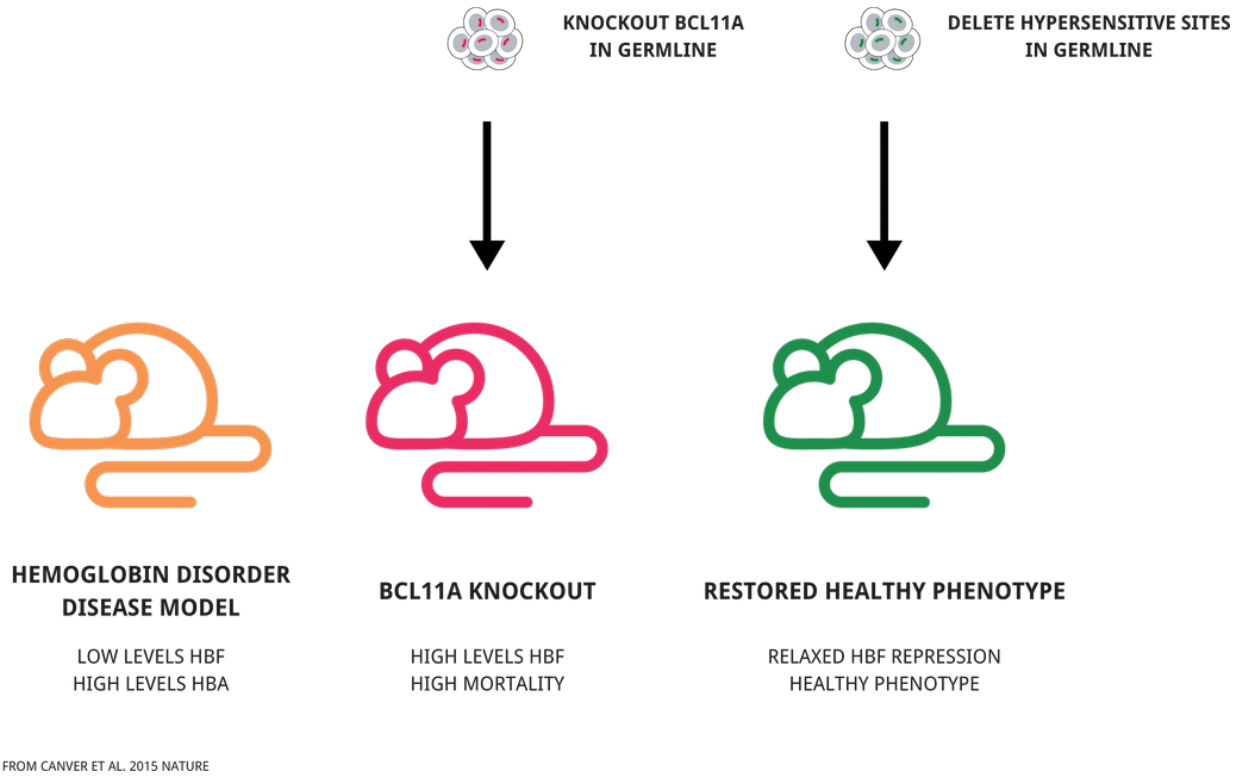


Figure 2. *Knocking out Bc11a led to increased HBF levels, but the mice did not live for more than a few hours. Targeting specific noncoding sites in the Bc11a led to increased HbF levels and did not seem to have the adverse consequences of whole gene knockout.*

Interestingly, targeting the two conserved hypersensitive sites led to two different phenotypes. In one region (m+55), targeting caused a twofold decrease in *Bcl11a* over the control. In the other (m+62), depletion levels approached the effect of total enhancer deletion — but without the caveats of possible inversions that comes with removing large pieces of DNA. This data supported the idea that precise CRISPR targeting within noncoding DNA can modulate phenotypic outcome.

After validating highly correlated sgRNAs in culture, Canver et al. moved to germline mouse cells and generated mice lacking the m+62 hypersensitive site (Figure 2). Previous *in vivo* work knocking out *Bcl11a* in mice produced animals which only lived for a few hours before dying due to neurologic and immunologic toxicity. Yet when the group only deleted m+62 site, they found that these mice developed and bred healthily. This offered significant hope for taking this approach into the clinic.

