## Resource

# ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4] Florencia Pauli,[5] Serafim Batzoglou,[3] Bradley E. Bernstein,[6] Peter Bickel,[7] James B. Brown,[7] Philip Cayting,[1] Yiwen Chen,[8] Gilberto DeSalvo,[2] Charles Epstein,[6] Katherine I. Fisher-Aylor,[2] Ghia Euskirchen,[1] Mark Gerstein,[9] Jason Gertz,[5] Alexander J. Hartemink,[10] Michael M. Hoffman,[11] Vishwanath R. Iyer,[12] Youngsook L. Jung,[13,14] Subhradip Karmakar,[15] Manolis Kellis,[4] Peter V. Kharchenko,[12] Qunhua Li,[16] Tao Liu,[8] X. Shirley Liu,[8] Lijia Ma,[15] Aleksandar Milosavljevic,[17] Richard M. Myers,[5] Peter J. Park,[13,14] Michael J. Pazin,[18] Marc D. Perry,[19] Debasish Raha,[20] Timothy E. Reddy,[5,27] Joel Rozowsky,[9] Noam Shoresh,[6] Arend Sidow,[1,21] Matthew Slattery,[15] John A. Stamatoyannopoulos,[11,22] Michael Y. Tolstorukov,[13,14] Kevin P. White,[15] Simon Xi,[23] Peggy J. Farnham,[24,28] Jason D. Lieb,[25,28] Barbara J. Wold,[2,28] and Michael Snyder[1,28]

[1–25][Author affiliations appear at the end of the paper.]

Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) has become a valuable and widely used approach for mapping the genomic location of transcription-factor binding and histone modifications in living cells. Despite its widespread use, there are considerable differences in how these experiments are conducted, how the results are scored and evaluated for quality, and how the data and metadata are archived for public use. These practices affect the quality and utility of any global ChIP experiment. Through our experience in performing ChIP-seq experiments, the ENCODE and modENCODE consortia have developed a set of working standards and guidelines for ChIP experiments that are updated routinely. The current guidelines address antibody validation, experimental replication, sequencing depth, data and metadata reporting, and data quality assessment. We discuss how ChIP quality, assessed in these ways, affects different uses of ChIP-seq data. All data sets used in the analysis have been deposited for public viewing and downloading at the ENCODE (http://encodeproject.org/ENCODE/) and modENCODE (http://www.modencode.org/) portals.

[Supplemental material is available for this article.]

Methods for mapping transcription-factor occupancy across the genome by chromatin immunoprecipitation (ChIP) were developed more than a decade ago (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Horak and Snyder 2002; Weinmann et al. 2002). In ChIP assays, a transcription factor, cofactor, or other chromatin protein of interest is enriched by immunoprecipitation from cross-linked cells, along with its associated DNA. Genomic DNA sites enriched in this manner were initially identified by DNA hybridization to a microarray (ChIP-chip) (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Horak and Snyder 2002; Weinmann et al. 2002),

and more recently by DNA sequencing (ChIP-seq) (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007). ChIP-seq has now been widely used for many transcription factors, histone modifications, chromatin modifying complexes, and other chromatin-associated proteins in a wide variety of organisms. There is, however, much diversity in the way ChIP-seq experiments are designed, executed, scored, and reported. The resulting variability and data quality issues affect not only primary measurements, but also the ability to compare data from multiple studies or to perform integrative analyses across multiple data-types.

The ENCODE and modENCODE consortia have performed more than a thousand individual ChIP-seq experiments for more than 140 different factors and histone modifications in more than 100 cell types in four different organisms (*D. melanogaster*, *C. elegans*, mouse, and human), using multiple independent data production and processing pipelines (The ENCODE Project Consortium 2004, 2011; Celniker et al. 2009). During this work, we developed guidelines, practices, and quality metrics that are applied to all ChIP-seq work done by the Consortium (Park 2009). Here we describe these, together with supporting data and illus-

trative examples. We emphasize issues common to all ChIP-seq studies: immunoprecipitation specificity and quality, impact of DNA sequencing depth, scoring and evaluation of data sets, appropriate control experiments, biological replication, and data reporting.
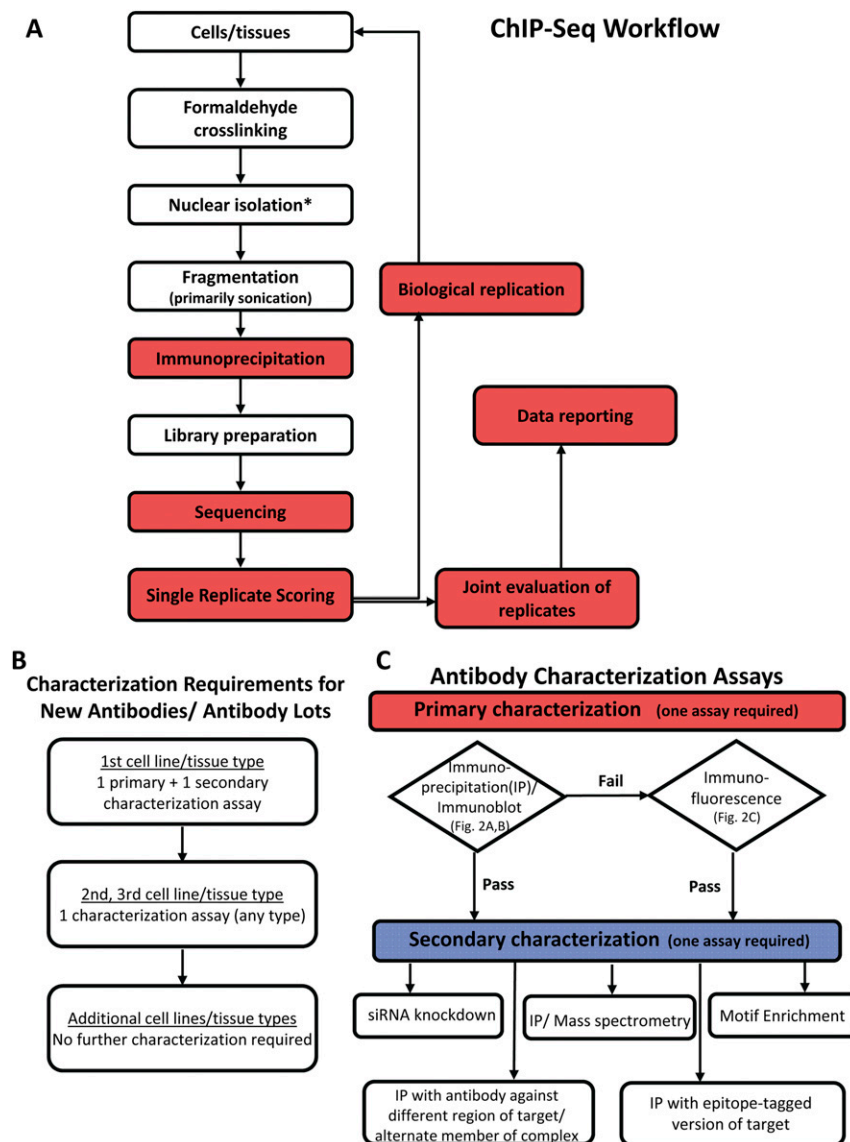
## ChIP overview

The goals of a genome-wide ChIP experiment are to map the binding sites of a target protein with maximal signal-to-noise ratio and completeness across the genome. The basic ChIP-seq procedure is outlined in Figure 1A, and detailed protocols (and data) from our two consortia can be obtained from the ENCODE and modENCODE production groups listed at the UCSC Genome Browser: http://

encodeproject.org/ENCODE/ and http://www.modencode.org/, respectively. Cells or tissues are treated with a chemical agent, usually formaldehyde, to cross-link proteins covalently to DNA. This is followed by cell disruption and sonication, or in some cases, enzymatic digestion, to shear the chromatin to a target size of 100–300 bp (Ren et al. 2000; Iyer et al. 2001). The protein of interest (transcription factor, modified histone, RNA polymerase, etc.) with its bound DNA is then enriched relative to the starting chromatin by purification with an antibody specific for the factor. Alternatively, cell lines expressing an epitope-tagged factor can be generated and the fusion protein immunoprecipitated via the epitope tag.

