Molecular Cell
Review

New Insights from Existing Sequence Data: Generating Breakthroughs without a Pipette

Alex M. Plocik[1] and Brenton R. Graveley[1,*]
[1]Department of Genetics and Developmental Biology, Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT 06030, USA
*Correspondence: graveley@neuron.uchc.edu
http://dx.doi.org/10.1016/j.molcel.2013.01.031

With the rapidly declining cost of data generation and the accumulation of massive data sets, molecular biology is entering an era in which incisive analysis of existing data will play an increasingly prominent role in the discovery of new biological phenomena and the elucidation of molecular mechanisms. Here, we discuss resources of publicly available sequencing data most useful for interrogating the mechanisms of gene expression. Existing next-generation sequence data sets, however, come with significant challenges in the form of technical and bioinformatic artifacts, which we discuss in detail. We also recount several breakthroughs made largely through the analysis of existing data, primarily in the RNA field.

Introduction

Recent technological breakthroughs in DNA sequencing have vastly accelerated the rate and greatly reduced the cost of generating high-throughput molecular data. The cost of nucleotide sequencing, for example, is falling faster than even Moore's law for integrated circuits (http://www.genome.gov/sequencingcosts/). Given sufficient download bandwidth and storage capacity and a desktop computer, anyone with the appropriate tools can analyze these existing molecular data sets to address myriad questions in molecular biology. Furthermore, increasing accessibility to supercomputers and cloud computing allows sophisticated analyses to be performed by an ever greater number of scientists.

Because individual laboratories or consortia producing and analyzing large-scale data sets do not (and typically cannot) explore every possible hypothesis that is supported by their data, opportunities abound to test new ideas that may have not been considered previously. In terms of individual data sets, the opportunities are vast. The NCBI Gene Expression Omnibus (GEO) has archived more than 32,000 microarray and sequencing studies that comprise more than 800,000 samples since 2001 (Barrett et al., 2013; http://www.ncbi.nlm.nih.gov/geo/). The Sequence Read Archive (SRA), which maintains sequence data that are either submitted directly to the SRA or extracted from GEO submissions, currently hosts over 1 petabase (Kodama et al., 2011; http://www.ncbi.nlm.nih.gov/sra) and is one of the largest data sets hosted by Google (http://www.dnanexus.com/). Clearly, access to molecular data has never been greater.

Before diving head first into this immense sea of data, it is essential to first identify publicly available data sets that constitute the equivalent of a properly controlled experiment. For example, can data sets be identified that are derived from matched biological samples? In some cases, answers to these questions can be easily obtained from metadata associated with each data set at the GEO and SRA databases. In other cases, however, this information may be difficult to find, incomplete, or even incorrect. Furthermore, expert technical knowledge of the experimental procedures used to generate the data sets is required to assess potential technical artifacts and other caveats. Below, we identify particularly useful collections of publicly available data and discuss common technical artifacts that should be taken into consideration when analyzing next-generation sequencing (NGS) data sets. To demonstrate the utility of analyzing existing data, we highlight successful approaches that have generated new ideas regarding the mechanisms that regulate gene expression.

Identifying Useful Data Sets

Related data set collections that are published together as a resource provide one solution to the problem of identifying comparable data sets. For example, Keji Zhao's group at the NIH has generated one of the most comprehensive ChIP-seq studies of epigenomic information from a single human cell type: resting CD4+ T cells (Barski et al., 2007; Schones et al., 2008; Wang et al., 2008b). These particular data sets have been analyzed by many other investigators to identify specific chromatin marks that combine to constitute "chromatin states" (Ernst and Kellis, 2010; Hon et al., 2009), domains (Shu et al., 2011), and boundaries (Wang et al., 2012), or those marks which are best correlated with tissue-specific gene expression (Pekowska et al., 2010; Visel et al., 2009) or gene architecture (Andersson et al., 2009; Hon et al., 2009; Huff et al., 2010; Schwartz et al., 2009; Spies et al., 2009; Tilgner et al., 2009). The studies demonstrate the range of information that can be gleaned from a single large, coherent collection of data sets. One reason that these studies were so successful was because all of the data sets were generated by a single group, using a consistent method, from a single cell type. Identifying similarly coherent data sets among the vast sea of the GEO, SRA, and other data repositories, however, can be a challenge.

Initiatives such as the 1000 Genomes (Clarke et al., 2012; The 1000 Genomes Project Consortium, 2010), The Cancer Genome Atlas, ENCODE (ENCODE Project Consortium et al., 2012), modENCODE (modENCODE Project Consortium et al.,

Table 1. Summary of Selected Data Sets Available from Ongoing Genome Projects

Project	Organism	Data Type	Assay Type	Cell Lines/Strains/ Individuals	Treatments	Highly Represented Samples
1000 Genomes	**Human** 1,000 individuals	Genomic	WGS	>179		
			WXS	>1,000		
			SNP genotype (up to 2 methods)	>1,000		
			WGS of parent–child trios	2		
ENCODE[a]	**Human** 266 primary and tissue cultures	Genomic	SNP Genotype	62		
		Transcriptome	CAGE	36		K562, GM12878, HeLa, HepG2, H1, A549, Huvec, MCF7
			Exon-array	123		
			RNA-Seq	51		
		Epigenomic	Chromatin conformation (up to 2 methods)	13		K562, GM12878, HepG2, HeLa, H1, A549
			Chromatin marks and transcription factors (up to 201 antibody targets)	119		
			DNA accessibility (up to 3 methods)	181		
			DNA methylation (up to 2 methods)	91		
			Nucleosomes	2		
	Mouse 37 primary tissues and cultures	Transcriptome	RNA-Seq	54		MEL
		Epigenomic	Chromatin marks and transcription factors (up to 56 antibody targets)	28		MEL, CH12, C2C12
			DNA accessibility	44		
modENCODE	**Fruit fly** Developmental stages, cell culture lines, and adult tissues	Genomic	Genotype and CNVs	19 cell lines		
		Transcriptome	RACE & CAGE	1		
			small RNA-Seq	18 stages		
			RNA-Seq	30 stages/29 tissues/22 cell lines	26 compounds / 59 RNAi depletions	S2
		Epigenomic	Nucleosomes (up to 5 methods)	5 cell lines/4 stages		
			Chromatin marks and transcription factors (up to 102 antibodies)	21 cell lines	4 cell lines />5 stages	
	Worm Strains and developmental stages	Transcriptome	RACE	5 strains		
			RNA-Seq	9 strains/>18 embryonic stages		N2
			Gene expression microarrays	12 stages/34 strains		
		Epigenomic	Nucleosomes (up to 2 methods)	4 strains		
			Chromatin marks and Transcription factors (up to 95 antibodies)			

(*Continued on next page*)

Table 1. *Continued*

Project	Organism	Data Type	Assay Type	Cell Lines/Strains/ Individuals	Treatments	Highly Represented Samples
Human Microbiome	**Human** Commensal microbiota	Genomic	16S ribosomal RNA	Metagenomes from 15–18 body sites		
			WGS			
			WGS	>800 reference strains		
TCGA[a]	**Human** Thousands of tumor samples from 27 cancers	Genomic	WGS	21		Breast invasive carcinoma
			CNVs[b]	21		
		Transcriptome	Gene expression profiling[b]	19		
		Epigenomic	DNA methylation[b]	19		
Roadmap Epigenome[a]	**Human** 261 primary tissues and cultures total	Epigenomic	Chromatin marks (up to 30 antibodies)	82		H9, H1, IMR90
			DNA accessibility (up to 2 methods)	65		
			DNA methylation (up to 4 methods)	80		
			RNA-Seq	21		

The data sets available from initiatives were largely derived from human cell lines, individuals, or commensal bacteria, with the exception of the modENCODE data sets, which were derived from fly and worm. The types of biological samples used in each project are briefly described in the second column. Focusing on genomic, transcriptome, and epigenomic applications, we describe the different types of data sets currently available. Parentheses indicate the number of experimental variations used for an assay. For example, the Roadmap Epigenome Project assessed chromatin marks by ChIP-seq using 30 different antibodies and DNA methylation by four related assays (e.g., MRE-seq and MeDIP-seq). We also summarize the number of different cell lines, strains, or individuals for which data sets have been made available and highlight a select group of experimental treatments. When appropriate, the cell type(s) that is highly represented by a particular application is indicated. Abbreviations are as follows: whole-genome sequencing (WGS), whole-exome sequencing (WXS), rapid amplification of cDNA ends (RACE), cap analysis of gene expression (CAGE).

