Martin Kircher and Kerstin U. Ludwig*

Systematic assays and resources for the functional annotation of non-coding variants

https://doi.org/10.1515/medgen-2022-2161

Abstract: Identification of genetic variation in individual genomes is now a routine procedure in human genetic research and diagnostics. For many variants, however, insufficient evidence is available to establish a pathogenic effect, particularly for variants in non-coding regions. Furthermore, the sheer number of candidate variants renders testing in individual assays virtually impossible. While scalable approaches are being developed, the selection of methods and resources and the application of a given framework to a particular disease or trait remain major challenges. This limits the translation of results from both genome-wide association studies and genome sequencing. Here, we discuss computational and experimental approaches available for functional annotation of non-coding variation.

Introduction

Over the past decade, technical capabilities to identify genetic alterations in individual genomes have increased substantially. However, two major challenges remain: (i) discriminating between pathogenic (causal) and benign variants; and (ii) understanding the effects of genetic variants at the molecular level. This is particularly true for the "non-coding" genome, which harbors both the majority of variants associated with common traits (as identified by genome-wide association studies [GWAS]) and an as yet unknown number of variants underlying monogenic disorders [1, 2].

To advance variant interpretation, large-scale collaborative efforts have generated extensive catalogs that document genetic variation and genomic elements, together with their respective molecular functions, across hundreds of cell types, including 3D genomic interactions [3]. Most of this information is deposited in public databases, which have emerged as knowledge hubs in human genomics. In addition, technological advances now provide the means to characterize the molecular effects of genetic variation at scale on an experimental basis. Nonetheless, the identification of appropriate resources and protocols, the assessment of relevant data, and the correct interpretation of experimental findings remains problematic [4, 5]. Here, we summarize how diverse experimental and computational approaches can be applied to advance interpretation of genetic variation in the non-coding genome (Figure 1).

***Corresponding author: Kerstin U. Ludwig,** Institute of Human Genetics, University Hospital Bonn, University of Bonn, Venusberg-Campus 1, Building 76, 53127 Bonn, Germany, e-mail: kerstin.ludwig@uni-bonn.de

Martin Kircher, Institute of Human Genetics, University of Lübeck, Lübeck, Germany; and Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany, e-mail: martin.kircher@uni-luebeck.de

Resources for the non-coding genome

Human genetic variation

Over the past decade, systematic global efforts have been made to catalog diverse types of genetic variants both across the allelic spectrum and across populations [3, 4, 6, 7]. These investigations have included the International HapMap Project, the 1000 Genomes Project, and the more recent Genome Aggregation Database (gnomAD) initiative. According to current estimates, each individual genome harbors 3 to 4 million small alterations of below 50 bp (the majority of which are single nucleotide variants [SNVs]), and around 15,000 structural variants (SVs, >50 bp) [8]. Additional variants will be identified as further progress is made towards completion of the human reference sequence (Telomere-to-Telomere (T2T) consortium; [9]).

Variant information is made accessible through web-based resources, which balance the issues of data sharing and privacy in order to benefit the medical genetics community (Table 1). The majority of variants are derived from observations in only one individual ("singletons"). Since some of these might represent either technical artifacts or disease-causing variants in unscreened individuals from population-based cohorts, researchers are encouraged to rely on cross-population allelic frequency, rather than on the mere presence of a variant. Importantly, variants identified from more than 100,000 individuals (across diverse populations) serve as a good basis to study sequence variation compatible with life [10]. The underrepresentation of