It has been shown that relieving HbF repression is sufficient to yield a therapeutic effect (Orkin 2016). Canver et al. noted that hemoglobin switching from HbF to HbA was significantly delayed in homozygous m+62 knockout mice as compared with wild type. For heterozygous mice, the delay was still apparent but more moderate. The phenotypic differences between homozygous, heterozygous and wild-type mice therefore validated the idea that dose-dependency — adjusting gene expression on a gradient with precise noncoding modification rather than an on-off binary — may prove a viable therapeutic alternative.

Follow the Transcription Factors

In 2016, Korkmaz et al. studied the noncoding genome by searching for known regulatory motifs to identify regions that may control expression. Specifically, the group interrogated noncoding DNA featuring putative binding sites for well-known transcription factors. They focused on p53 (*TP53*), one of the most famous tumor suppressors, and estrogen receptor alpha (ERα, *ESR1*), a hormone-dependent regulator of proliferation in breast cancer.

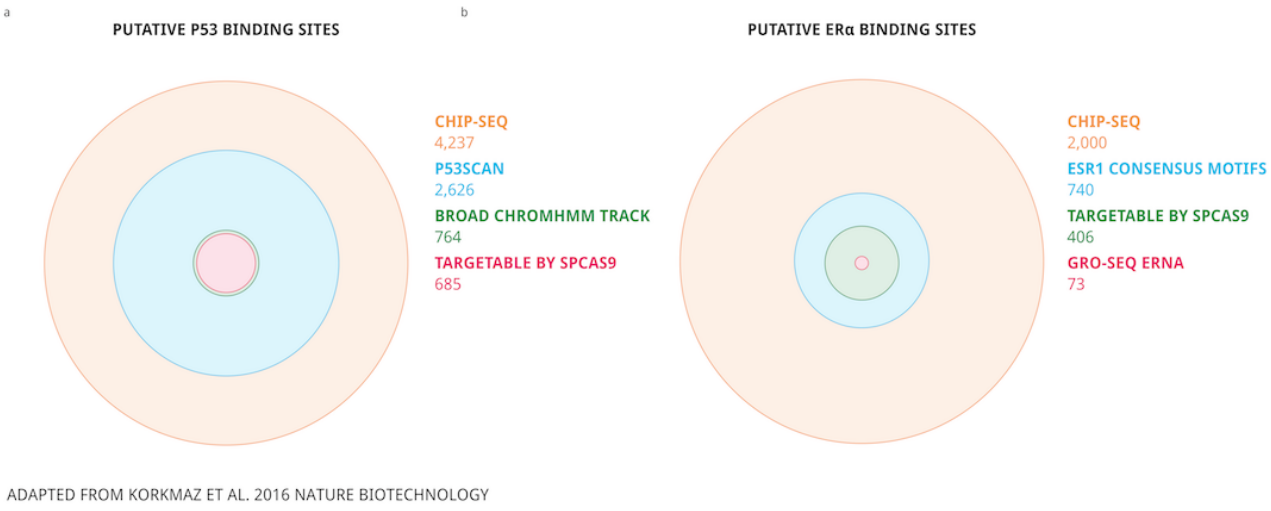


Figure 3. *a) Putative p53 and b) ERα binding regions were filtered based on histone markers, motif analysis and other key factors according to a series of bench data and in silico web tools.*

In order to find regions to tile, the group first looked for p53 binding sites throughout the genome (Fig. 3a). They generated a list of 4,237 putative loci by ChIP-seq. They narrowed these to 2,626 regions using in silico prediction by p53scan binding motif analysis. From there, they further pared the list down to 764 predicted p53-binding enhancers based on histone markers from Broad ChromHMM Track via the UCSC Genome Browser. Within this group, 685 regions were targetable with CRISPR-Cas9 (i.e. contained NGG PAMs). The group finally designed 1,116 sgRNAs to target these loci.