[a]ENCODE data sets as of January 14, 2013, http://genome.ucsc.edu/ENCODE/dataSummary.html, http://genome.ucsc.edu/ENCODE/dataSummaryMouse.html; TCGA as of June 18, 2012, https://tcga-data.nci.nih.gov/tcga/; Roadmap Epigenome Project as of June 19, 2012, http://www.genboree.org/epigenomeatlas/index.rhtml.

[b]Various methods were used.

2010; Gerstein et al., 2010), and the Epigenomics Roadmap (Chadwick, 2012) projects provide additional resources that are ideal for data integration (Table 1). Because these initiatives are specifically tasked with providing high-quality resources, common sets of biological samples, reagents, and methods are defined for each project. In the case of ENCODE and modENCODE data sets, established standards must be met for data to be released (The ENCODE Project Consortium, 2011; http://genome.ucsc.edu/ENCODE/protocols/dataStandards/). For example, the modENCODE project has evaluated the specificity and efficiency of commercial antibodies that are commonly used to generate ChIP-seq data sets (of which fewer than 75% passed muster) (Egelhofer et al., 2011). Thus, when using data from one of these public projects, users can be reasonably confident that the quality of the data meets or exceeds certain standards.

These initiatives also generate important control data sets as resources. For example, it is most appropriate to align sequence reads from RNA-seq and ChIP-seq experiments to the cell line-, strain-, or individual-specific genome rather than to a generic reference genome. Otherwise, single-nucleotide polymorphisms (SNPs) may be mistaken for RNA editing sites, and copy number variations (CNVs) may be mistaken for differential gene expression or changes in chromatin structure (Pickrell et al., 2011; Schrider et al., 2011). Accordingly, many of these initiatives have resequenced the genomes of the cell lines, strains, and individuals used in the projects.

Lastly, more than 1,500 curated databases are described in the *Nucleic Acids Research* online Molecular Biology Database Collection, many of which collect and integrate existing data to produce user-friendly, searchable websites (Fernández-Suárez and Galperin, 2013). As the field of bioinformatics has evolved to more effectively tackle specific questions in biology (Butte, 2009), cutting-edge databases have been designed to place mechanistic hypotheses within arm's reach of investigators by automating novel data integration strategies. For example, the HaploReg database integrates user-defined genome-wide association study results with linkage disequilibrium (The 1000 Genomes Project Consortium, 2010), sequence conservation information (Lindblad-Toh et al., 2011), and chromatin structure (Ernst et al., 2011) to link disease-associated genetic variation with putative regulatory elements (Ward and Kellis, 2012). Similarly, many of the large initiatives (e.g., ENCODE, modENCODE, etc.) also provide unified exploratory tools (e.g., UCSC and IGV genome browsers) that allow straightforward evaluation of genomic regions of interest.

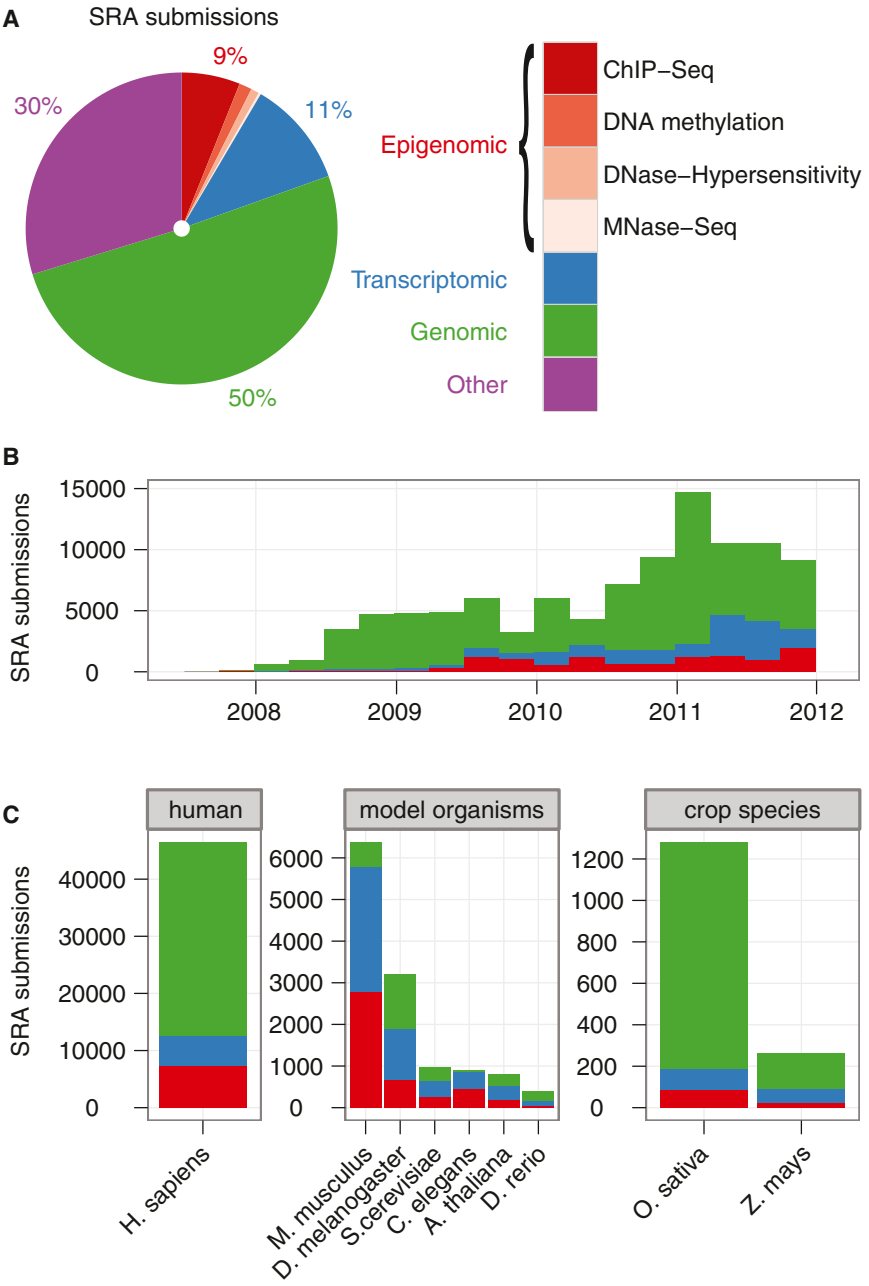

Figure 1. An Overview of the Publicly Available Data at the Sequence Read Archive Based on User-Submitted Metadata

(A) Half of the Sequence Read Archive (SRA) submissions have been generated by genomic library strategies, mostly whole-genome sequencing (green). The second half is composed of library strategies from transcriptome (blue), epigenomic (red), and other applications (purple). Epigenomic applications are diverse, despite composing <10% of all SRA submissions. These include numerous library strategies used to assess accessible or methylated DNA and nucleosomes (MNase-seq) or their posttranslational modifications (ChIP-seq) (shades of red).

(B) The SRA's growth rate, which is greatly increasing over time.

(C) Human SRA submissions, mostly whole-genome sequences, outnumber submissions from most other species by orders of magnitude.

NGS Data Sets: Prospects and Best Practices

New advances in NGS technologies are greatly expanding the current volume and the range of existing data (Metzker, 2010). As there is no evidence that innovations in sequencing technology are slowing down, it can only be anticipated that the pace of generating sequence data will continue to increase and the cost will decrease. By the start of 2012, approximately 75,000 genomic, 15,000 transcriptome, and 15,000 epigenomic submissions had been contributed to the SRA (Figure 1A). However, that volume of data represents only the tip of the iceberg, as transcriptome and epigenomic applications will be applied to include a greater range of cell types and species. Indeed, the number of transcriptome and epigenomic submissions has been steadily increasing, particularly in recent years (Figure 1B).