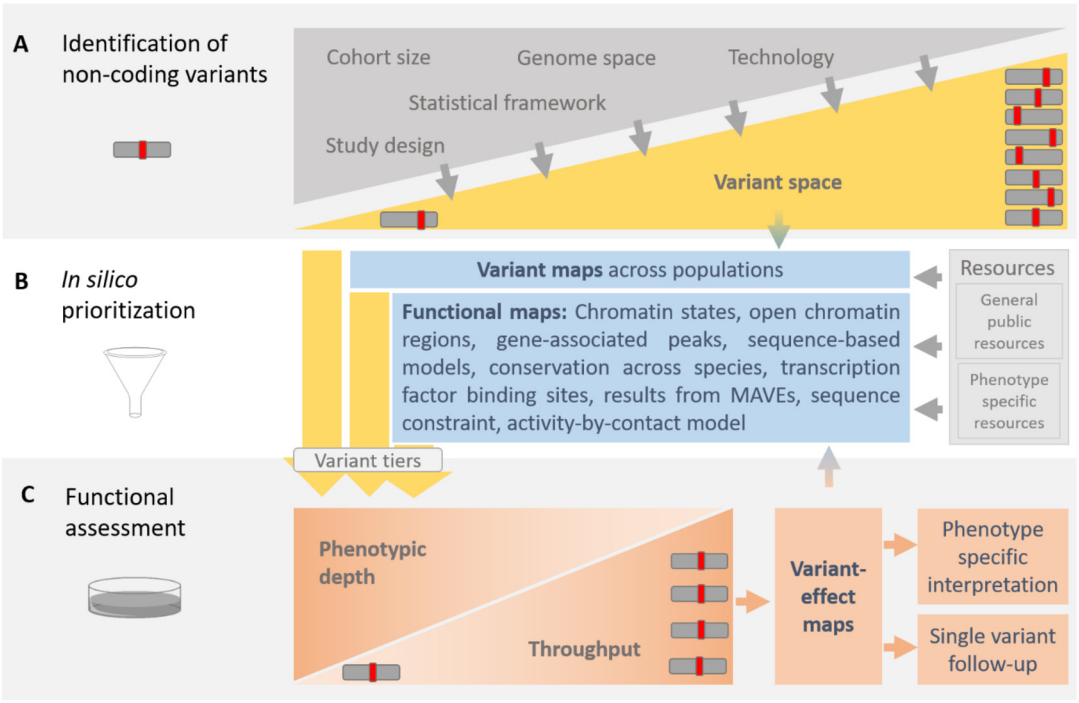

Figure 1: Functional genomics of non-coding variants. (A) Defining the variant space. Technological advances such as array- or sequencing-based methods have enabled the systematic identification of genetic variants in individual genomes. The variant space encompasses all identified candidate variants, for example, all variants observed in an individual patient in a clinical panel or exome, variants identified from a genome-sequenced case-parent trio, or those variants that meet a specific statistical threshold in cohort-based analyses (e. g., significance in genome-wide association studies). Thus, the size of this variant space is largely driven by the specific study design. **(B) Ranking of variants.** In situations where the variant space is small, these can be taken directly to functional assessment, or they can be ranked based on variant frequencies in different populations (variant tiers, yellow arrows). However, in most study designs, prioritization approaches are required. Here, functional maps (of experimental or computational origin) are integrated with the variant space to reduce numbers. Functional maps are drawn from publicly available web resources and databases, which can be general or specific to a certain phenotype. For a selection of resources, see Table 1. **(C) Mapping variant effects.** The experimental approach to investigate the functional effects of the prioritized variants is a trade-off between depth of molecular assessment and throughput (i. e., number of variants tested in parallel). Ideally, results of variant analyses are collected in variant-effect maps, which can then be used for interpretation in the phenotypic context and/or prioritize a limited number of variants for in-depth investigation, for instance in animal or organoid models. Importantly, variant-effect maps may inform prioritization in future studies of the same phenotype, if they are deposited in publicly available resources. Abbreviations: MAVE, multiplex analysis of variant effects.

variant alleles, or the clustering of trait-associated variants within a genomic region, can facilitate the prioritization of variants that are of functional relevance [11–13].

Together with information from NCBI ClinVar or HGMD, "variant prioritization" based on allelic frequencies has been central to the identification of causal genes for many Mendelian syndromes [14]. However, few of the variants that are reported as causal in curated clinical databases are located outside of gene sequences, thus limiting the interpretation of variants identified in non-coding regions by genome sequencing. Furthermore, while public variant maps have facilitated the identification of common risk variants for multifactorial traits, functional interpretation is difficult, due to the fact that, again, most are located outside of protein-coding genes [15].

Genomic architecture of the non-coding genome

By definition, the "non-coding genome" encompasses all of the sequence located outside of protein-coding elements (i. e., around 98 % of the human genome). It contains the majority of variants associated with common traits, as well as an as yet unknown number of causal variants for Mendelian diseases. Although widely used, the term "non-coding genome" does not reflect its true complexity, as illustrated by the wide diversity of molecular functions that have been associated with distinct sequence elements.

Non-coding elements can be located in close proximity to coding regions and are considered part of the respec-

Table 1: Important resources for the annotation of functional variants in the human genome. Various online resources are listed in the categories "Genetic variation and associations," "Large-scale functional genomics data sources," "Browser and meta databases," and "Enhancer, transcription factor, and element databases." We tried to list the major resources, so this list must be incomplete and can only represent a limited view of available resources. Many of these websites provide data access through interactive searches and visualizations, while other sites serve as portals for data download and offline analysis.

Name	Description	URL
Genetic variation and associations		
BRAVO/TOPMed	BRAVO variant browser provides alleles, functional annotations, and allele frequencies from variants identified across genomes in the TOPMed project	https://bravo.sph.umich.edu/
gnomAD	Genome Aggregation Database providing aggregated and harmonized variants (incl. SVs) and their annotation from various large-scale exome and genome sequencing projects	https://gnomad.broadinstitute.org/
ClinVar	NCBI repository clinically annotated variant effects	https://www.ncbi.nlm.nih.gov/clinvar/
COSMIC	Catalogue of Somatic Mutations in Cancer and their annotations	https://cancer.sanger.ac.uk/cosmic/
GWAS Catalog	NHGRI-EBI Catalog of human genome-wide association studies, collecting region/variant associations from thousands of publications	https://www.ebi.ac.uk/gwas/
IGSR	The International Genome Sample Resource, incl. the Human Genome Diversity Project (HGDP) and the Simons Genome Diversity Project (SGDP)	https://www.internationalgenome.org/
1000 Genomes Project	International Genome Sample Resource of the 1000 Genomes Project providing links to individual level data from various populations	https://www.internationalgenome.org/
Large-scale functional genomics data sources		
ENCODE	Data portal of the Encyclopedia of DNA Elements, including for example TF and histone ChIP, open chromatin, and expression data	https://www.encodeproject.org/
EMBL-EBI Single cell atlas	Single-cell expression atlas across species, including the Human Cell Atlas	https://www.ebi.ac.uk/gxa/sc/home
FANTOM	Functional Annotation of the Mammalian Genome, including atlases of promoters, enhancers, long non-coding RNAs, and microRNAs	https://fantom.gsc.riken.jp/
GTEx	Genotype-Tissue Expression (GTEx) Portal with tissue-specific gene expression and regulation data	https://www.gtexportal.org/home/
HuBMAP	Human BioMolecular Atlas resource for discovery, visualization, and download of single-cell tissue data	https://portal.hubmapconsortium.org/
IHEC	Data portal of the International Human Epigenome Consortium incl. methylome, transcriptome, histone, and other data	https://ihec-epigenomes.org/
Roadmap Epigenomics	Integrative analysis of 111 reference human epigenomes	http://www.roadmapepigenomics.org/
pyschENCODE	Integrated resource of regulatory genomic elements in individuals with neuropsychiatric disorders	http://resource.psychencode.org/
4DN	4D Nucleome Network provides nuclear organization data as well as a platform to search, visualize, and download them	https://data.4dnucleome.org/
Browser and meta databases		
Ensembl Regulation	Ensembl Regulation provides computational annotation of regulatory features in the genome, incl. genome segmentation and annotation of regulatory features	http://www.ensembl.org/info/genome/funcgen/index.html
Gene Expression Omnibus (GEO)	NCBI repository for all kinds of functional genomics datasets and their structured metadata	https://www.ncbi.nlm.nih.gov/geo/
UCSC Genome Browser	Popular genome browser integrating data from various sources	https://genome.ucsc.edu/
WashU Epigenome Browser	Genome browser integrating epigenetic, 3D genome visualization, and image data	https://epigenomegateway.wustl.edu/
Enhancer, transcription factor, and element databases		
Altius Index	Human DHS index of about 3.6 million sites, providing a common coordinate system for regulatory DNA	https://index.altius.org/ (browser), https://www.meuleman.org/research/dhsindex/ (data)
ENCODE SCREEN	Registry of Candidate cis-Regulatory Elements from the ENCODE project	https://screen.encodeproject.org/
EnhancerAtlas	Experimentally derived enhancer annotation in nine species	http://www.enhanceratlas.org/
Gene Transcr. Regulation Db	GTRD provides uniformly processed ChIP-seq data for identification of transcription factor binding sites in human or mouse	https://gtrd.biouml.org/
GeneHancer	Genome-wide integration of enhancers and target genes in GeneCards	https://www.genecards.org/