After immuno-enrichment, cross-links are reversed, and the enriched DNA is purified and prepared for analysis. In ChIP-chip, the DNA is fluorescently labeled and hybridized to a DNA microarray, along with differentially labeled reference DNA (Ren et al. 2000; Iyer et al. 2001). In ChIP-seq, the DNA is analyzed by high-throughput DNA sequencing. The ENCODE Consortium chose ChIP-seq for human and mouse experiments because it permits comprehensive coverage of large genomes and increases site resolution ( Johnson et al. 2007; Robertson et al. 2007). For organisms with small genomes, the modENCODE Consortium has used both ChIP-chip and ChIP-seq, as modern arrays can provide high-resolution coverage of small genomes (Gerstein et al. 2010; Roy et al. 2010). In all formats, we identified putatively enriched genomic regions by comparing ChIP signals in the experimental sample with a similarly processed reference sample prepared from appropriate control chromatin or a control immunoprecipitation.

Different protein classes have distinct modes of interaction with the genome that necessitate different analytical approaches (Pepke et al. 2009):

1. *Point-source* factors and certain chromatin modifications are localized at specific positions that generate highly localized ChIP-seq signals. This class includes most sequence-specific transcription factors, their cofactors, and, with some caveats, transcription start site or enhancer-associated histone marks. These comprise the majority of ENCODE and modENCODE determinations and are therefore the primary focus of this work.

2. *Broad-source* factors are associated with large genomic domains. Examples include certain chromatin marks (H3K9me3, H3K36me3, etc.) and chromatin proteins associated with transcriptional elongation or repression (e.g., ZNF217) (Krig et al. 2007).

3. *Mixed-source* factors can bind in point-source fashion to some locations of



**Figure 1.** Overview of ChIP-seq workflow and antibody characterization procedures. (*A*) Steps for which specific ENCODE guidelines are presented in this document are indicated in red. For other steps, standard ENCODE protocols exist that should be validated and optimized for each new cell line/tissue type or sonicator. (*) A commonly used but optional step. (*B*) Flowchart for characterization of new antibodies or antibody lots. (*C*) Flowchart for use of antibody characterization assays.

the genome, but form broader domains of binding in others. RNA polymerase II, as well as some chromatin modifying proteins (e.g., SUZ12) behave in this way (Squazzo et al. 2006).

Below, we report our experience with ChIP-seq experimental design, execution, and quality assessment. We offer specific recommendations, based on current experience, as summaries in boxes.

## ChIP-seq experimental design considerations

### Antibody and immunoprecipitation specificity

The quality of a ChIP experiment is governed by the specificity of the antibody and the degree of enrichment achieved in the affinity precipitation step. The majority of ENCODE/modENCODE ChIP experiments in human cells and in *Drosophila* embryos were performed with antibodies directed against individual factors and histone modifications. A total of 145 polyclonal and 43 monoclonal antibodies had been used to successfully generate ChIP-seq data as of October 2011.

Antibody deficiencies are of two main types: poor reactivity against the intended target and/or cross-reactivity with other DNA-associated proteins. For these reasons, we have developed a set of working standards and reporting guidelines designed to provide measures of confidence that the reagent recognizes the antigen of interest with minimal cross-reactivity toward other chromosomal proteins. Widely accessible methods for measuring antibody specificity and sensitivity range from semiquantitative to qualitative, and each can have noise and interpretation issues. We therefore emphasize reporting of antibody characterization data so that users of the ChIP data, or the reagent itself, can make informed judgments. We also recognize that a successful experiment can be performed with reagents that fail to strictly comply with these guidelines. For example, cross-reacting proteins detected in an immunoblot assay might not interfere in ChIP, because the protein is not attached to chromatin. Secondary tests of diverse types can help to provide confidence concerning the acceptability of an antibody that fails an initial assessment.

Two tests, a primary and a secondary test, are used to characterize each monoclonal antibody or different lots of the same polyclonal antibody. The ordering of the primary and secondary tests are influenced by the effort required to execute each, with the primary assay being easier to perform on large numbers of antibodies. The tests differ for antibodies against transcription factors vs. those against histone modifications. A detailed description of the tests is provided in Box 1, and a typical workflow is presented in Figure 2, B and C. For transcription-factor antigens, we adopted the immunoblot as our primary assay, with immunostaining as the alternative. The former can give more information about cross-reacting material or multiple isoforms; the latter is typically less sensitive, but provides information about nuclear location. Examples of antibodies that pass and fail these tests are shown in Figure 2A.

Our consortia also include one of five criteria as a secondary characterization: (1) factor "knockdown" by mutation or RNAi, (2) independent ChIP experiments using antibodies against more than one epitope on a protein or against different members of the same complex, (3) immunoprecipitation using epitope-tagged constructs, (4) affinity enrichment followed by mass spectrometry, or (5) binding-site motif analysis. Motif enrichment is the easiest assay to perform, but requires pre-existing information about the

sequences to which a protein binds and assumes that the motif is uniquely recognized in a given cell source by the factor of interest. ChIP with a second antibody or against an epitope-tagged construct and siRNA experiments coupled with ChIP provide independent evidence that the target sites are bound by the factor of interest. We found that mass spectrometry is particularly useful for cases where multiple or unexpected bands are observed on an immunoblot and the presence of spliced isoforms, post-translational modification, or degradation is suspected. Additionally, it can precisely identify potential alternate sources of ChIP signal, often with novel biological implications, which can be tested by additional ChIP experiments. Due to the significant effort and expense required to perform these assays, our standard for the consortia requires only one secondary assay. We found that ~20% (44 of 227) of the tested commercially available antibodies against transcription factors meet these characterization guidelines and also function in ChIP-seq assays.

To date, 55% of consortia antibodies have been submitted with mass spectrometry data, 28% with ChIP data using a second antibody, epitope tag, or alternate member of a known complex, 10% with data from motif analysis (this standard has only been used by ENCODE for 1 yr), and 7% with siRNA knockdown data. A summary of motif detection for all data sets is in preparation (P Kheradpour and M Kellis, in prep.).

Validating histone modification antibodies involves multiple issues (Egelhofer et al. 2011): (1) specificity with respect to other nuclear/chromatin proteins, (2) specificity with respect to unmodified histones and off-target modified histone residues (e.g., H3K9me vs. H3K27me), (3) specificity with respect to mono-, di-, and trimethylation at the same residue (e.g., H3K9me1, H3K9me2, and H3K9me3), and (4) lot-to-lot variation. For all consortia histone measurements, we set the standard that immunoblot analysis and one of the following secondary criteria are applied: Peptide-binding tests (dot blots), mass spectrometry, immunoreactivity analysis in cell lines containing knockdowns of a relevant histone modification enzyme or mutants histones, or genome annotation enrichment. The details of these standards are in Box 1.