Transcriptome and epigenomic applications have been applied most liberally to humans and model organisms such as mouse, fly, worm, and yeast (Figure 1C). While the transcriptome data sets consist almost entirely of RNA-seq experiments, the epigenomic data sets are generated using a large collection of methods that interrogate various aspects of chromatin structure. The epigenomic experiments include methods to assess DNA accessibility (Boyle et al., 2008), DNA methylation (Laird, 2010), the genomic locations of transcription factors and chromatin marks (ChIP-seq) (Park, 2009; Schones and Zhao, 2008), and nucleosome positions (MNase-seq) (Jiang and Pugh, 2009) (Figure 1A, red bar graph). Additionally, specialized applications have also been submitted to the SRA that assess chromatin conformation (de Wit and de Laat, 2012; Dekker et al., 2002; Fullwood et al., 2010), RNA:protein interactions (Licatalosi and Darnell, 2010; Ule et al., 2005), RNA polymerase elongation (Churchman and Weissman, 2011; Core et al., 2008), and ribosome occupancy (Ingolia et al., 2009). In theory, any process related to nucleic acid metabolism can be assessed with the proper biochemical preparation, which makes NGS applications a rich and powerful source for integrative data analysis (Hawkins et al., 2010).

Furthermore, NGS data sets are extraordinarily rich. The sequence reads from a single experiment can provide a vast array of quantitative, positional, and sequence information. For instance, RNA-seq data sets provide sufficient information to measure mRNA expression levels and alternative splicing, to identify transcriptional start site and polyadenylation sites,

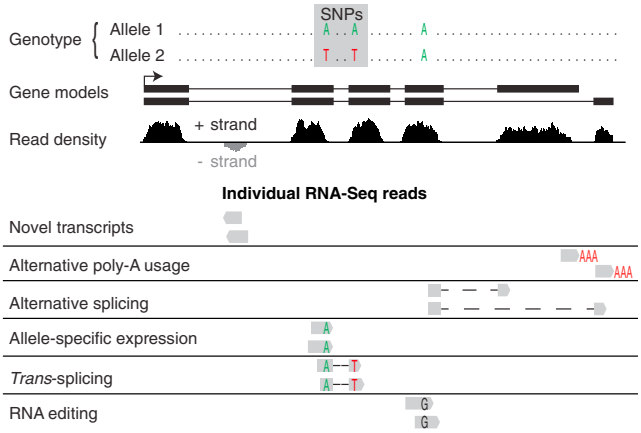

Figure 2. RNA-Seq Data Sets Are Information Rich
Due to the single nucleotide resolution and annotation-independent nature of RNA-seq data, many different processes can be analyzed from a single RNA-seq data set. For example, novel, unannotated transcripts can be identified as expressed regions that do not overlap with annotated gene models. Poly(A) sites can be identified by mining RNA-seq data sets for reads that span poly(A) sites (where part of the read contains genomic sequence and the rest of the read corresponds to the poly[A] tail). Alternative splicing can be identified and measured by identifying reads that map to splice-junction sequences specific to one isoform or another. Allele-specific expression can be quantified from reads that map in an allele-specific manner as determined by allele-specific SNPs. Similarly, *trans*-splicing can be monitored by identifying mate pairs where one read maps to one allele and the other read maps to the other allele. Finally, RNA editing sites can be identified and quantified by identifying RNA-seq reads that map accurately but contain a sequence different from the reference genome. In the case of allele-specific expression, *trans*-splicing, and RNA editing, it is critical to use a high-quality genome sequence containing confident SNP calls specific to the sample being studied, to use appropriate experimental and bioinformatic controls, and to validate the findings.

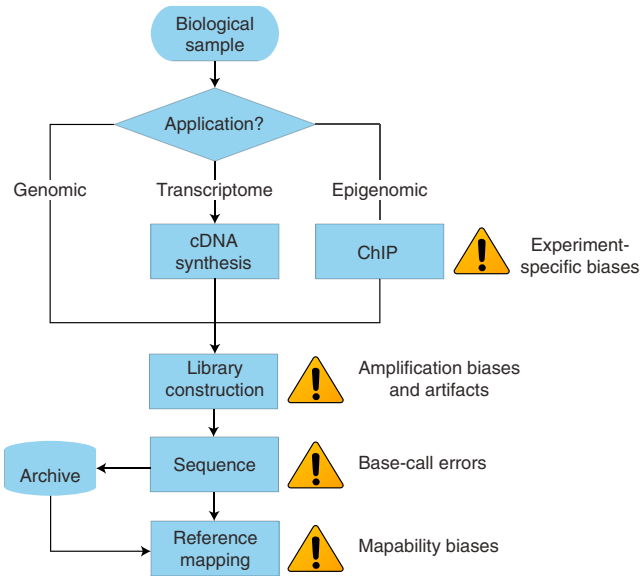

Figure 3. Stages at which Artifacts, Errors, and Biases Can Be Introduced in NGS Experiments and Analysis
Refer to text for details.

and to identify instances of RNA editing. In certain cases, allele-specific gene expression, allele-specific splicing, and even *trans*-splicing can be measured. Further, RNA-seq can be used as a discovery tool to annotate novel coding and noncoding transcripts as well as chimeric transcripts that result from genomic rearrangements (Figure 2) (Martin and Wang, 2011; McManus et al., 2010; Ozsolak and Milos, 2011; Wang et al., 2009). Though these data sets are very rich, they must be analyzed carefully. Essentially every step involved in generating NGS data introduces detectable, sometimes substantial, biases or errors (Figure 3). This presents a particular challenge for data integration, since different sequencing platforms, biochemical procedures, and data processing methods are associated with unique caveats. With the proper controls, however, these effects can often be identified and accounted for in downstream analyses.

NGS platforms have been in use long enough that biases attributable to library construction and sequencing have been evaluated in great detail. Data generated by the Illumina platform, for example, are subject to base-call errors that increase with read position due to phasing issues (Dohm et al., 2008) and underrepresentation of high and low GC content reads (Dohm et al., 2008; Risso et al., 2011). As these technical issues are well characterized, popular analysis packages attempt to correct for such nucleotide biases. For example, Cufflinks, a popular program used to measure differential gene expression, empirically determines nucleotide biases present in RNA-seq data sets and corrects for them (Meacham et al., 2011; Trapnell et al., 2012). While this strategy vastly improves comparisons between independently generated data sets, and even from different sequencing platforms, third-generation sequencing platforms, such as those from IonTorrent, PacificBiosciences and Oxford Nanopore (and other lurking companies), use radically different chemistries, the biases of which will need to be identified and remedied.

Less recognized are the myriad experiment-specific biases or artifacts that are introduced at nearly every step involved in preparing libraries and sequencing them. For example, it has clearly been shown that RNA-seq libraries prepared using random-hexamer priming display a systematic nontemplated sequence profile at the beginning of reads which is primarily due to first-strand synthesis (Hansen et al., 2010). This technical artifact appears to be a major source of the controversial ~10,000 "RNA DNA differences" (proposed RNA editing sites) identified from human cell lines that were recently reported (Li et al., 2011) and subsequently called into question (Kleinman and Majewski, 2012; Lin et al., 2012; Pickrell et al., 2012). Similarly, ChIP-seq experiments overrepresent regions of open chromatin, which can create false positives (Chen et al., 2012).

Another technical artifact associated with library preparations is template switching that occurs during the amplification steps, which can give rise to molecules that did not exist in the initial biological sample. For RNA-seq experiments, this type of artifact can generate "chimeric" RNAs that appear to be synthesized by *trans*-splicing or some unknown biological process (Gingeras, 2009; McManus et al., 2010). Sequence data from control libraries that assess the frequency of template switching are absolutely essential to distinguish biologically derived

chimeric RNAs from those generated artifactually (McManus et al., 2010).