For the *TP53* portion of their study, Korkmaz et al. used an enrichment screen (Figure 4). They targeted p53 binding sites around the genome and then measured whether the cells bypassed oncogene-induced senescence (OIS). They found eight sgRNAs that yielded OIS bypass, then validated these sgRNAs in culture with positive and negative controls

and found that four guides replicated the phenotype. Mapping these sgRNAs to specific loci in the enhancer, just as Canver et al. did, yielded precise identification of functional regulatory elements.

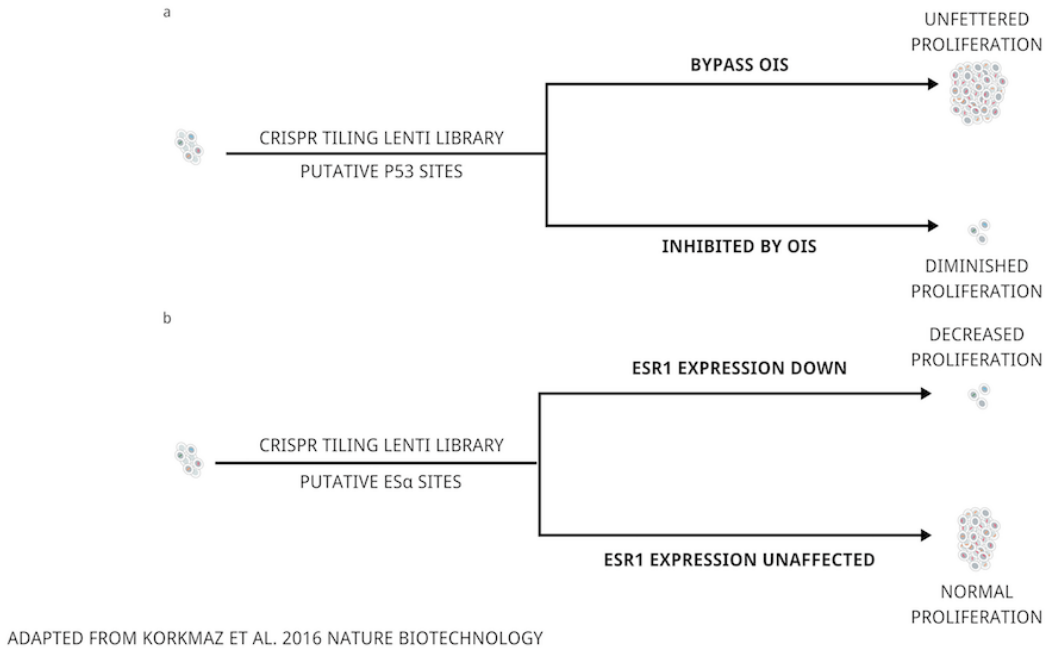


Figure 4. a) After tiling putative p53 binding regions in BJ-RASG12V cells with a lentiviral CRISPR library, some of the population escaped OIS and proliferated. b) Tiling putative ERα sites yielded precise mapping of functional enhancer domains controlling CCND1-mediated proliferation through a dropout screen.

The group then performed a similar investigation with ERα binding sites. They sorted through 2,000 loci from ChIP-Seq data. These were reduced to 740 regions based on *ESR1* consensus motifs (allowing for up to one mismatch), 406 of which could be targeted by SpCas9. This was intersected with GRO-seq enhancer RNA (eRNA) data generating 73 putatively active enhancers to tile. 97 sgRNAs were designed for a lentiviral pool.

In this investigation, the group used a dropout screen that depended on knocking out the cells' ability to proliferate. Based on sgRNA representation, three regions were identified which seemed to play a role in proliferation. One of those, ERα-enh588, was previously identified as a putative enhancer of the cyclin D1 gene, *CCND1*. The sgRNA targeting this enhancer was tested individually with non-targeting sgRNAs as negative controls. ERα-enh588 was confirmed in a biological context for the first time, bolstered by mRNA, protein and eRNA expression data. Finally, the group tested the reliance of *CCND1* on the 17B-estradiol hormone and found that activation required the ERα-enh588 domain.