Table 1 (continued)

Name	Description	URL
JASPAR	Open-access database of transcription factor binding profiles	https://jaspar.genereg.net/
MaveDB	Open-source platform to distribute and interpret data from multiplex assays of variant effects (MAVEs)	https://www.mavedb.org
MPRAbase	Repository and uniform processing of massively parallel reporter assay (MPRA) datasets across several organisms	https://www.mprabase.com/
ORegAnno	Open resource for curated regulatory annotation, incl. about transcription factor binding sites, RNA binding sites, regulatory variants, haplotypes, and other regulatory elements	http://www.oreganno.org/
RegulomeDB	Annotation of SNVs with known and predicted regulatory elements in the intergenic regions of the human genome	https://regulomedb.org/
VISTA Enhancer Browser	Resource for experimentally validated human and mouse non-coding fragments with gene enhancer activity as assessed in transgenic mice	https://enhancer.lbl.gov/

tive genes. These elements include the 3′/5′ untranslated regions, the core promoter, and (deep) intronic splice regions, all of which have an established role in gene regulation [16]. In addition, "non-coding genes" provide the sequence for diverse RNA species, which are generally not translated into proteins (e. g., long non-coding RNA, microRNA). While they contribute to transcriptional and post-transcriptional regulation of their target genes, the map of non-coding genes remains incomplete [17].

However, most regulatory sequence elements are located outside of genic regions and are difficult to predict from sequence alone. Regulatory sequence elements are composed of: (i) "proximal regulatory elements," which are located close to transcription start sites; and (ii) "distal regulatory elements," which are located further away. Both are in contact with their target genes through spatial interaction and, based on different resources (Table 1), they cover an estimated 5–20 % of the genome. Their interactions occur predominantly within the context of regulatory units [18], with topologically associating domains (TADs) representing the basic domains of the 3D genome architecture [19]. Importantly, the activity state of a specific regulatory element (e. g., an active "enhancer" and a repressive "silencer") is largely dependent on the presence of cell type-specific binding proteins (e. g., transcription factors), and can vary between cell types.

Functional maps of the "non-coding" genome

In contrast to the technical ease of identifying human variation at the individual and population levels, current capabilities for understanding the functional effects of non-coding variants at the molecular, cellular, organismal, and ultimately phenotypic level remain limited. This is largely attributable to our limited understanding of sequence elements in the non-coding genome, which is mainly due to: (i) the lack of a universal translation code comparable to the amino acid sequence in protein-coding genes; (ii) the temporal and spatial activity of regulatory elements complicating their study; (iii) limited understanding of general gene regulation processes; and (iv) the presumably small, but as yet unknown, effect sizes of most non-coding variants.

To improve understanding of the non-coding genome, functionally relevant genomic elements must first be cataloged and annotated in a systematic manner [5]. Here, "functionally relevant" refers to sequence blocks of variable length that contribute to the spatio-temporal expression of genes, hence the term "regulatory." The Encyclopedia of DNA Elements (ENCODE) represents the first attempt towards the global mapping of sequence elements that are indicative of regulatory activity (i. e., DNA accessibility, histone modifications, methylation patterns), across diverse human tissues and developmental stages. Additional projects, such as the EU BLUEPRINT Epigenome initiative, the International Human Epigenome Consortium (IHEC), NIH RoadMap Epigenomics, and the Functional Annotation of the Mouse/Mammalian Genome (FANTOM) consortium (Table 1), have further advanced the "functional genome map" by providing additional tissues and assay types (e. g., immunoprecipitation of DNA binding proteins [transcription factors and histones], 3D organization, and interaction of DNA elements).

Importantly, these datasets are often enriched for specific tissues and cell types, with a major bias towards those that are easily obtainable (e. g., immune cell types from blood). In addition, funding for specific projects has often been provided within the context of specific research areas (e. g., the focus of the BLUEPRINT Epigenome is the

hematopoietic system). Therefore, the use of public data for prioritization approaches may be hampered by inherent biases in data acquisition, and thus a lack of general applicability.

Functional maps are often accessible through the graphical interfaces of web-based portals (Table 1). In contrast to the situation for genetic variation, here, data privacy is only of limited concern, and both unprocessed and processed data are often made available. Although a large number of molecular assays are performed, each provides only one functional dimension, and integrated approaches are often required to provide functional annotation (e. g., diverse histone modifications combined into chromatin segments [20]).

Visualization of functional maps is most helpful when used for specific chromosomal regions (e. g., a GWAS risk locus) and/or when a disease-relevant cell type is already known (and available). If this is not the case, functional maps can be used to identify the disease-relevant cell types by calculating enrichments of non-coding variants in regulatory elements. To enable such systematic analyses, summary level data on the entirety of regulatory elements across different cells and tissues can be downloaded (e. g., from the SCREEN database [21] or Ensembl Regulatory Build [22]) and used for enrichment approaches. Methods for enrichment analyses have been excellently reviewed elsewhere [23].