### Immunoprecipitation using epitope tagged constructs

Given the challenges in obtaining antibodies for suitable ChIP, an attractive alternative is to tag the factor with an exogenous epitope and immunoprecipitate with a well-characterized monoclonal reagent specific for the tag. Epitope-tagging addresses the problems of antibody variation and cross-reaction with different members of multigene families by using a highly specific reagent that can be used for many different factors. However, this introduces concerns about expression levels and whether tagging alters the activity of the factor. The level of expression is typically addressed by using large clones (usually fosmids and BACs) carrying as much regulatory information as possible to make the level of expression nearly physiological (Poser et al. 2008; Hua et al. 2009). Higher expression is known to result in occupancy of sites not necessarily occupied at physiological levels (DeKoter and Singh 2000; Fernandez et al. 2003). In ENCODE/modENCODE, tagged factors have been used most extensively thus far for *C. elegans* studies, where factors have been tagged with GFP and shown to complement null mutants; six of six tested to date have been found to complement (Zhong et al. 2010; V Reinke, unpubl.). In some cases, information regarding expression is not available and expression from an exogenous promoter has been used (P Farnham, unpubl.).

---

**Box 1:** ENCODE guidelines for antibody and immunoprecipitation characterization

## Characterization of antibodies directed against transcription factors

Antibodies directed against transcription factors must be characterized using both a primary and secondary characterization; characterizations must be repeated for each new antibody or antibody lot number that is used for ChIP-seq (Fig. 1B,C).

### Primary mode of characterization

Antibodies are characterized by one of two primary methods, immunoblot analysis, or immunofluorescence.

*Immunoblot analyses*

Immunoblot analyses are performed on protein lysates from either whole-cell extracts, nuclear extracts, chromatin preparations, or immunoprecipitated material (before proceeding to ChIP assays, it is helpful to demonstrate that the protein of interest can be efficiently immunoprecipitated from a nuclear extract, see Fig. 2B). We use the guideline that the primary reactive band should contain at least 50% of the signal observed on the blot. Ideally, this band should correspond to the size expected for the protein of interest (Fig. 2A). However, the electrophoretic mobility of many factors can deviate significantly from the expected size due to modifications, isoform differences, or intrinsic properties of the factor. Therefore, antibodies for which the main band differs from the expected size by >20% or for which multiple bands are seen (such that no band represents >50% of the signal) can be used under certain circumstances. In these cases, further criteria must be met, such as (1) the unexpected mobility must have been properly documented in published studies using the same antibody lot, (2) the signal in the band(s) is reduced by siRNA knockdown or mutation, or (3) the factor can be identified in all band(s) by mass spectrometry.

*Immunofluorescence*

Some antibodies that work well for ChIP do not work well in immunoblots. If immunoblot analysis is not successful, immunofluorescence can be used as an alternative method. Staining should be of the expected pattern (e.g., nuclear and only in cell types or under specific growth conditions that express the factor) (Fig. 2C). Because immunofluorescence does not provide evidence that the antibody detects only one protein, this validation method should be combined with a method that reduces the level of the protein, such as siRNA- or shRNA-mediated knockdown, or used with a knockout cell line or organism (see below).

### Secondary mode of characterization

In addition to the primary mode of characterization, the consortia performs at least one of the following five assays as an additional secondary test:

*Knockdown or knockout of the target protein*

Immunoblots or immunoprecipitations are performed in duplicate using extracts from siRNA or shRNA knockdowns or from knockout mutant cell lines or organisms. We use the guideline that the primary immunoblot (or immunofluorescence) signal, along with additional immunoreactive bands, should be reduced to no more than 30% of the original signal and any signal remaining after genetic mutation, RNAi, or siRNA is noted. As an alternative, knockdown can also be measured with ChIP experiments. ENCODE data can be submitted if reduction of ChIP-chip or ChIP-seq signals by >50% relative to control is observed. A suitable control knockdown (e.g. using "scrambled" siRNA sequences) should also be performed and the data should be submitted; reduction of signal should not be observed in the control knockdown data set. The methodology used for binding-region signal normalization (for instance, normalization against total read counts or using values from reference peaks quantified by qPCR under all experimental conditions) should also be reported.

*Immunoprecipitation followed by mass spectrometry*

All immunoreactive bands identified by immunoblot analysis are analyzed (Fig. 2D). ENCODE passes such analyses if the protein of interest is identified in such bands; if additional chromosomal proteins are identified in an immunoreactive band, the Consortium accepts the experiment as long as they are present at lower prevalence than the desired protein (as measured by peptide counts or other methods) or can be demonstrated to arise from nonspecific immunoprecipitation (e.g., also present in a control immunoprecipitation). All proteins identified by mass spectrometry and the number of peptide counts for each are reported.

*Immunoprecipitation with multiple antibodies against different parts of the target protein or members of the same complex*

Different antibodies against different parts of the same protein or other members of a known protein complex can be used in analyzing the specificity of antibodies. In the ENCODE Consortium, results of the different ChIP experiments are compared and significant overlap of enriched loci is expected (ChIP-seq experiments are compared using the IDR-based standards in Box 3). Note that for different proteins that are members of a complex, there may be some functions that are independent of one another. Thus, the targets lists for two different proteins may not entirely overlap. In this case, specific evidence about limited overlap of binding specificity in the literature is presented to justify the significance of the overlap observed between data sets for the factors in question.

*Immunoprecipitation with an epitope-tagged version of the protein*

An epitope-tagged version of the target protein may be used, preferably expressed from the endogenous gene promoter. ENCODE conducts and analyzes such experiments as described above for the use of multiple antibodies.

*Motif enrichment*

For transcription factors, if a factor has a well-characterized motif derived from in vitro binding studies or another justifiable method, and if either no paralogs are expressed in the cell lines being analyzed or if the antibody is raised to a unique region of the factor, motif enrichment can be used for validation. Motif analysis can be performed using a defined set of high-quality peaks (a 0.01 IDR threshold is used), and for ENCODE data to be submitted, motifs should be enriched at least fourfold compared with all accessible regions (e.g., DNase hypersensitive regions) and present in >10% of analyzed peaks. Analysis of data sets deposited as of January 2011 identified data sets that meet these standards for 49 of 85 factors (Fig. 2E). We note that due to differences in transcription-factor recruitment mechanisms, failure of a data set to meet the motif enrichment threshold does not necessarily indicate poor quality data.

*(continued)*

---

**Box 1:** *Continued*

*Other considerations*

1. For antibodies directed against members of a multigene family, the best practice is to prepare or obtain antibodies that recognize protein regions unique to individual family members. For an ENCODE validated antibody, any potential cross-reaction is noted when reporting data collected using that antibody.
2. For antibodies that have been previously characterized for one cell type, ENCODE has used only one validation method (such as immunoblot analysis) when the antibody is used to perform ChIP in a new cell type or organism. If an antibody has been validated in at least three different cell types, we do not require further validation for ChIP-seq experiments with additional cell types for ENCODE submission. Similarly, for whole organisms, if the antibody has been characterized in three growth stages, no further characterization is required.
3. If antibodies derived from the same lot are used by different groups in ENCODE, they only need to be characterized once. However, antibodies from different lots of the same catalog number are characterized as if they were new antibodies.

**Epitope-tagged proteins**

Epitope-tagged factors are introduced into cells by transfection of an expression construct. To help ensure that ChIP-seq results obtained using the tagged factor are comparable to those expected for the endogenous factor, ENCODE uses the criteria that tagged factors are expressed at a comparable amount to the endogenous factor. This is usually achieved by cloning into a low-copy number vector and using the natural promoter to drive expression. If the tagged protein is expressed from a heterologous promoter, data comparing expression levels of the tagged and endogenous proteins (i.e., immunoblots to measure protein levels or qPCR to measure RNA levels) are needed. There are special cases in which ChIP cannot be obtained at endogenous protein levels, and here, elevated expression can provide useful information. ENCODE's recommended control for epitope-tagged measurements is an immunoprecipitation using the same antibody against the epitope tag in otherwise identical cells that do not express the tagged factor.

**Histone modifications**

For ENCODE data to be submitted, all commercial histone antibodies are validated by at least two independent methods, as described below, and new lots of antibody are analyzed independently. These validations are performed by the ENCODE laboratory performing the ChIP-seq or by the antibody supplier, but only if the supplier provides data for the specific lot of antibody. The tests need only be performed once for each antibody lot.