In addition to experimental artifacts, bioinformatic artifacts can severely impact data interpretation. A major source of these artifacts is the mappability of NGS sequence reads, which are typically 25–100 nt in length. Mapping these sequences to a reference genome can be particularly problematic due to the plethora of repetitive elements present in most genomes. Repetitive elements such as LINEs and SINEs have always presented difficulties for correctly mapping sequences, but the short size of NGS reads significantly amplifies this problem—as read length decreases, so does the number of unique regions that can be mapped within a reference genome (Treangen and Salzberg, 2012). Consequently, ''mappability'' differs depending on read length. Such biases can create illusory nonrandom associations with biological features (e.g., exons) in ChIP- and MNase-seq experiments. For example, with 32 bp reads, tiny but common genomic features, such as coding starts, ends, exons, and splice sites accumulate greater read densities than other local features (e.g., introns) (Schwartz et al., 2011). Thus, mappability must be carefully considered when interpreting any type of alignment data.

A second bioinformatic artifact is caused by genetic variation that has not been accounted for. CNVs that differ between experimental samples and reference genomes, for example, can create false-positive enrichment regions in ChIP-seq experiments (Pickrell et al., 2011). Studies of RNA editing are particularly susceptible to high false discovery rates if SNPs are not accounted for in the analysis. In the case of RNA DNA differences (Li et al., 2011), 55% match the genome of at least one of the 27 individual genomes used in the original analysis, which suggests that the relatively low coverage (2–6×) of these genomes was not sufficient to identify and eliminate confounding SNPs (Schrider et al., 2011). Thus, a matched reference genome—with sufficiently deep coverage—should be used when mapping short reads, since experimental samples such as cell lines, strains, and individuals may differ in their SNPs, CNVs, and chromosome number. Even then, care must be taken to ensure that interesting results are not correlated with regions of shallow sequence depth.

NGS technologies are rapidly evolving. Consequently, robust computational methods lag behind this moving target. Thus, data generated by the early adoption of exciting new technologies should be evaluated, first and foremost, with critical attention to sequence biases. It is our opinion that novel phenomena will increasingly be discovered by the use of existing data. However, these phenomena are only as compelling as the support for an underlying mechanism. The vast potential for technical artifacts in NGS data are more than enough reason for caution, particularly since some technical artifacts are, in fact, not random and may correlate strongly with known biological features. Here, we are reminded of a sage warning made by Daniel MacArthur in reference to remarkable results obtained by analyzing large NGS data sets: ''The more surprising a result seems to be, the less likely it is to be true'' (MacArthur et al., 2012). We implore data analysts to heed this warning and perform extensive validation of remarkable findings, or they may indeed fall victim to MacArthur's rule.

Successful Uses of Existing Data

The Watson-Crick model of DNA—the double helix—was developed largely through model building informed by existing data (Watson and Crick, 1953a, 1953b). Although structural evidence supporting the model required meticulous experimentation (Franklin and Gosling, 1953; Wilkins et al., 1953), the model alone suggested the basis of genetic inheritance as well as DNA replication and recombination (Watson and Crick, 1953a, 1953b). Despite lacking any knowledge of the complex protein machines responsible for replication and recombination, the central predictions were, nonetheless, essentially proven within a decade (Alberts, 2003). Thus, the double helix exemplifies how the insightful analysis of existing data can revolutionize a field.

There are numerous examples in which insights into the molecular mechanisms of various biological processes have been gleaned from analyzing existing data. Below we discuss several of these and attempt to highlight the general aspects of each study as a lesson for how each approach can be applied to other problems. We focus on problems related to various aspects of RNA biology, though these approaches can be used for other molecular processes as well.

Identifying Functional Elements through Conservation

Functionally important sequence elements are expected to be conserved over time. Thus, one way of investigating a particular process is to identify conserved sequence elements using alignments of multiple whole-genome sequences. Conservation plots can be generated from such alignments using several software packages that calculate nucleotide substitution rates. Conveniently, conservation scores based on whole-genome alignments using phastCons (Siepel et al., 2005), phyloP (Pollard et al., 2010), or SiPhy (Garber et al., 2009) can be downloaded from the UCSC genome browser for many model organisms, including yeast, worm, fly, and various mammals including mouse and human (Kent et al., 2002). Analyzing the identity, characteristics, and locations of conserved sequences can provide tremendous mechanistic insight. Below, we highlight two such examples, in the fields of RNA editing and alternative splicing, that utilized this approach.

Adenosine-to-inosine (A-to-I) editing of RNA is an evolutionarily conserved process catalyzed by the ADAR family of adenosine deaminases (reviewed in Rieder and Reenan, 2011). A mystery that dogged the field for some time was the paucity of known endogenous RNA targets—only a few chance discoveries had been described—despite evidence that the inosine content of mRNA isolated from brain tissues might be as high as 1 in every 17,000 nucleotides (Paul and Bass, 1998). Armed with the knowledge that ADAR mutations resulted in neurological defects (Palladino et al., 2000) and that ADAR editing required an RNA duplex formed between the targeted region (containing the edited adenosine) and a complementary sequence (Higuchi et al., 1993), Hoopengardener and colleagues (Hoopengardner et al., 2003) searched for new targets of RNA editing among the neuronally expressed genes of *Drosophila*. In the case of *para*, which encodes a Na^+ channel, the exon containing a known editing site is very highly conserved near the editing site, as is a region in the adjacent intron which base pairs with the edited exon. Based on this observation, Hoopengardner et al. (2003)

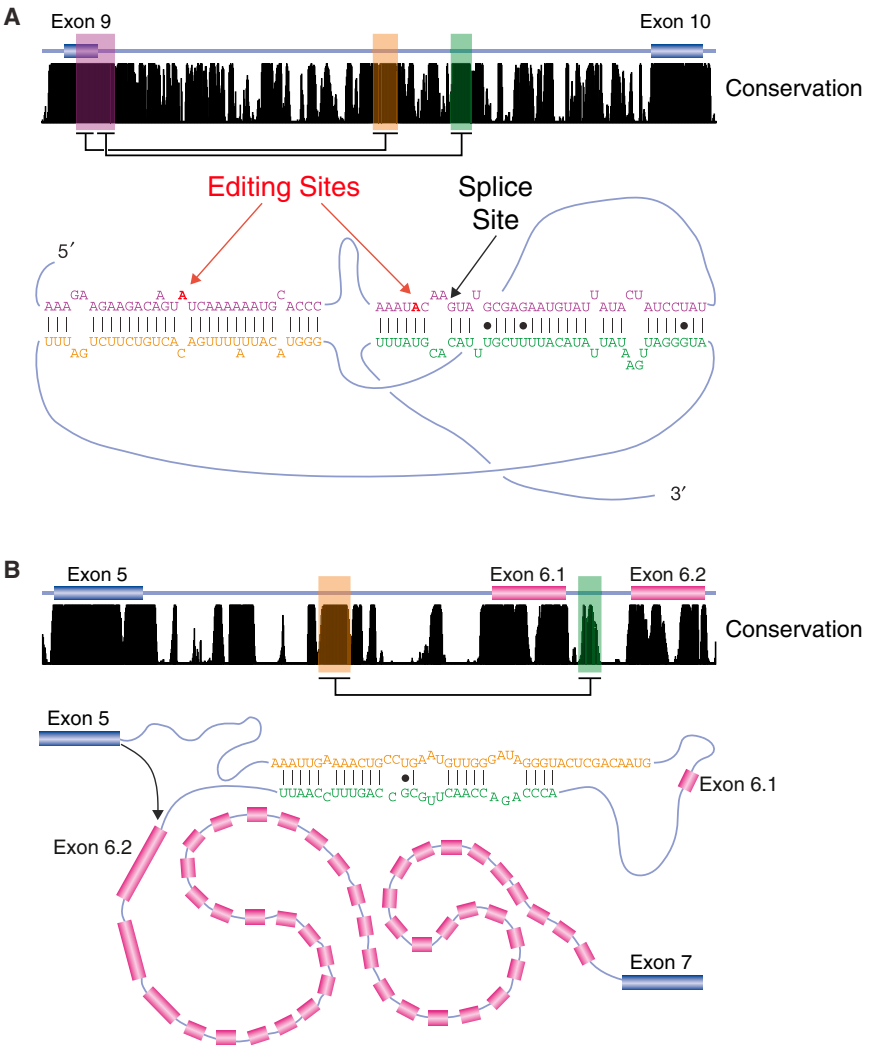

Figure 4. Comparative Approaches Reveal Insights into the Mechanisms of RNA Editing and Alternative Splicing

(A) The *Drosophila synaptotagmin* (*syt*) gene contains two editing sites in exon 9 (indicated as red adenosine [A] residues [bottom]). On the top, the region between exons 9 and 10 is shown along with the insect conservation track. Editing within this exon was first identified by the high extent of conservation (top) (Hoopengardner et al., 2003). Subsequent studies revealed that editing is directed by RNA duplexes formed between the highly conserved sequences highlighted in purple and the two highly conserved intronic sequences highlighted in orange and green. These sequence elements form a pseudoknot structure that places the edited residues into a double-stranded RNA structure that can be recognized by dADAR.