Integration of genotypes and (molecular) phenotypes

Individual projects have also analyzed primary tissues and embryonic cell types, as well as cells derived from non-European populations, since the influence of population background on cell type-specific gene regulation remains unclear. While these studies complete the existing functional maps, they often lack the resources required to maintain web portals and thus release their data in general databases such as NCBI Gene Expression Omnibus (GEO). Tens of thousands of functional genomics datasets are available in GEO, including data from diverse model organisms, which enable interspecies analyses as promising orthogonal avenues [24, 25]. The widespread availability of functional maps in public domains such as ENCODE and GEO provides enormous potential for advancing the interpretation of non-coding risk variants. However, one current challenge in this respect is the requirement for the uniform reprocessing of data when multiple studies are combined.

To facilitate the investigation of the impact of common genetic variants on molecular functions, the Genotype-Tissue-Expression (GTEx, [26]) Project was established in order to generate maps of quantitative trait loci (QTLs), in which genotypes are statistically correlated with molecular measures at the population level. Following their application to investigate gene expression in bulk (i. e., expression QTLs [eQTLs]), QTL studies have since been extended towards the investigation of a variety of molecular functions, such as splicing (sQTLs), methylation (meQTL), chromatin accessibility (caQTL), and even the regulation of protein abundance (pQTL). Importantly, the observed correspondence between genotype and molecular measure represents only a statistical correlation between two traits, and orthogonal evidence is required to delineate the biological mechanism. The latter might include Bayesian approaches, such as colocalization, or experimental manipulation [23].

Beyond single nucleotide variants

SNVs are the most abundant form of genetic variation and are the most comprehensively cataloged to date due to the technical ease of their identification. Unsurprisingly, therefore, most annotation efforts for non-coding variants focus on SNVs. However, variant types that encompass a larger number of nucleotides, such as SVs, are probably more powerful in terms of causing functional effects of regulatory elements. For instance, whereas the absence of an entire transcription factor binding motif is more likely to abolish binding, an SNV within the motif is likely to modify binding affinity in a quantitative manner. Studies of patient-specific SVs have already suggested disease-causing mechanisms in non-coding regions, particularly within the context of TADs [19, 27]. Since the identification of SVs is becoming ever easier due to technological and algorithmic advances (e. g., long-read sequencing, optical mapping) [8], an increasing number of high-quality SVs is anticipated. While these provide the opportunity to study regulatory effects, interpretation might also become more complex, as SVs are likely to harbor multiple genomic elements simultaneously, each of which might contribute to a different molecular function [28, 29].

Prioritization approaches

Intuitively, the number of candidate non-coding variants identified by large-scale GWAS might be larger than those

identified by studies on *de novo* mutations (DNMs) in rare diseases. However, the exact number of candidate variants (i. e., the "variant space") is largely influenced by study design, sample size, and the genetic architecture of the trait (Figure 1). For example, an analysis of genome-wide data from 1,000 trios would result in around 80,000 candidate DNMs, whereas GWAS for traits with limited biological complexity (e. g., orofacial clefting) have identified a few dozen risk loci with a few hundred candidate variants to date. The aim of *in silico* prioritization approaches is to provide a relative ranking of candidate variants, thereby allowing a reduction in the number of variants forwarded for experimental follow-up.

Non-coding *in silico* scores

Two major types of *in silico* scores are currently available, i. e., "specialized scores" and "broadly applicable scores." Specialized scores assess the impact of a variant on specific molecular functions and are particularly powerful for candidate variants with *a priori* functional hypotheses, e. g., those located in splice regions, within binding sites of transcription factors or miRNAs, or in regions of open chromatin. In contrast, "broadly applicable scores" make use of general annotations such as sequence constraint, i. e., conservation across species, or metrics derived from variant density at a certain region. Sequence constraint is a particularly powerful measure, as it integrates diverse molecular effects through organismal fitness and survival, at the cost of not necessarily providing a base pair resolution. This concept is commonly used in the context of annotating deleterious variants in protein-coding regions (e. g., missense Z-score in gnomAD), and can be readily transferred to non-coding regions. Nonetheless, at writing, no single score from either group is an effective predictor across all variant types. To improve predictions, multiple scores can be integrated into one measure of deleteriousness [30]. Examples of tools that use genomically broad (and less biased) datasets include Eigen, LINSIGHT, and CADD.

Again, the generalizability of *in silico* scores is limited by the available data. Particularly in the case of scores that are derived from specific experimental assays via machine learning, the predictive power is limited to those functions that are represented in the training data. For instance, if no feature covers the effect of a cell type-specific transcription factor, the resulting score will not be predictive of such molecular effects when annotating candidate variants. Furthermore, *in silico* scores that rely on conservation may fail in the prediction of "gain-of-function" variants, e. g., the generation of new transcription factor binding sites. It was previously demonstrated that results from experimental data of regulatory variant effects (e. g., from multiplex assays of variant effects [MAVEs], see below) are not well captured by any of the existing *in silico* scores [31].

To overcome these limitations, novel computational methods must be developed [31, 32]. Currently, the best results are obtained from sequence models that learn active and inactive motif representations from large collections of open chromatin data and histone marks (e. g., DeepBind [33], gkmSVM [34], DeepSEA [35], and Enformer [36]). These sequence models are publicly available and can be applied to variants of interest, although they may be biased by the representation of cell and tissue types in the respective training data. Model specialization ("transfer learning") on matching cell type data might reduce such effects, and this is an area of active development within the computational field.

Layered prioritization approaches

A wide range of computational pipelines are available, thus creating a plethora of prioritization options for any given list of variants. Therefore, consecutive ("layered") approaches have become popular. These include: (i) considering variants with a specific predicted molecular effect only; (ii) removing variants above a certain allele frequency; (iii) applying the requirement for colocalization with certain histone modifications or open chromatin annotation; and (iv) filtering for conservation. Importantly, each of these layers reinforces the applied assumptions (e. g., inverse correlation of allele frequency and effect size), despite our still incomplete biological understanding of, and substantial evidence for exceptions to, all these proposed rules. To enable a more systematic characterization, the effects of the individual prioritization layers must be investigated one criterion at a time with the inclusion of a random set of all variants, or using a fully crossed design. Ideally, a compendium of scalable assays, each addressing a certain aspect of molecular read-out, should also be available. MAVEs (see below) and advances in synthetic biology represent initial steps in this direction.