**Primary test**

All antibodies used in ENCODE ChIP experiments are checked for reactivity with nonhistone proteins and with unmodified histones by performing immunoblot analysis on total nuclear extract and recombinant histones. To enable visual quantification of reactivity, a concentration series of both extract and recombinant histones are analyzed using recombinant histone levels that are comparable to those of the target histone in nuclear extract. Since cross-reactivity may vary between species, this test is performed using nuclear extracts from each species to be studied by ChIP. To pass the criteria for submission in ENCODE, the specific histone band should constitute at least 50% of the signal in western blots of nuclear extract, show at least 10-fold enrichment relative to any other single band, and show at least 10-fold enriched signal relative to unmodified histone.

**Secondary test**

In addition to the primary test, antibody specificity is verified by at least one additional test. The pros and cons of each test are described. The first two are the most commonly used.

*Peptide binding tests*

Peptide binding and peptide competition assays provide a fast method to initially evaluate the specificity and relative binding strength of antibodies to histone tails with different modifications (e.g., H3K9 or H3K27 and me1, me2, and me3 levels of methylation). A potential drawback is that antibodies may differ in their binding specificity toward histone tail peptides in vitro versus toward full-length histones in the context of chromatin in IP experiments. Nevertheless, observing at least a 10-fold enriched binding signal for the modification of interest relative to other modifications provides confidence in the antibody specificity. For these assays, histone tail peptides with particular modifications can be purchased commercially. Alternatively, peptide binding and/or competition assays using the same lot of antibody can be performed by the company from which the antibody is purchased.

*Mass spectrometry*

For antibodies generated against related and historically problematic modifications, the ability of the antibody to effectively distinguish between similar histone marks (e.g., H3K9me and H3K27me) and between different levels of methylation (e.g., H3K9me1, H3K9me2, and H3K9me3) can be tested by mass spectrometry analysis of material immunoprecipitated from histone preparations. For ENCODE data, the target modification constitutes at least 80% of the immunoprecipitated histone signal. This test may often not be successful because IP for one modification can simultaneously isolate coassociated histones with other modifications. Thus, only a positive result (i.e., a specific modification) is interpretable.

*Mutants defective in modifying histones*

Strains or cell lines harboring knockouts or catalytically inactive mutants of enzymes responsible for particular histone modifications offer the opportunity to test antibody specificity. Such mutants exist for *S. cerevisiae*, *S. pombe*, *Drosophila*, *C. elegans* and can, in cases where the modifying enzymes are nonredundant, be created for mammalian cells. For submitted ENCODE/modENCODE data, antibody signal is reduced to below 10% of wild-type signal in mutant samples, compared with wild type. RNAi or siRNA depletion of histone modifying activity may be substituted for mutants. Mutant or RNAi or siRNA reduction of signal can be assayed by immunoblot analysis or by immunofluorescence staining. Mutant/RNAi/siRNA tests usually do not allow testing antibodies for the ability to discriminate between mono-, di-, and trimethylation. In cases where more than one enzyme modifies the same residue (e.g., H3K9 methylation in *Drosophila*), double mutants or RNAi may be required. Replicates of this test are encouraged but not required for ENCODE/modENCODE data to be submitted. However, positive controls showing that the antibody works on

*(continued)*

---

**Box 1:** *Continued*

wild-type samples processed in parallel, and positive controls showing that the mutant extract is amenable to the assay employed are included for data to be submitted.

*Mutant histones*
Mutant histones (e.g., histone H3 with Lys4 mutated to Arg or Ala) expressed in yeast provide another avenue to test specificity by immunoblot analysis or even by ChIP. When analyzing a strain containing a mutated histone that cannot be modified, we expect at least a 10-fold reduction in immunoblot or IP signal relative to wild-type histone preparations. Mutant histone tests cannot distinguish whether antibodies discriminate between mono, di, and trimethylation.

*Annotation enrichment*
Enrichment at annotated features (e.g., transcription start sites) can be used as a validation criterion for certain chromatin-associated modifications and proteins. If a well-characterized modification (e.g. H3K4me3) is analyzed, the observed localization to annotations are expected to be similar to that of known overlap standards derived from the literature or existing ChIP-seq data sets (for point source peaks, overlap with known annotations can be assessed using the IDR guidelines in Box 3).

*Use of two different antibodies*
Even if antibodies pass the specificity tests described above, observing similar ChIP results with two independent antibodies provides added confidence. We therefore aspire to obtain ChIP-seq data from two independent antibodies whenever possible, providing statistical comparisons of the results and presenting the intersection of the peak sets obtained with the two antibodies. The reasons for a significant discordance can be either biological or technical, and merit further dissection.

---

## Replication, sequencing depth, library complexity, and site discovery

Biological replicate experiments from independent cell cultures, embryo pools, or tissue samples are used to assess reproducibility. Initial RNA polymerase II ChIP-seq experiments showed that more than two replicates did not significantly improve site discovery (Rozowsky et al. 2009). Thus, the ENCODE Consortium set as our standard that all ChIP measurements would be performed on two independent biological replicates. The irreproducible discovery rate (IDR) analysis methodology (Li et al. 2011) is now used to assess replicate agreement and set thresholds (discussed further below). For experiments with poor values for quality metrics described in Section III, additional replicate(s) have been generated.
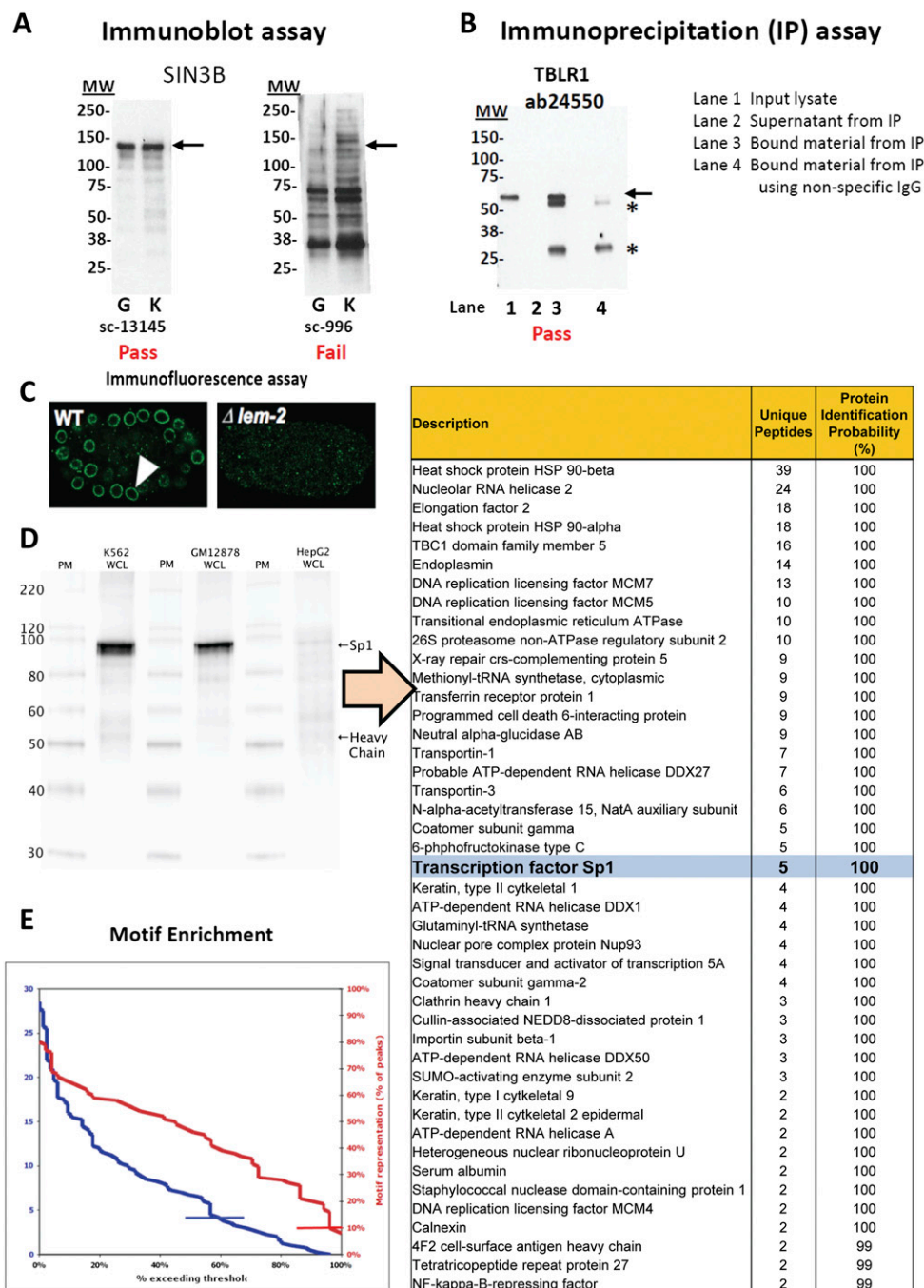
For a typical point-source DNA-binding factor, the number of ChIP-seq positive sites identified typically increases with the number of sequenced reads (Myers et al. 2011). This result is expected, as studies of numerous factors by ENCODE and by other groups have repeatedly found a continuum of ChIP signal strength, rather than a sharply bounded and discrete set of positive sites (Rozowsky et al. 2009; Myers et al. 2011). Weaker sites can be detected with greater confidence in larger data sets because of the increased statistical power afforded by more reads. Figure 3 shows an analysis of peak calls for 11 human ENCODE ChIP-seq data sets for which deep-sequence data (30–100 million mapped reads) were obtained. Clear saturation of peak counts was observed for one factor with few binding sites, but counts continued to increase at varying rates for all other factors, including a case in which >150,000 peaks were called using 100 million mapped reads. Examination of peak signals reveals that the signal enrichments consistently plateau at greater sequencing depths. At 20 million mapped reads, which we currently use as a minimum for all ENCODE ChIP experiments for point-source transcription factors (Box 2), five- to 13-fold median enrichments are the norm; new peaks identified after 20 million reads give enrichments that are ~20% of the enrichment of the strongest peaks (Fig. 3C). Interestingly, many additional peaks, with enrichment values of three- to sevenfold, can still be found by sequencing to much greater depths. It is likely that many of these regions correspond to low-affinity sites and/or regions of open chromatin that bind TFs less specifically.