(B) The exon 6 cluster of the *Drosophila Down syndrome cell adhesion molecule* (*Dscam*) gene contains 48 alternative exons, only one of which is included in each mRNA. On the top, the region from constitutive exon 5 and the first two alternative exons is shown along with the insect conservation track. The intron between exons 5 and 6.1 contains a highly conserved sequence called the docking site (orange), which can base pair with a selector sequence located upstream of each variable exon. In this case, the selector sequence for exon 6.2 is highlighted in green. On the bottom, the base pairing between the docking site and the exon 6.2 selector sequence is shown.

reasoned that they might be able to identify new RNA editing targets by identifying very highly conserved exons. They therefore assessed 914 neuronally expressed genes to identify exons with a high level of sequence constraint between *Drosophila melanogaster* and *Drosophila pseudoobscura*. This approach proved highly productive, as 16 new editing sites were identified that were validated by cDNA sequencing (Hoopengardner et al., 2003) (Figure 4A illustrates one such example). Importantly, this use of comparative genomics demonstrated a previously unanticipated degree of phylogenetic conservation between A-to-I editing sites, solidified the RNA duplex-dependent mechanism of ADAR function, and provided a facile bioinformatic strategy for editing-site identification. Indeed, improved variations of this approach using existing EST:genome alignments (Levanon et al., 2004) or archived sequence chromatograms (Zaranek et al., 2010) have now greatly increased the number of high-confidence A-to-I editing candidates (reviewed in Wulff et al., 2011).

A similar approach has been used to uncover novel mechanisms that regulate alternative splicing. In one particularly illustrative case, the *Drosophila Dscam* gene, conservation within introns was used to uncover a novel mechanism of mutually exclusive splicing (Graveley, 2005). *Dscam* is a textbook example of the importance of alternative splicing in increasing protein diversity, as it may generate more than 38,000 different protein isoforms (Schmucker et al., 2000). Each time the *Dscam* gene is transcribed, the pre-mRNA is spliced such that each mRNA contains one and only one exon from each of four exon clusters (exons 4, 6, 9, and 17, specifically). But at the time of its discovery, no previously described mechanism of mutually exclusive splicing could explain how the many variable exons of *Dscam* are spliced in a mutually exclusive manner.

Compared to exons, which are often highly conserved due to their coding potential, intronic regions typically have little conservation except at sites that have noncoding function, such as RNA splicing. Nucleotide alignments from 15 insect species revealed two types of conserved sequence elements in the introns of the exon 6 cluster (Graveley, 2005). The first element, the docking site, was located between exon 5 and the first exon 6 variant. Importantly, the docking site was found only once in the exon 6 cluster, but was present in every species examined, even species that diverged more than 450 million years ago. Selector sequences, the second type of sequence element, were located in introns upstream of each exon 6 variant. Based on their complementary sequences, the docking site and selector sites appeared to base pair with one another;

however, one and only one of the selector sequences could base pair with the docking site at a time (Figure 4B). Thus, base pairing of one selector sequence with the docking site would be predicted to promote inclusion of that exon while simultaneously inhibiting the splicing of the 47 other exon 6 variants. Though this mechanism was discovered purely by comparative genomics and bioinformatics, the elegance and universal conservation among insects lent credence to the proposed mechanism. Experimental confirmation of the model was subsequently obtained using mutagenized BACs containing the entire *Dscam* gene (May et al., 2011). Further demonstrating the power of this approach, additional docking site:selector sequences within the other *Dscam* clusters and even other alternatively spliced genes have been identified largely on the basis of intronic conservation (Yang et al., 2011).

These two examples illustrate how conservation can be used to identify functional elements that provided insight into the mechanisms of RNA editing and alternative splicing. However, these approaches can be used to study many other processes. For instance, candidate functions have been assigned to ~60% of the conserved sequences in mammals (Lindblad-Toh et al., 2011), yet 40% of these elements have unknown functions. Moreover, another 10,000 regions of mammalian coding sequences are predicted to have overlapping functions; yet again, these functions are mostly unknown. Thus, a fruitful avenue of research is to use existing multiple sequence alignments to identify the conserved sequence elements associated with a gene or process of interest. The function of conserved sequences will likely require experimental approaches to determine their functions, but they may also be inferred based on their sequence features or locations alone.

Functional Relationships Identified through Data Integration

More recently, analyses of existing data have made a significant impact on the burgeoning study of cotranscriptional splicing. A growing body of evidence now supports the notion that transcription and splicing are not only concurrent but also coupled, such that transcriptional dynamics profoundly influence RNA splicing (Neugebauer, 2002). For example, the elongation rate of RNA polymerase can influence the propensity for exon skipping (Kornblihtt et al., 2004). Until recently, however, little was known about the relationship between cotranscriptional splicing and the chromatin context in which it takes place. By integrating existing epigenomic data sets with known splicing patterns, recent studies have generated exciting new hypotheses that intimately connect chromatin structure to RNA splicing.

A major challenge associated with studying chromatin structure is its immense complexity. Even the most fundamental unit of chromatin, the nucleosome, can differ between genomic regions in occupancy, positioning, and myriad posttranslational modifications (a.k.a. chromatin marks). It has long been observed in *S. cerevisiae*, for instance, that the chromatin mark, histone H3 lysine 36 trimethyation (H3K36me3), is enriched within the bodies of active genes (reviewed in Li et al., 2007). Thus H3K36me3, and many other marks, are thought to be intimately associated with transcriptional processes. Questions concerning whether chromatin marks might affect, or be affected by, splicing were rarely discussed until genome-wide ChIP surveys in *C. elegans* demonstrated higher levels of H3K36me3 at exons compared to nearby introns within the same gene (Kolasinska-Zwierz et al., 2009). As this finding was entirely unanticipated, Kolasinska-Zwierz et al. turned to publicly available H3K36me3 ChIP-seq data (Barski et al., 2007) to test the relevance of their findings to humans. By aggregating all of the H3K36me3 sequence reads that aligned near exons (a so-called "metagene" analysis), the authors observed strikingly similar H3K36me3 enrichment at the average human exon (Figures 5A and 5B, blue panel).

Furthermore, integrating H3K36me3 ChIP data with annotations of alternative and constitutive exons revealed a potential connection between the degree of chromatin marking and alternative splicing. A modest but suggestive decrease in H3K36me3 reads at alternatively spliced exons was observed in both worm and mouse (Kolasinska-Zwierz et al., 2009). Similar analyses have been somewhat conflicting (Hon et al., 2009; Spies et al., 2009), but this may simply indicate that the differences between constitutive and alternatively spliced exons are subtle. Nonetheless, the notion that chromatin marks, and in particular H3K36me3, can affect alternative splicing is consistent with several recently reported experiments. Indeed, two H3K36me3-binding proteins, MRG15 and PSIP1, have been implicated in mediating alternative splicing regulation by the splicing factors PTB and SRSF1 (ASF/SF2), respectively (Luco et al., 2010; Pradeepa et al., 2012). Additional marks, such as H3K4me3 and H3 acetylations, have also been shown to influence splicing (Gunderson and Johnson, 2009; Sims et al., 2007). Most likely, chromatin marks function in splicing by modulating RNA polymerase elongation rates or by recruiting specific splicing factors to active genes (Luco et al., 2011, and references therein; Nilsen and Graveley, 2010). Intriguingly, depletion of the only known H3K36me3 methyltransferase influences the splicing of only a small, but significant, subset of PTB-dependent splicing events (Luco et al., 2010). Such gene-specific affects might also explain why genome-wide correlations between H3K36me3 and alternative splicing have been modest.