Approaches to link non-coding elements to target genes

Regulatory sequence elements can be located upstream or downstream – or even on a different chromosome [37] – than its target gene(s). This renders the assignment of

links between genes and regulatory elements (gene-to-regulatory element link [GRL]) difficult, which is often required for the performance of gene set analyses and the generation of biological hypotheses. The most commonly used approaches to assign GRLs are on the basis of proximity, considering either the closest gene or both neighboring genes (potentially within certain distance limits). Alternatively, all genes within certain genomic windows, or all genes within TADs, are considered target genes. GRLs may also be inferred from experimental data, e. g., from diverse chromatin capture datasets; coexpression/coactivity data obtained from matched open chromatin and expression data across multiple cell-types; or the Activity-By-Contact (ABC) model, which considers both chromatin interaction and accessibility/histone acetylation data [38]. Once GRLs are established, a wide variety of gene annotation-based methods can be applied, ranging from gene set analyses to pathway and network enrichments.

Multiplex assays of variant effects

The necessity for experimental testing of thousands of variants, coupled with advances in next-generation sequencing (NGS), has driven the development of MAVEs [39] and the collection of their results in an open repository, MAVEdb [40]. At their core, MAVEs allow systematic screening of variants in a single quantitative experiment and are primarily intended to identify variants with the potential for a specific molecular effect. We describe two major types of MAVEs for non-coding regions below. Importantly, the results of MAVEs typically require subsequent validation in an organismal system, often including organoid or animal models [41].

Massively parallel reporter assays

Massively parallel reporter assays (MPRAs, alternatively CRE-seq or STARR-seq) test the capability of putative regulatory elements to trigger gene expression in a specific cellular context. In MPRAs, thousands of short sequences (typically 150–300 bp) containing the regions (or variants) of interest are first created by oligo synthesis, which is the method of choice for the assessment of many independent variants, e. g., those located across loci. Alternatively, existing variation in cell lines can be utilized [42, 43], or error-prone PCR ("saturation mutagenesis") can create all possible SNVs within a region of interest. The oligo-pool containing all candidate sequences is cloned into plasmid vectors, which are subsequently introduced into an *in vitro* system (Figure 2). Here, the plasmids either remain episomal, or are integrated into genomes through a lentiviral or other system [44]. The latter has been proven to be more powerful for cell types that are difficult to infect (e. g., neurons) or when a nucleosome context is required for read-out. To minimize the large positional effects of random integration [45], flanking insulators to the vector can be included [46].

Regulatory effects are assessed (or "read out") by either quantification of the relative abundance of individual reporter RNAs by NGS compared to their DNA abundance, or a molecular phenotype (e. g., cell proliferation/death, fluorescence of the reporter), and can be correlated with specific variants within the tested element. To date, large MPRA datasets of regulatory variant effects ("variant-effect maps") have been created in specific cell types, for specific loci [31], and for common variants identified by QTL studies or GWAS [42, 43, 47]. In a recent study, MPRAs were applied in clinical research to implicate differing transcriptional networks in two phenotypically similar neurodegenerative disorders [48].

CRISPR/Cas9 approaches

The introduction of CRISPR/Cas9 paved the way for the development of in-genome MAVEs that retain the original local genomic context. In the first study to apply CRISPR/Cas9 genome editing in the context of MAVEs, Findlay et al. generated all possible SNVs in exon 18 of *BRCA1* [49], and used effects on nonsense-mediated decay, exonic splicing, and cellular growth as read-out. This application to coding regions influenced the way in which the non-coding genome is investigated by CRISPR/Cas9 genome editing [50], and recent developments include the integration of single-cell technologies for functional read-out [51]. Based on its capabilities to create highly specific sequence alterations, future applications of the CRISPR prime editing system are anticipated to replace other systems for sequence perturbations [52]. In combination with multiplexed read-outs, this may ultimately replace current vector-based MPRA studies.

CRISPR-based alternatives to genome editing include different CRISPR activation (CRISPRa) and interference (CRISPRi) screens [53]. Here, a specific locus or allele is targeted by a modified CRISPR fusion protein which no longer introduces strand breaks (e. g., dCas9 fusions), but instead serves as a sequence-specific probe. This probe tows an epigenetic modifier – which either increases or impairs gene expression – to a region of interest. With (single-cell) RNA-seq as the molecular read-out and combinatorial