The relationship of ChIP signal strength to biological regulatory activity is a current area of active investigation. The biological activity of known enhancers, defined in the literature independently of ChIP data, is distributed quite broadly relative to ChIP-seq signal strength (Ozdemir et al. 2011; G DeSalvo, G Marinov, K Fisher, A Kirilusha, A Mortazavi, B Williams, and B Wold, in prep.). Some highly active transcriptional enhancers reproducibly display modest ChIP signals (Fig. 4B). This means that one cannot a priori set a specific target threshold for ChIP peak number or ChIP signal strength that will assure inclusion of all functional sites (see Discussion). Therefore, a practical goal is to maximize site discovery by optimizing immunoprecipitation and sequencing deeply, within reasonable expense constraints. For point-source factors in mammalian cells, a minimum of 10 million uniquely mapped reads are used by ENCODE for each biological replicate (providing a minimum of 20 million uniquely mapped reads per factor); for worms and flies a minimum of 2 million uniquely mapped reads per replicate is used. For broad areas of enrichment, the appropriate number of uniquely mapped reads is currently under investigation, but at least 20 million uniquely mapped reads per replicate for mammalian cells and 5 million uniquely mapped reads per replicate for worms and flies is currently being produced for most experiments.

Site discovery and reproducibility are also affected by the complexity of a ChIP-seq sequencing library (Fig. 4A). We define library complexity operationally as the fraction of DNA fragments that are nonredundant. With increased depth of sequencing of a library, a point is eventually reached where the complexity will be exhausted and the same PCR-amplified DNA fragments will be sequenced repeatedly. Low library complexity can occur when very low amounts of DNA are isolated during the IP or due to problems with library construction.

A useful complexity metric is the fraction of nonredundant mapped reads in a data set (nonredundant fraction or NRF), which we define as the ratio between the number of positions in the genome that uniquely mappable reads map to and the total number of uniquely mappable reads; it is similar to a recently published redundancy metric (Heinz et al. 2010). NRF decreases with se-

**A** Immunoblot assay

SIN3B

G K
sc-13145
**Pass**

G K
sc-996
**Fail**

**B** Immunoprecipitation (IP) assay

TBLR1
ab24550

Lane 1  Input lysate
Lane 2  Supernatant from IP
Lane 3  Bound material from IP
Lane 4  Bound material from IP
using non-specific IgG

Lane 1 2 3 4
**Pass**

**C** Immunofluorescence assay

WT    Δ lem-2

**D**

K562 WCL   GM12878 WCL   HepG2 WCL
←Sp1
←Heavy Chain

| Description | Unique Peptides | Protein Identification Probability (%) |
|---|---|---|
| Heat shock protein HSP 90-beta | 39 | 100 |
| Nucleolar RNA helicase 2 | 24 | 100 |
| Elongation factor 2 | 18 | 100 |
| Heat shock protein HSP 90-alpha | 18 | 100 |
| TBC1 domain family member 5 | 16 | 100 |
| Endoplasmin | 14 | 100 |
| DNA replication licensing factor MCM7 | 13 | 100 |
| DNA replication licensing factor MCM5 | 10 | 100 |
| Transitional endoplasmic reticulum ATPase | 10 | 100 |
| 26S proteasome non-ATPase regulatory subunit 2 | 10 | 100 |
| X-ray repair crs-complementing protein 5 | 9 | 100 |
| Methionyl-tRNA synthetase, cytoplasmic | 9 | 100 |
| Transferrin receptor protein 1 | 9 | 100 |
| Programmed cell death 6-interacting protein | 9 | 100 |
| Neutral alpha-glucidase AB | 9 | 100 |
| Transportin-1 | 7 | 100 |
| Probable ATP-dependent RNA helicase DDX27 | 7 | 100 |
| Transportin-3 | 6 | 100 |
| N-alpha-acetyltransferase 15, NatA auxiliary subunit | 6 | 100 |
| Coatomer subunit gamma | 5 | 100 |
| 6-phphofructokinase type C | 5 | 100 |
| **Transcription factor Sp1** | **5** | **100** |
| Keratin, type II cytkeletal 1 | 4 | 100 |
| ATP-dependent RNA helicase DDX1 | 4 | 100 |
| Glutaminyl-tRNA synthetase | 4 | 100 |
| Nuclear pore complex protein Nup93 | 4 | 100 |
| Signal transducer and activator of transcription 5A | 4 | 100 |
| Coatomer subunit gamma-2 | 4 | 100 |
| Clathrin heavy chain 1 | 3 | 100 |
| Cullin-associated NEDD8-dissociated protein 1 | 3 | 100 |
| Importin subunit beta-1 | 3 | 100 |
| ATP-dependent RNA helicase DDX50 | 3 | 100 |
| SUMO-activating enzyme subunit 2 | 3 | 100 |
| Keratin, type I cytkeletal 9 | 2 | 100 |
| Keratin, type II cytkeletal 2 epidermal | 2 | 100 |
| ATP-dependent RNA helicase A | 2 | 100 |
| Heterogeneous nuclear ribonucleoprotein U | 2 | 100 |
| Serum albumin | 2 | 100 |
| Staphylococcal nuclease domain-containing protein 1 | 2 | 100 |
| DNA replication licensing factor MCM4 | 2 | 100 |
| Calnexin | 2 | 100 |
| 4F2 cell-surface antigen heavy chain | 2 | 99 |
| Tetratricopeptide repeat protein 27 | 2 | 99 |
| NF-kappa-B-repressing factor | 2 | 99 |