Numerous studies have since analyzed more than 41 chromatin marks using publicly available epigenomic data sets, which have yielded a strong consensus: chromatin structure reflects gene architecture. In humans, three types of exon/intron boundaries have been shown to be associated with particular chromatin marks: (1) H3K4me3 and H3K9Ac throughout the length of the first exon (Bieberstein et al., 2012), (2) H3K79me2 (among several other marks) throughout the length of the first intron (Huff et al., 2010), and most notably (3) H3K36me3 enrichment at internal exons (Andersson et al., 2009; Hon et al., 2009; Huff et al., 2010; Spies et al., 2009). Because these chromatin marks are intimately associated with transcription, these results suggest a much closer connection between the splicing and transcription machineries than previously thought.

Similarly, analysis of published MNase-seq data (Schones et al., 2008; Valouev et al., 2008) revealed that nucleosomes themselves were also highly associated with internal exons (Andersson et al., 2009; Huff et al., 2010; Schwartz et al., 2009; Spies et al., 2009; Tilgner et al., 2009). Based on this observation,

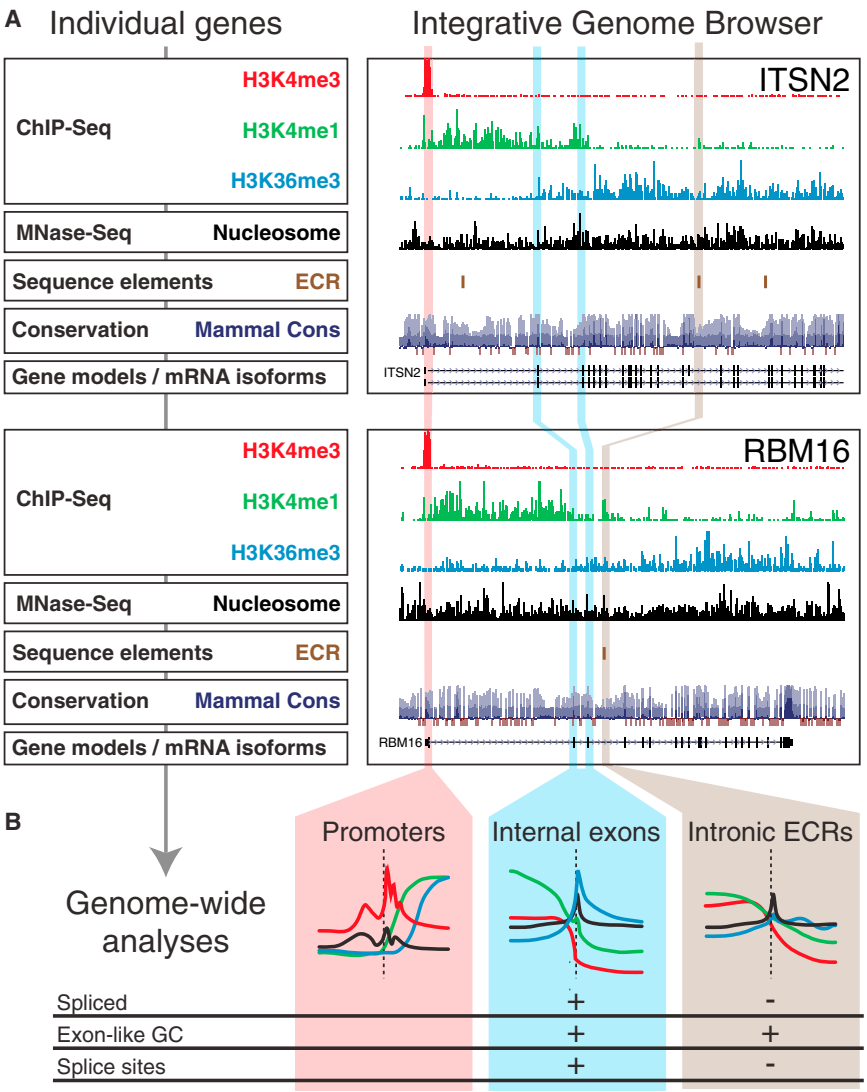

Figure 5. From Observations at Individual Genes to Genome-wide Analyses

(A) Data integration using the UCSC genome browser. ChIP-seq (Barski et al., 2007), MNase-seq (Schones et al., 2008), and ECR (Spies et al., 2009) positions were uploaded to the UCSC genome browser to compare the chromatin structures of the ITSN2 (top panel) and RBM16 (bottom panel) genes. ChIP-seq reads for H3K4me3 (red), but not H3K36me3 (blue), are enriched at the promoter of each gene.

(B) Genome-wide aggregate "metagene" analyses demonstrate that chromatin structure reflects gene architecture. For example, H3K4me3 (red line) and H3K36me3 (blue line) are enriched at the average promoter (red panel) and the average internal exon (blue panel), respectively. Similarly, metagene analyses show that nucleosomes are enriched at the average internal exon (blue panel) and the average ECR (brown panel), which suggests that exon-like sequence content alone is sufficient for high nucleosome occupancy (Spies et al., 2009).

a novel mechanism for exon definition has been proposed (but yet to be proven), whereby nucleosome occupied exons serve as RNA polymerase "speed bumps" that provide additional time for the spliceosomal machinery to recognize nearby splice sites (Schwartz et al., 2009; Tilgner et al., 2009).

In theory, elevated nucleosome occupancy at exons alone could explain the previously identified enrichment of H3K36me3 at exons (Schwartz et al., 2009; Tilgner et al., 2009). However, bioinformatic analyses comparing nucleosome occupancy at exons, exon-like composition regions (ECRs), and pseudoexons have uncovered evidence that the mechanisms determining nucleosome occupancy and H3K36me3 enrichment at exons are distinct. To discern which aspects of exon sequence might be necessary and sufficient for high nucleosome occupancy, sequence characteristics of exons were analyzed separately for high nucleosome occupancy. In this case, the average ECR, which is not flanked by splice sites but does have the same GC content of the average exon, displayed nucleosome occupancies equal to that of the average exon (Spies et al., 2009) (Figure 5B, brown panel). Conversely, the average pseudoexon, which has lower GC content than the average exon, was depleted for nucleosome occupancy (Tilgner et al., 2009). Thus, the DNA sequence composition of exons alone may be sufficient for exon-like nucleosome occupancy. Lastly, ECRs were not enriched for H3K36me3, which suggests that exon marking reflects some aspect of splicing rather than exon-like sequence composition (Huff et al., 2010) (Figure 5B, brown panel). Further supporting a role for the spliceosome in specifying chromatin structure, recent experiments in which splicing was inhibited by splice site mutations or spliceostatin exposure have indeed caused changes in H3K36me3 marking (de Almeida et al., 2011; Kim et al., 2011).

The above analyses demonstrate the power of data integration to establish new connections between related processes whose mechanistic links may yet be unclear. In some cases, these relationships can be readily observed at single genes using a genome browser to facilitate comparisons between data sets. In such cases, moving from observations at single loci to genome-wide analyses can be accomplished by aggregating values from genome-wide data sets at specific features of interest (Figure 5). Genome-wide relationships can also be revealed by plotting all relevant loci in a high-density heatmap that is aligned and sorted to highlight features of interest (Hawkins et al., 2010). The latter approach is particularly useful in instances in which summary statistics like the genome-wide averages might be deceptive (e.g., the mean of a bimodal distribution). Thus, by integrating and aggregating existing data, a mere anecdote can be transformed into a global principle.