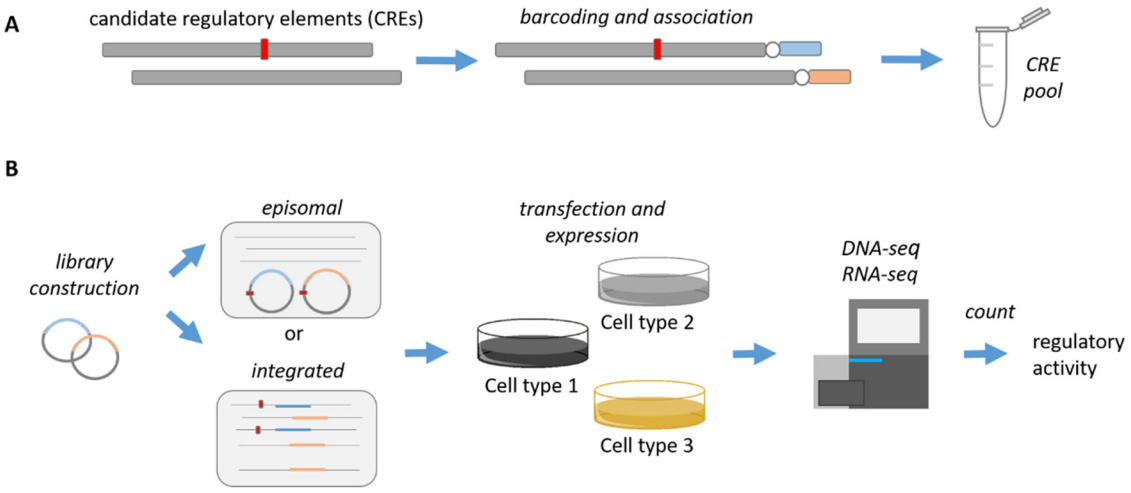

Figure 2: Principle of massively parallel reporter assays (MPRAs). MPRAs are used to simultaneously test hundreds to thousands of variants for potential regulatory effects in one assay. **(A) Generation of MPRA pools.** First, the genomic sequence around each variant (candidate regulatory elements [CREs]) is synthesized or otherwise derived and then combined with an individual barcode. All barcoded sequences are combined into one pool. **(B) MPRA reporter assay.** This pool is then cloned into vectors. The vectors contain a reporter gene (potentially driven by a minimal promoter) and place the CRE upstream of the transcriptional start site, while the barcode becomes part of the transcript's 3′ or 5′ untranslated region. Depending on the assay type, the vectors are designed to either remain episomal or integrate into the genome (e. g., by lentivirus). The vector pool is then transfected (or transduced) into cell types of interest where the reporter genes are expressed. Following extraction of DNA and RNA from those cells, barcodes can be converted into highly complex sequencing libraries and read out on a high-throughput sequencing device. A regulatory effect of a certain CRE can be inferred from the number of detected barcode sequences at the RNA level, corrected by the number of transfected plasmids (detected by the barcode abundance in DNA). Allelic effects are derived from comparing the inferred expression effect of CREs with and without the allele of interest.

CRISPR targeting, the functional impact of candidate loci or allelic variants can be explored within regulatory networks [54].

Limitations of MAVEs

MAVEs share the same limitations as low-throughput functional assays. First, they are performed in individual cellular systems, which require *a priori* knowledge regarding the most appropriate cell type for the trait of interest. Second, the results and interpretation are specific to the applied cell type [55–57], and do not capture organismal effects that might originate from the interaction of various cell types. Third, even if the relevant cell type(s) are known, the respective cell models might be unavailable and can only be replaced in part, e. g., by immortalized cell lines, since the applied alternatives do not capture the true biological identity in its entirety. Performing MAVEs that test effects across a number of cell types and/or conditions might generate the most robust results.

To maximize the biological insights provided by MAVEs, certain technical aspects also require further improvement. First, complementary high-resolution readouts at the molecular and cellular levels are required to measure phenotypes at scale. This is particularly relevant for the phenotypic effects of the more common alleles, which are likely to be subtle for broader phenotypes, but will become detectable with more precise molecular readouts. Second, current technical restrictions in DNA synthesis technology limit DNA fragment sizes, while for longer fragments, the capacity of the plasmid vectors imposes an artificial size limit. Here, novel approaches in synthetic biology that enable the analysis of larger fragments would provide superior coverage of the broad size range of regulatory elements, including the assessment of 3D interactions.

From GWAS to molecular mechanism: The *FTO* locus in obesity

Early GWAS identified an extended haplotype block of 89 common variants, located in introns 1 and 2 of a gene named *Fat Mass And Obesity Associated* (*FTO*), as a risk locus for obesity (as measured by high body mass index [BMI]). *FTO* encodes a protein involved in the oxidative