**E** Motif Enrichment

% exceeding threshold
Motif representation (% of peaks)

**Figure 2.** Representative results from antibody characterization assays. (*A*) Immunoblot analyses of antibodies against SIN3B that (*left*) pass quality control (Santa Cruz sc13145) and (*right*) fail quality control (Santa Cruz sc996). Lanes contain nuclear extract from GM12878 cells (G) and K562 cells (K). Arrows indicate band of expected size of 133 kDa. Molecular weights (MW) are in kilodaltons. (*B*) Immunoblot analysis of an antibody against TBLR1 (Abcam ab24550) that passes quality control and can be used for immunoprecipitation. Immunoprecipitations (IPs) were performed from nuclear lysates of K562 cells. Arrow indicates band of expected size (56 kDa) that is detected in the input lysate (lane *1*) and is efficiently (cf. lanes *3* and *2*) and specifically (absent in lane *4*) immunoprecipitated. (*) IgG light and heavy chains. (*C*) Immunofluorescence analyses of antibodies that pass (*left*) and fail (*right*) quality control. (*D*) Immunoprecipitation/mass spectrometry analysis of an antibody against SP1 (Santa Cruz sc-17824). Whole-cell lysates (WCL) of K562, GM12878, and HepG2 were immunoprecipitated, and a band of expected size (~106 kDa) was detected on a Western blot with SP1 primary antibody. The immunoprecipitation was repeated in K562 WCL, separated on a gel, stained with Coomassie Blue, and the band previously detected on the Western blot was excised and analyzed by mass spectrometry. Peptides were identified using MASCOT (Matrix Science) with probability-based matching at $P < 0.05$. Subsequent analysis was performed in Scaffold (Proteome Software, Inc.) at 0.0% protein FDR and 0.0% peptide FDR. SP1 protein was detected (along with common contaminants that are often obtained in control experiments) (data not shown) and is highlighted in bold and light blue. (*E*) Histogram depicting motif fold-enrichment (blue) for all transcription factors for which ENCODE ChIP-seq data is available (85 factors). Enrichments are relative to all DNase-accessible sites and were corrected for sequence bias using shuffle motifs. Motif searches were conducted with a matching stringency of 4–6. Where multiple data sets are available for a factor, the data set with the highest enrichment was counted. Data sets that meet the ENCODE standard of fourfold enrichment (indicated by blue line) were found for 60% of factors. Motif representation, as a percentage of all analyzed peaks, is shown in red for all factors for which a data set exists that exceeds the enrichment standard. A total of 96% of these data sets meet the ENCODE standard of >10% motif representation (red line). All calculations were carried out on peaks identified by IDR analysis (0.01 cut-off).
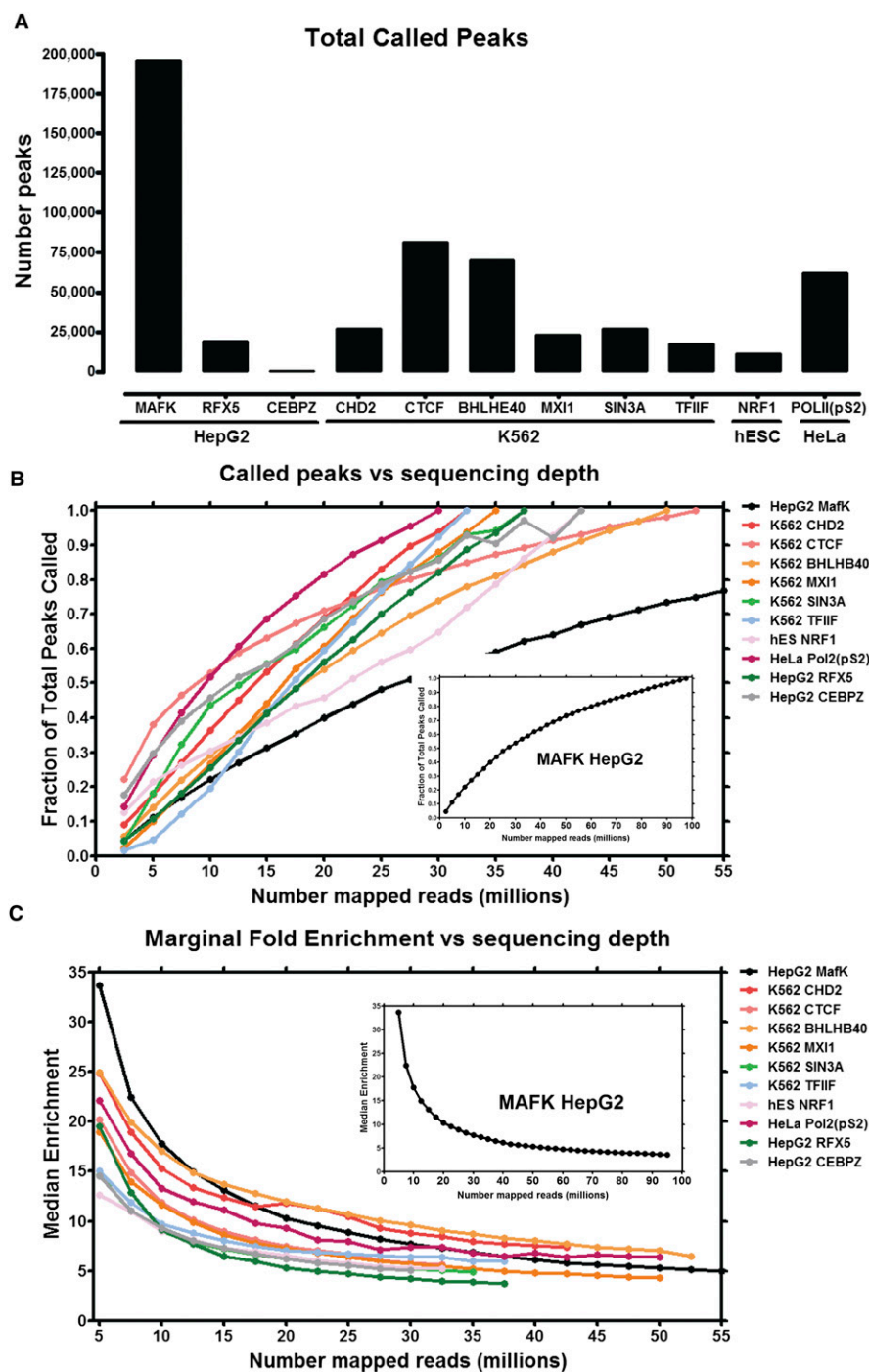
**Figure 3.** Peak counts depend on sequencing depth. (*A*) Number of peaks called with Peak-seq (0.01% FDR cut-off) for 11 ENCODE ChIP-seq data sets. (*B*) Called peak numbers for 11 ChIP-seq data sets as a function of the number of uniquely mapped reads used for peak calling. (*Inset*) Called peak data for the MAFK data set from HepG2 cells, currently the most deeply sequenced ENCODE ChIP-seq data set (displayed separately due to the significantly larger number of reads relative to the other data sets). Data sets are indicated by cell line and transcription factor (e.g., cell line HepG2, transcription factor MAFK). (*C*) Fold-enrichment for newly called peaks as a function of sequencing depth. For each incremental addition of 2.5 million uniquely mapped reads, the median fold-enrichment for newly called peaks as compared with an IgG control data set sequenced to identical depth is plotted.

quencing depth, and for point source TFs, our current target is NRF $\geq 0.8$ for 10 million (M) uniquely mapped reads (Box 2). We expect that, as sequencing technology improves and read numbers in the hundreds of millions per lane become feasible, even complex libraries from point-source factor libraries may be sequenced at depths greater than necessary. To maximize information that can be obtained for each DNA-sequencing run and to prevent oversequencing, barcoding and pooling strategies can be used (Lefebvre et al. 2010).