Estimating the Frequency and Functional Consequences of Poorly Characterized Biological Phenomena

Publicly available data can also be used to assess the prevalence and functional consequences of previously ignored biological phenomena. In animals, alternatively spliced genes are the norm; more than 92% of human genes produce at least one alternatively spliced transcript (Pan et al., 2008; Wang et al., 2008a). While many of these alternatively spliced transcripts are predicted to encode functionally distinct protein isoforms, others encode protein isoforms whose biological relevance is questionable. Thus, the perennial question: which of these splicing events are regulated and which are stochastic?

For instance, alternative splicing may introduce premature termination codons (PTCs) that target the message for degradation by nonsense-mediated decay (NMD). Such unproductively spliced transcripts could be regulated to function as posttranscriptional on/off switches, or merely splicing mistakes in need of triage. Only a few genes were previously known to be regulated by unproductive splicing. Publicly available EST data, on the other hand, suggested that nearly one-third of alternatively spliced transcripts were potential NMD targets (Lewis et al., 2003). This unexpectedly high prevalence brought new attention to the hypothesis that unproductive splicing might posttranscriptionally regulate the expression of entire classes of genes. Controversy initially surrounded the original EST-based estimates because experiments that depleted NMD factors to identify stabilized unproductively spliced transcripts yielded more conservative estimates. By microarray, only ~10% of cassette exons substantially elicited NMD (Pan et al., 2006), and tissue-specific regulation was found to be rare. More recent approaches using RNA-seq, which allow for all major forms of splicing to be considered, have brought these numbers closer to initial estimates, but only for some tissues (Weischenfeldt et al., 2012). Nonetheless, the question as to whether a significant portion of unproductive splicing regulates the expression of entire classes of genes was answered through analyses that showed that unproductively spliced transcripts were enriched for genes encoding splicing factors and other RNA-binding proteins (Ni et al., 2007; Saltzman et al., 2008). Another experimental study demonstrated that the entire family of human SR proteins was associated with unproductive splicing (Lareau et al., 2007). These studies satisfyingly confirmed and extended previous reports of autoregulation by unproductive splicing (reviewed in McGlincy and Smith, 2008). Thus, as with the previous examples, an initial breakthrough was achieved through the analysis of existing data, followed by further refinement and proof through experimental studies.

The functional consequences of alternative splicing decisions that produce nearly identical protein isoforms has also been assessed using publicly available data sets. In this case, introns ending in NAGNAG (a tandem duplication of the 3′ splice site NAG) have been previously shown to be alternatively spliced such that their protein isoforms differ by only a single amino acid based on publicly available EST data (Hiller et al., 2004; Hiller and Platzer, 2008). However, whether these small differences are regulated or stochastic has been questioned (Chern et al., 2006). By first analyzing their own experimental data, Bradley et al. (2012) confirmed that there is broad use of NAGNAG splicing in human and mouse tissues. Motivated by these findings, the authors also mined the extensive collection of RNA-seq data sets generated by the *Drosophila* and *C. elegans* modENCODE projects. Strikingly, 500 NAGNAGs splice sites were found to be alternatively spliced in at least one of 30 developmental time points in *Drosophila*, while NAGNAG splicing in *C. elegans* was considerably less dynamic. Approximately 5%–14% of alternatively spliced NAGNAGs were found to be developmentally regulated and conserved, such that the most dynamically spliced NAGNAGs were associated with the greatest intronic sequence conservation. While the mechanisms regulating NAGNAG splicing remain unclear, these analyses provide the best evidence to date that even small changes in splicing are commonly regulated (at least in some animals).

The above examples demonstrate the utility of large data sets to assess the prevalence of intriguing phenomena. With large collections of data sets, the prevalence of tissue-specific or developmental regulation can be estimated with tremendous breadth. In such pursuits, however, it is essential to accurately assess the false discovery rate of the analysis. In the above example of NAGNAG splicing, the mean false discovery rates for technical and biological replicates were estimated at a very reasonable 4.4% and 1.1%, respectively. This is another area in which heeding MacArthur's rule is well advised.

Unsupervised Methods

Here we have highlighted successful strategies for supervised data mining of existing data. Although not discussed at length here, unsupervised methods like machine learning algorithms also demonstrate great promise for researchers seeking to unbiasedly analyze large data sets. A machine learning algorithm has recently been used to write a first draft of the splicing code—a set of rules so expansive that it is best referenced with the aid of a computer (Barash et al., 2010). This approach led Barash et al. to identify novel sequence motifs associated with regulated alternative splicing as well as a new example of developmentally regulated unproductive splicing. As the quantity of available data sets increases, it is almost certain that unsupervised methods will become more broadly used for analyses.

Analytical Skills Needed for the Future

In a recent poll, most scientists reported that they "rarely" accessed data or used data sets from the published literature for their original research papers (Science Staff, 2011). However, this will undoubtedly change. The opportunities and challenges of "Big Data" are being felt not only in the biological sciences, but also in society at large (Lohr, 2012). Thus, students will gain transferable skills from exercises that teach basic workflows using large data sets and scripting languages.

At a minimum, we envision teaching exercises that require biology students to devise an experimental design using only existing data; access relevant data sets from archives; parse and integrate data using programing languages such as Perl, Python, Ruby, or R; and apply an appropriate visualization technique. Laboratory protocols for the use of analytic software currently exist to aid these pursuits (e.g., Cufflinks) (Trapnell et al., 2012). Such exercises will empower students to explore

and assess the quantitative data published in the manuscripts that they read, which can no longer be assessed at a glance like the qualitative gel-based results on which molecular biology was founded. Ultimately, it will be equally important to know how to write code as it is to pipette.

ACKNOWLEDGMENTS

We thank members of the Graveley lab and Jason Huff for discussions and comments on the manuscript. Work in the Graveley lab is supported by NIH grants R01GM067842, R01GM095296, and U54HG007005 to B.R.G. and U54HG006994 and U01HG004271 to Susan E. Celniker (co-principal investigator to B.R.G.) and a by Ruth L. Kirschstein National Research Service Award (F32GM105264) to A.M.P.

REFERENCES

Alberts, B. (2003). DNA replication and recombination. Nature *421*, 431–435.

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. *19*, 1732–1741.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*(D1), D991–D995.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep. *2*, 62–68.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311–322.

Bradley, R.K., Merkin, J., Lambert, N.J., and Burge, C.B. (2012). Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol. *10*, e1001229. http://dx.doi.org/10.1371/journal.pbio.1001229.

Butte, A.J. (2009). Translational bioinformatics applications in genome medicine. Genome Med. *1*, 64. http://dx.doi.org/10.1186/gm64.

Chadwick, L.H. (2012). The NIH Roadmap Epigenomics Program data resource. Epigenomics *4*, 317–324.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J.O., Slattery, M., Liu, T., Zhang, Y., Kim, T.K., He, H.H., Zieba, J., et al. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. Nat. Methods *9*, 609–614.

Chern, T.M., van Nimwegen, E., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Zavolan, M. (2006). A simple physical model predicts small exon length variations. PLoS Genet. *2*, e45. http://dx.doi.org/10.1371/journal.pgen.0020045.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M., et al.; 1000 Genomes Project Consortium. (2012). The 1000 Genomes Project: data management and community access. Nat. Methods *9*, 459–462.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., et al. (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat. Struct. Mol. Biol. *18*, 977–983.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev. *26*, 11–24.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. *36*, e105. http://dx.doi.org/10.1093/nar/gkn425.

Egelhofer, T.A., Minoda, A., Klugman, S., Lee, K., Kolasinska-Zwierz, P., Alekseyenko, A.A., Cheung, M.-S., Day, D.S., Gadel, S., Gorchakov, A.A., et al. (2011). An assessment of histone-modification antibody quality. Nat. Struct. Mol. Biol. *18*, 91–93.

ENCODE Project Consortium; Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Fernández-Suárez, X.M., and Galperin, M.Y. (2013). The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. Nucleic Acids Res. *41*(D1), D1–D7.

Franklin, R.E., and Gosling, R.G. (1953). Molecular configuration in sodium thymonucleate. Nature *171*, 740–741.