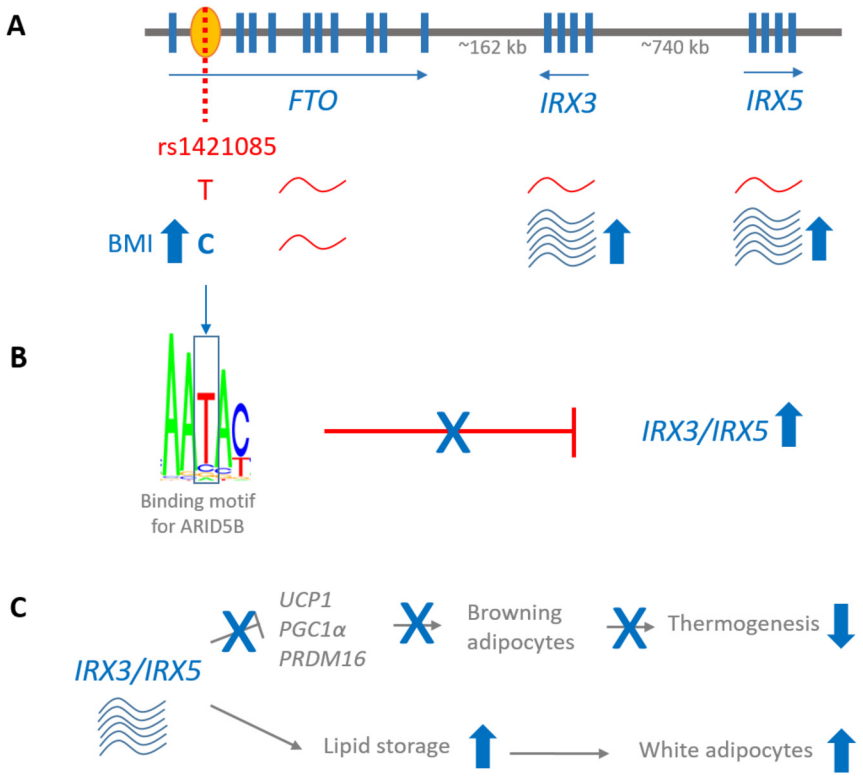

Figure 3: Functional dissection of the *FTO* locus and its role in determining body mass index (BMI). (A) Genetic locus. Genome-wide association studies have revealed an association between a region on chromosome 16q12 and BMI. Three genes are located within a 1-Mb genomic interval. One of the associated non-coding SNVs, rs1421085 (red line), is located within a regulatory sequence (yellow oval). Correlating individual genotypes and expression levels in adipocytes revealed that the C-allele is associated with increased expression of the genes *IRX3* and *IRX5*, while no genotype-dependent effect on *FTO* expression was observed. **(B) Molecular pathway leading to *IRX3*/*IRX5* overexpression.** The C-allele disrupts the binding motif of the transcription factor ARID5B, which is a repressor of *IRX3*/5 expression. In the presence of the C-allele, ARID5B does not bind, and the expression levels of *IRX3* and *IRX5* increase. This pathway explains the observed correlation between the C-allele and *IRX3*/5 expression as depicted in panel (A). **(C) Cellular mechanism.** In pre-adipocytes, IRX3 and IRX5 influence the cellular fate of pre-adipocytes; these can either develop into brown adipocytes, which contribute to energy consumption through thermogenesis, or white adipocytes, in which energy is stored through lipid accumulation. In the presence of an increased amount of *IRX3*/5, there is an increased number of cells shifting to the white adipocyte trajectory, while few brown adipocytes are generated. This results in reduced energy consumption in the presence of increased lipid storage.

demethylation of various RNA species, and thereby contributes to post-transcriptional gene regulation. The absence of any deleterious coding variants suggested a regulatory effect of the risk haplotype, and *FTO* was considered the major positional candidate gene based on functional evidence from mice [58].

In a seminal study, the causative molecular pathway was identified via functional genomics ([59]; Figure 3). First, the authors intersected the regional association statistics with chromatin state annotations using 127 samples from the Roadmap Epigenomics project (see Table 1). A putative enhancer region of 12.8 kb in size was identified and found to be active in adipocytes. The integration of expression data suggested a key role for pre-adipocytes. These represent a specific adipocyte type, which develops along one of two trajectories into either (i) white adipocytes, i. e., fat-storing cells, or (ii) beige/brown adipocytes, which contribute to fat consumption via heat generation ("thermogenesis"). Integration of eQTL and chromatin interaction data from adipocytes suggested the genes *Iroquois Homeobox 3* (*IRX3*) and *IRX5* – both of which are master regulators of thermogenesis – as the enhancer's targets. Having established both the implicated cell type and the regulatory targets, the causal variant was then determined via analysis of sequence conservation, transcription factor binding motifs, and gene coexpression. The analyses highlighted one single SNV (rs1421085) from the long haplotype block, involving a T-to-C transition within the region encoding the binding motif of the transcriptional repressor ARID5B.

Ultimately, the authors presented robust evidence that in the presence of the C-allele of rs1421085, ARID5B cannot bind to its target motif within the enhancer region. This results in increased expression of *IRX3*/*IRX5* in preadipocytes, which are then prompted to shift their developmental trajectory towards white adipocytes. This results in increased lipid storage and a simultaneous reduction in fat consumption via thermogenesis in beige adipocytes. This study reinforces the earlier notion that *a priori* reliance on the "nearest" gene might misguide functional follow-up.

Concluding remarks

The interpretation of variants in the non-coding genome is a key challenge in the path towards personalized medicine. While researchers now appreciate the diversity of the molecular functions of non-coding elements, our knowledge of the full extent of regulatory principles and their complex interactions remains incomplete. In addition, currently available data are restricted to specific cell types (developmental stages, cellular conditions) and molecular assays, which limits efforts to predict variant effects.

In situations in which genomic datasets from different labs must be combined, joint analyses are complicated by cross-study differences in experimental and computational pipelines. In this respect, the importance of large-scale consortia such as ENCODE or gnomAD should be emphasized. These provide standards for experimental protocols, reagents, and terminology, and have pioneered data accessibility via the initiation of data portals with versioned and uniform data processing pipelines, genome browsers, and/or application programming interfaces (APIs), which enable scripted access and data download. However, the maintenance and sustainability of these databases is problematic due to the time-limited and project-specific nature of the respective funding periods.

Over the next decade, a major aim of research will be the efficient engineering of genomic alterations in order to assess their functional read-out in biological systems. Due to the technical challenges associated with MAVEs, the routine performance of these high-throughput approaches across multiple laboratories is unlikely. A more plausible scenario is that specialized academic centers will perform these analyses for a particular region of the genome or a disease of interest, and that the generated data will then be made available to the wider research community. However, to perform these experiments at scale and to enable sustained data accessibility, substantial funding will be required. An initial effort towards this goal is the recent foundation of the Impact of Genomic Variation on Function (IGVF) Consortium [60]. This was established in order to evaluate the function and phenotypic outcomes of coding and non-coding genomic variation using currently available approaches, and to develop improved experimental and computational strategies.

The coming years will see an enormous expansion in functional genomics datasets at all levels, i. e., with respect to novel experimental read-outs, additional annotations, and a variety of computational tools including scores, analysis pipelines, and machine learning approaches. While this opens up substantial opportunities for the field, the enormous challenges associated with data aggregation will complicate the use of these resources by the research community. A specific aim of consortia such as IGVF is to also facilitate data access by establishing variant-to-effect catalogs, including options to visualize variant impacts within the context of the underlying data, tools, and models. Together with additional key players, such as the Global Alliance for Genomics and Health (GA4GH) and European infrastructure projects such as ELIXIR, joined efforts must be established to build global resources for the interpretation of the non-coding genome. This will be required to introduce precision medicine across the broad medical genetics community.

Acknowledgment: We thank current and previous members of the Kircher and Ludwig labs for helpful discussions and suggestions, and the reviewers for their comments and feedback.

Research funding: M. K. is supported by the NIH/NHGRI IGVF effort (1UM1HG011966-01). K. U. L. is supported by the German Research Council (Deutsche Forschungsgemeinschaft [DFG]; LU-1944/3-1).

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: Authors state no conflict of interest.

Informed consent: Not applicable.

Ethical approval: Not applicable.

References

[1] Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24:R102–10.