## Control sample

An appropriate control data set is critical for analysis of any ChIP-seq experiment because DNA breakage during sonication is not uniform. In particular, some regions of open chromatin are preferentially represented in the sonicated sample (Auerbach et al. 2009). There are also platform-specific sequencing efficiency biases that contribute to nonuniformity (Dohm et al. 2008). There are two basic methods to produce control DNA samples, each of which mitigates the effects of these issues on binding-site identification: (1) DNA is isolated from cells that have been cross-linked and fragmented under the same conditions as the immunoprecipitated DNA ("Input" DNA); and (2) a "mock" ChIP reaction is performed using a control antibody that reacts with an irrelevant, non-nuclear antigen ("IgG" control). For both types of controls, ENCODE groups sequence to a depth at least equal to, and preferably larger than, that of the ChIP sample. While the IgG control mimics a ChIP experiment more closely than does an "input" control, it is important that IgG control immunoprecipitations recover enough DNA to build a library of sufficiently high complexity to that of the experimental samples; otherwise, binding-site identifications made using this control can be significantly biased.

Regardless of the type of control used, ENCODE and modENCODE groups perform a separate control experiment for each cell line, developmental stage, and different culture condition/treatment because of known and unknown differences in ploidy, genotype, and epigenetic features that affect chromatin preparation. To serve as a valid control, we use identical protocols to build ChIP and control sequencing libraries (i.e., the same as the number of PCR amplification cycles,

Downloaded from [genome.cshlp.org](genome.cshlp.org) on September 15, 2018 - Published by [Cold Spring Harbor Laboratory Press](Cold Spring Harbor Laboratory Press)

---

**Box 2:** ChIP experimental design guidelines

**Sequencing and library complexity**

For each ChIP-seq point-source library, ENCODE's goal is to obtain ≥10 million uniquely mapping reads per replicate experiment for mammalian genomes, with a target NRF (nonredundancy fraction) ≥0.8 for 10 million reads. The corresponding objective for modENCODE point-source factors is to obtain ≥2 M uniquely mapped reads per replicate, ≥0.8 NRF. The modENCODE target for broad-source ChIP-seq in *Drosophila* is ≥5 million reads, and the ENCODE provisional target for mammalian broad-source histone marks is ≥20 million uniquely mapping reads at NRF ≥0.8. The distribution of NRF values for all ENCODE data sets is shown in Figure 7.

**Control libraries**

ENCODE generates and sequences a control ChIP library for each cell type, tissue, or embryo collection and sequences the library to the appropriate depth (i.e., at least equal to, and preferably greater than, the most deeply sequenced experimental library). If cost constraints allow, a control library should be prepared from every chromatin preparation and sonication batch, although some circumstances can justify fewer control libraries. Importantly, a new control is always performed if the culture conditions, treatments, chromatin shearing protocol, or instrumentation is significantly modified.

**Reproducibility**

Experiments are performed at least twice to ensure reproducibility. For ENCODE data to pass criteria for submission, concordance is determined from analysis using the IDR methodology (current ENCODE criteria are in Box 3), and a third replicate is performed if the standard is not reached. Cut-offs for identifying highly reproducible peaks for use in subsequent analyses can be determined by IDR (typically using a 1% threshold).

---

fragment size, etc.). Although rare in our experience, control libraries with particularly strong sonication biases have been observed and they can adversely affect peak calling (Supplemental Fig. S1). As much as possible, ENCODE/modENCODE groups also generate a separate control for each batch of sonicated samples to control for possible sonication variation.

### Peak calling

After mapping reads to the genome, peak calling software is used to identify regions of ChIP enrichment. We have used several peak calling algorithms and corresponding software packages, including SPP, PeakSeq, and MACs (Ji et al. 2008; Valouev et al. 2008; Zhang et al. 2008; Rozowsky et al. 2009). The resulting output of these algorithms generally ranks called regions by absolute signal (read number) or by computed significance of enrichment (e.g., *P*-values and false discovery rates). Because ChIP signal strength is a continuum with many more weak sites than strong ones (Fig. 4B), the composition of the final peak list depends heavily on the specific parameter settings and the algorithm used as well as the quality of the experiment itself. Thresholds that are too relaxed lead to a high proportion of false positives for each replicate, but as discussed below, subsequent analysis can strip false positives from a final joint peak determination. Different peak-calling algorithms rely on different statistical models to calculate *P*-values and false discovery rates (FDR), meaning that significance values from different software packages are not directly comparable. When using standard peak-calling thresholds, successful experiments generally identify thousands to tens of thousands of peaks for most TFs in mammalian genomes, although some exceptions are known (Frietze et al. 2010; Raha et al. 2010). In all cases, it is important to use an appropriate control experiment in peak calling.

Calling discrete regions of enrichment for *Broad-source* factors or *Mixed-source* factors is more challenging and is at an earlier stage of development. Methods to identify such regions are emerging (e.g., ZINBA [Rashid et al. 2011] [installation package at http://code.google.com/p/zinba/], Scripture [Guttman et al. 2010], and MACS2, an updated version of MACS that is specifically designed to process mixed signal types [https://github.com/taoliu/MACS]). Standards for the identification of broad enrichment regions are currently in development.

### Evaluating ChIP-seq data

The quality of individual ChIP-seq experiments varies considerably and can be especially difficult to evaluate when new antibodies are being tested or when little is known about the factor and its binding motif. The ENCODE Consortium has developed and uses metrics for several aspects of ChIP-seq quality, together with traditional site-inspection-based evaluation. When applied and interpreted as a group, these metrics and approaches provide a valuable overall assessment of experimental success and data quality.

### Browser inspection and previously known sites

A first impression about ChIP-seq quality can be obtained by local inspection of mapped sequence reads using a genome browser. Although not quantitative, this approach is very useful, especially when a known binding location can be examined; read distribution shape and signal strength relative to a control sample can provide a sense of ChIP quality. A true signal is expected to show a clear asymmetrical distribution of reads mapping to the forward and reverse strands around the midpoint (peak) of accumulated reads. This signal should be large compared with the signal of the same region from the control library. Of course it is not feasible to inspect the whole genome in this manner, and evaluating a limited number of the strongest sites may overestimate the quality of the entire data set (Supplemental Fig. S2). The genome-wide metrics discussed below provide more objective and global assessments.