Fullwood, M.J., Han, Y., Wei, C.L., Ruan, X., and Ruan, Y. (2010). Chromatin interaction analysis using paired-end tag sequencing. Curr. Protoc. Mol. Biol. *Chapter 21*, Unit 21.15, 21–25.

Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics *25*, i54–62. http://dx.doi.org/10.1093/bioinformatics/btp190.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al.; modENCODE Consortium. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science *330*, 1775–1787.

Gingeras, T.R. (2009). Implications of chimaeric non-co-linear transcripts. Nature *461*, 206–211.

Graveley, B.R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. Cell *123*, 65–73.

Gunderson, F.Q., and Johnson, T.L. (2009). Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. PLoS Genet. *5*, e1000682. http://dx.doi.org/10.1371/journal.pgen.1000682.

Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. *38*, e131. http://dx.doi.org/10.1093/nar/gkq224.

Hawkins, R.D., Hon, G.C., and Ren, B. (2010). Next-generation genomics: an integrative approach. Nat. Rev. Genet. *11*, 476–486.

Higuchi, M., Single, F.N., Köhler, M., Sommer, B., Sprengel, R., and Seeburg, P.H. (1993). RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. Cell *75*, 1361–1370.

Hiller, M., and Platzer, M. (2008). Widespread and subtle: alternative splicing at short-distance tandem sites. Trends Genet. *24*, 246–255.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. (2004). Widespread occurrence of alternative splicing

at NAGNAG acceptors contributes to proteome plasticity. Nat. Genet. *36*, 1255–1257.

Hon, G., Wang, W., and Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput. Biol. *5*, e1000566. http://dx.doi.org/10.1371/journal.pcbi.1000566.

Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R. (2003). Nervous system targets of RNA editing identified by comparative genomics. Science *301*, 832–836.

Huff, J.T., Plocik, A.M., Guthrie, C., and Yamamoto, K.R. (2010). Reciprocal intronic and exonic histone modification regions in humans. Nat. Struct. Mol. Biol. *17*, 1495–1499.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. *10*, 161–172.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kim, S., Kim, H., Fong, N., Erickson, B., and Bentley, D.L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc. Natl. Acad. Sci. USA *108*, 13564–13569.

Kleinman, C.L., and Majewski, J. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302, author reply 1302.

Kodama, Y., Shumway, M., and Leinonen, R.; International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. *40*, D54–D56. http://dx.doi.org/10.1093/nar/gkr854.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. *41*, 376–381.

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., and Nogues, G. (2004). Multiple links between transcription and splicing. RNA *10*, 1489–1498.

Laird, P.W. (2010). Principles and challenges of genomewide DNA methylation analysis. Nat. Rev. Genet. *11*, 191–203.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., and Brenner, S.E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature *446*, 926–929.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. *22*, 1001–1005.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. USA *100*, 189–192.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M., and Cheung, V.G. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. Science *333*, 53–58.

Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat. Rev. Genet. *11*, 75–87.

Lin, W., Piskol, R., Tan, M.H., and Li, J.B. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302, author reply 1302.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al.; Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

Lohr, S. (2012). Big data's impact in the world. The New York Times, February 11, 2012.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996–1000.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell *144*, 16–26.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al.; 1000 Genomes Project Consortium. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. *12*, 671–682.

May, G.E., Olson, S., McManus, C.J., and Graveley, B.R. (2011). Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster. RNA *17*, 222–229.

McGlincy, N.J., and Smith, C.W. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? Trends Biochem. Sci. *33*, 385–393.

McManus, C.J., Duff, M.O., Eipper-Mains, J., and Graveley, B.R. (2010). Global analysis of trans-splicing in Drosophila. Proc. Natl. Acad. Sci. USA *107*, 12975–12979.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D.I., Singer, M., and Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics *12*, 451. http://dx.doi.org/10.1186/1471-2105-12-451.

Metzker, M.L. (2010). Sequencing technologies—the next generation. Nat. Rev. Genet. *11*, 31–46.

modENCODE Project Consortium; Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., et al. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787–1797.

Neugebauer, K.M. (2002). On the importance of being co-transcriptional. J. Cell Sci. *115*, 3865–3871.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M., Jr. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. *21*, 708–718.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Palladino, M.J., Keegan, L.P., O'Connell, M.A., and Reenan, R.A. (2000). A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. Cell *102*, 437–449.

Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J., and Blencowe, B.J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev. *20*, 153–158.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

Paul, M.S., and Bass, B.L. (1998). Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. EMBO J. *17*, 1120–1127.

Pekowska, A., Benoukraf, T., Ferrier, P., and Spicuglia, S. (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. Genome Res. *20*, 1493–1502.

Pickrell, J.K., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics *27*, 2144–2146.

Pickrell, J.K., Gilad, Y., and Pritchard, J.K. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302, author reply 1302.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Pradeepa, M.M., Sutherland, H.G., Ule, J., Grimes, G.R., and Bickmore, W.A. (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. PLoS Genet. *8*, e1002717. http://dx.doi.org/10.1371/journal.pgen.1002717.

Rieder, L.E., and Reenan, R.A. (2011). The intricate relationship between RNA structure, editing, and splicing. Semin. Cell Dev. Biol. *23*, 281–288.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. BMC Bioinformatics *12*, 480. http://dx.doi.org/10.1186/1471-2105-12-480.

Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol. Cell. Biol. *28*, 4320–4330.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*, 671–684.

Schones, D.E., and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. Nat. Rev. Genet. *9*, 179–191.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Schrider, D.R., Gout, J.F., and Hahn, M.W. (2011). Very few RNA and DNA sequence differences in the human transcriptome. PLoS ONE *6*, e25842. http://dx.doi.org/10.1371/journal.pone.0025842.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16*, 990–995.

Schwartz, S., Oren, R., and Ast, G. (2011). Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS ONE *6*, e16685. http://dx.doi.org/10.1371/journal.pone.0016685.

Science Staff. (2011). Dealing with data. Challenges and opportunities. Introduction. Science *331*, 692–693.

Shu, W., Chen, H., Bo, X., and Wang, S. (2011). Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. Nucleic Acids Res. *39*, 7428–7443.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Sims, R.J., 3rd, Millhouse, S., Chen, C.-F., Lewis, B.A., Erdjument-Bromage, H., Tempst, P., Manley, J.L., and Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol. Cell *28*, 665–676.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol. Cell *36*, 245–254.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

The ENCODE Project Consortium. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046. http://dx.doi.org/10.1371/journal.pbio.1001046.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16*, 996–1001.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46.

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods *37*, 376–386.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. *18*, 1051–1063.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008a). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., and Zhao, K. (2008b). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Wang, J., Lunyak, V.V., and Jordan, I.K. (2012). Genome-wide prediction and analysis of human chromatin boundary elements. Nucleic Acids Res. *40*, 511–529.

Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930–D934.

Watson, J.D., and Crick, F.H. (1953a). Genetical implications of the structure of deoxyribonucleic acid. Nature *171*, 964–967.

Watson, J.D., and Crick, F.H. (1953b). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J.S., Kristiansen, K., Krogh, A., Wang, J., and Porse, B.T. (2012). Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome Biol. *13*, R35. http://dx.doi.org/10.1186/gb-2012-13-5-r35.

Wilkins, M.H., Stokes, A.R., and Wilson, H.R. (1953). Molecular structure of deoxypentose nucleic acids. Nature *171*, 738–740.

Wulff, B.-E., Sakurai, M., and Nishikura, K. (2011). Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. Nat. Rev. Genet. *12*, 81–85.

Yang, Y., Zhan, L., Zhang, W., Sun, F., Wang, W., Tian, N., Bi, J., Wang, H., Shi, D., Jiang, Y., et al. (2011). RNA secondary structure in mutually exclusive splicing. Nat. Struct. Mol. Biol. *18*, 159–168.

Zaranek, A.W., Levanon, E.Y., Zecharia, T., Clegg, T., and Church, G.M. (2010). A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing. PLoS Genet. *6*, e1000954. http://dx.doi.org/10.1371/journal.pgen.1000954.