[2] Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, Campbell C et al. Recommendations for clinical interpretation

of variants found in non-coding regions of the genome. Genome Med. 2022;14:73.
[3] Spielmann M, Kircher M. Computational and experimental methods for classifying variants of unknown clinical significance. Cold Spring Harb Mol Case Stud. 2022;8:a006196.
[4] Krude H, Mundlos S, Øien NC, Opitz R, Schuelke M. What can go wrong in the non-coding genome and how to interpret whole genome sequencing data. Med Genet. 2021;33:121–31.
[5] Garda S, Schwarz JM, Schuelke M, Leser U, Seelow D. Public data sources for regulatory genomic features. Med Genet. 2021;33:167–77.
[6] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
[7] Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC et al. A structural variation reference for medical and population genetics. Nature. 2020;581:444–51.
[8] Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372:eabf7117.
[9] Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A et al. The complete sequence of a human genome. Science. 2022;376:44–53.
[10] Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
[11] di Iulio J, Bartha I, Wong EHM, Yu H-C, Lavrenko V, Yang D et al. The human noncoding genome defined by genetic diversity. Nat Genet. 2018;50:333–7.
[12] Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019;51:88.
[13] Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 2019;363:eaau1043.
[14] Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: Fast and furious with no end in sight. Am J Hum Genet. 2019;105:448–55.
[15] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies. targeted arrays and summary statistics 2019 Nucleic Acids Res. 2019;47:D1005–12.
[16] Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell Mol Life Sci CMLS. 2012;69:3613–34.
[17] Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet. 2014;15:7–21.
[18] Hafner A, Boettiger A. The spatial organization of transcriptional control. Nat Rev Genet. 2022;1–16.
[19] Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. Nat Rev Genet. 2018;19:453–67.
[20] Vu H, Ernst J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. Genome Biol. 2022;23:9.
[21] Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.
[22] Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The Ensembl Regulatory Build Genome Biol. 2015;16:56.
[23] Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet. 2020;11:424.
[24] Wangler MF, Yamamoto S, Chao H-T, Posey JE, Westerfield M, Postlethwait J et al. Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research. Genetics. 2017;207:9–27.
[25] Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2020;48:D704–15.
[26] GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
[27] Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.
[28] Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 2022;185:3426–3440.e19.
[29] Gong T, Jaratlerdsiri W, Jiang J, Willet C, Chew T, Patrick SM et al. Genome-wide interrogation of structural variation reveals novel African-specific prostate cancer oncogenic drivers. Genome Med. 2022;14:100.
[30] Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13:31.
[31] Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun. 2019;10:1–15.
[32] Shigaki D, Adato O, Adhikari AN, Dong S, Hawkins-Hooker A, Inoue F et al. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. Hum Mutat. 2019;40:1280–91.
[33] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.
[34] Beer MA. Predicting enhancer activity and variant impact using gkm-SVM. Hum Mutat. 2017;38:1251–8.
[35] Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022;54:940–9.
[36] Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
[37] Delaneau O, Zazhytska M, Borel C, Giannuzzi G, Rey G, Howald C, et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. Science. 2019;364:eaat8266.
[38] Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT,

Subramanian V et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.
[39] Findlay GM. Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. Hum Mol Genet. 2021;30:R187–97.
[40] Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. 2019;20:223.
[41] Kvon EZ, Zhu Y, Kelman G, Novak CS, Plajzer-Frick I, Kato M, et al. Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. Cell. 2020;180:1262–1271.e15.
[42] van Arensbergen J, Pagie L, FitzPatrick VD, de HM, Baltissen MP, Comoglio F et al. High-throughput identification of human SNPs affecting regulatory element activity. Nat Genet. 2019;51:1160–9.
[43] Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG et al. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res. 2015;25:1206–14.
[44] Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat Methods. 2020;1–9.
[45] Akhtar W, Pindyurin AV, de Jong J, Pagie L, Ten Hoeve J, Berns A et al. Using TRIP for genome-wide position effect analysis in cultured cells. Nat Protoc. 2014;9:1255–81.
[46] Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. 2017;27:38–52.
[47] Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK et al. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016;165:1519–29.
[48] Cooper YA, Teyssier N, Dräger MN, Guo Q Q, Davis JE, Sattler SM, et al. Functional regulatory variants implicate distinct transcriptional networks in dementia. Science. 2022;377:eabi8654.
[49] Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature. 2014;513:120–3.
[50] Przybyla L, Gilbert LA. A new era in functional genomics screens. Nat Rev Genet. 2021. https://doi.org/10.1038/s41576-021-00409-w.
[51] Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell. 2019;176:377–390.e19.
[52] Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature. 2019;576:149–57.
[53] Dominguez AA, Lim WA, Qi LS. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nat Rev Mol Cell Biol. 2016;17:5–15.
[54] Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167:1853–1866.e17.
[55] Maricque BB, Dougherty JD, Cohen BA. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. Nucleic Acids Res. 2017;45:e16.
[56] Inoue F, Kreimer A, Ashuach T, Ahituv N, Yosef N. Identification and massively parallel characterization of regulatory elements driving neural induction. Cell Stem Cell. 2019;25:713–727.e10.
[57] Griesemer D, Xue JR, Reilly SK, Ulirsch JC, Kukreja K, Davis JR, et al. Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. Cell. 2021;184:5247–5260.e19.
[58] Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, JC B et al. Inactivation of the Fto gene protects from obesity. Nature. 2009;458:894–8.
[59] Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C et al. FTO obesity variant circuitry and adipocyte browning in humans. N Engl J Med. 2015;373:895–907.
[60] National Human Genome Research Institute (NHGRI). Impact of Genomic Variation on Function (IGVF). Consortium Genome gov. 2021. https://www.genome.gov/Funded-Programs-Projects/Impact-of-Genomic-Variation-on-Function-Consortium. Accessed 7 Jan 2022.

Martin Kircher
Institute of Human Genetics, University of Lübeck, Lübeck, Germany
Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
martin.kircher@uni-luebeck.de

Kerstin U. Ludwig
Institute of Human Genetics, University Hospital Bonn, University of Bonn, Venusberg-Campus 1, Building 76, 53127 Bonn, Germany
kerstin.ludwig@uni-bonn.de