Varied Approaches Depend on Experimental Need

Canver et al. focused on genes of interest and discovered cis-acting regulatory regions. Korkmaz et al. used putative regulatory elements to find genes of interest, then isolated

essential regions in the corresponding enhancers with CRISPR. Both approaches have helped to further illuminate the purpose of noncoding DNA. By exploring these regions, researchers can then associate variants from disease populations with essential functional domains.

CRISPR tiling presents new opportunities for noncoding interrogation and broader genome editing. Population variants appear in the noncoding genome more often than they do in genes, and investigators are now able to systematically evaluate those regions with CRISPR. In translational research, investigators are no longer restricted to editing genes. They can now use fine-tuned genetic manipulation to effect a dose-dependent response.

Building Effective Noncoding CRISPR Libraries

A few themes have emerged from CRISPR tiling studies. One key trend in Korkmaz et al. and Canver et al., as well as a review by Zhou & Wei in 2016, is the call to use Cas9 orthologs to more completely saturate unknown regions. SpCas9 is limited to NGG PAM sites and therefore cannot cover all regions throughout the genome. This can introduce bias in tiling studies; Korkmaz et al. found that Cas9 restrictions meant they could only explore 90% and 60% of candidate enhancers containing p53 and ESR1 sites, respectively. By including nucleases like NmCas9, Cpf1 and others, investigators can close gaps in targeted regions and evaluate all functional regulatory domains.

Investigating the noncoding genome also makes model characterization more important than ever. While protein coding regions tend not to vary to the same degree, it has been shown that cell line to cell line variation can occur in noncoding DNA (Sanjana et al. 2016). As we've found at Desktop Genetics, differences even in the protein coding genome can affect 1 in 20 guides — a significant number in the context of a large CRISPR library. Based on the research by Sanjana et al. and others, this number is bound to increase when tiling the other 98% of the genome.

Sources

Canver MC, Smith EC, Sher F et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. 2015 Nov 12;527(7577):192-7. doi: 10.1038/nature15521. PubMed PMID: 26375006.

Elgar G, Vavouri T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet. 2008 Jul;24(7):344-52. doi: 10.1016/j.tig.2008.04.005. Review. PubMed PMID: 18514361.

Giani FC, Fiorini C, Wakabayashi A et al. Targeted Application of Human Genetic Variation Can Improve Red Blood Cell Production from Stem Cells. Cell Stem Cell. 2016 Jan 7;18(1):73-8. doi: 10.1016/j.stem.2015.09.015. PubMed PMID: 26607381.

Korkmaz G, Lopes R, Ugalde AP et al. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat Biotechnol. 2016 Feb;34(2):192-8. doi: 10.1038/nbt.3450. PubMed PMID: 26751173.

Kurita R, Suda N, Sudo K et al. Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. PLoS One. 2013;8(3):e59890. doi: 10.1371/journal.pone.0059890. PubMed PMID: 23533656.

Melnikov A, Murugan A, Zhang X et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012 Feb 26;30(3):271-7. doi: 10.1038/nbt.2137. PubMed PMID: 22371084.

Orkin SH. Recent advances in globin research using genome-wide association studies and gene editing. Ann N Y Acad Sci. 2016 Mar;1368(1):5-10. doi: 10.1111/nyas.13001. PubMed PMID: 26866328.

Sanjana NE, Wright J, Zheng K et al. High-resolution interrogation of functional elements in the noncoding genome. Science. 2016 Sep 30;353(6307):1545-1549. PubMed PMID: 27708104.

van der Harst P, Zhang W, Mateo Leach I et al. Seventy-five genetic loci influencing the human red blood cell. Nature. 2012 Dec 20;492(7429):369-75. doi: 10.1038/nature11677. PubMed PMID: 23222517.

Zhou Y, Wei W. Mapping regulatory elements. Nat Biotechnol. 2016 Feb;34(2):151-2. doi: 10.1038/nbt.3477. PubMed PMID: 26849519.

Søren Hough, Leigh Brody, Ayokunmi Ajetunmobi, Bill Morrison and Edward Perello contributed to this Resource.