### Measuring global ChIP enrichment (FRiP)

For point-source data sets, we calculate the fraction of all mapped reads that fall into peak regions identified by a peak-calling algorithm (Ji et al. 2008). Typically, a minority of reads in ChIP-seq experiments occur in significantly enriched genomic regions (i.e., peaks); the remainder of the read represents background. The fraction of reads falling within peak regions is therefore a useful and simple first-cut metric for the success of the immunoprecipitation, and is called FRiP (fraction of reads in peaks). In general, FRiP values correlate positively and linearly with the number of called regions, although there are exceptions, such as REST (also known as NRSF) and GABP, which yield a more limited number of
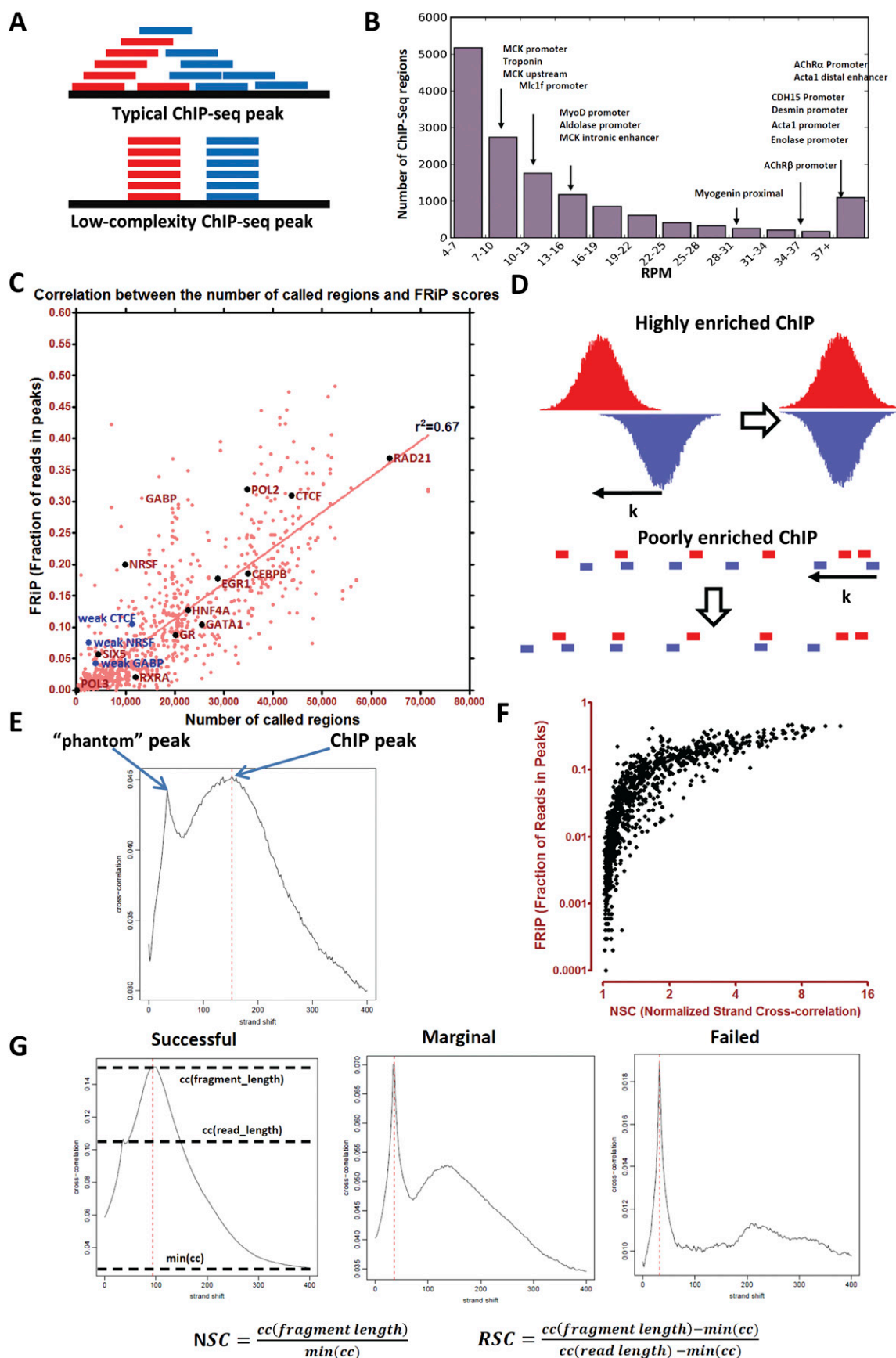
**Genome Research** 1821
www.genome.org

**Figure 4.** (Legend on next page)

$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

called regions but display very high enrichment (Fig. 4C). Most (787 of 1052) ENCODE data sets have a FRiP enrichment of 1% or more when peaks are called using MACS with default parameters. The ENCODE Consortium scrutinizes experiments in which the FRiP falls below 1%.

The 1% FRiP guideline works well when there are thousands to tens of thousands of called occupancy sites in a large mammalian genome. However, passing this threshold does not automatically mean that an experiment is successful and a FRiP below the threshold does not automatically mean failure. For example, ZNF274 and human RNA polymerase III have very few true binding sites (Frietze et al. 2010; Raha et al. 2010), and a FRiP of <1% is obtained. At the other extreme, ChIP experiments using antibody/factor pairs capable of generating very high enrichment (such as REST and GABP mentioned above) and/or binding-site numbers (CTCF, RAD21, and others) can result in FRiP scores that exceed those obtained for most factors (Fig. 5C), even for experiments that are suboptimal. Thus, FRiP is very useful for comparing results obtained with the same antibody across cell lines or with different antibodies against the same factor. FRiP is sensitive to the specifics of peak calling, including the way the algorithm delineates regions of enrichment and the parameters and thresholds used. Thus, all FRiP values that are compared should be derived from peaks uniformly called by a single algorithm and parameter set.

## Cross-correlation analysis

A very useful ChIP-seq quality metric that is independent of peak calling is strand cross-correlation. It is based on the fact that a high-quality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest, and that the sequence tag density accumulates on forward and reverse strands centered around the binding site. As illustrated in Figure 5D, these "true signal" sequence tags are positioned at a distance from the binding site center that depends on the fragment size distribution (Kharchenko et al. 2008). A control experiment, such as sequenced input DNA, lacks this pattern of shifted stranded tag densities (Supplemental Fig. S1). This has made it possible to develop a metric that quantifies fragment clustering (IP enrichment) based on the correlation between genome-wide stranded tag densities (A Kundaje, Y Jung, P Kharchenko, B Wold, A Sidow, S Batzoglou, and P Park, in prep.). It is computed as the Pearson linear correlation between the Crick strand and the Watson strand, after shifting Watson by $k$ base pairs (Fig. 5E). This typically produces two peaks when cross-correlation is plotted against the shift value: a peak of enrichment corresponding to the predominant fragment length and a peak corresponding to the read length ("phantom" peak) (Fig. 4E; Heinz et al. 2010; A Kundaje, Y Jung, P Kharchenko, B Wold, A Sidow, S Batzoglou, and P Park, in prep.).

The normalized ratio between the fragment-length cross-correlation peak and the background cross-correlation (normalized strand coefficient, NSC) and the ratio between the fragment-length peak and the read-length peak (relative strand correlation, RSC) (Fig. 4G), are strong metrics for assessing signal-to-noise ratios in a ChIP-seq experiment. High-quality ChIP-seq data sets tend to have a larger fragment-length peak compared with the read-length peak, whereas failed ones and inputs have little or no such peak (Figs. 4G, 5A,B; Fig. 7, below). In general, we observe a continuum between the two extremes, and broad-source data sets are expected to have flatter cross-correlation profiles than point-sources, even when they are of very high quality. As expected, the NSC/RSC and FRiP metrics are strongly and positively correlated for the majority of experiments (Fig. 4F). As with the other quality metrics, even high-quality data sets generated for factors with few genuine binding sites tend to produce relatively low NSCs.

These measures form the basis for one of the current quality standards for ENCODE data sets. We repeat replicates with NSC values <1.05 and RSC values <0.8 and, if additional replicates produce low values, we include a note with the reported data set (Box 3). We illustrate the application of our ChIP-seq quality metrics to a failed pair of replicates in Figure 5, A–E. Initially, two EGR1 ChIP-seq replicates were generated in the K562 cell line. Based on the cross-correlation profiles, FRiP score, and number of called regions, these replicates were flagged as marginal in quality. The experiments were repeated, with all quality control metrics improving considerably. On this basis, the superior measurements replaced the initial ones in the ENCODE database.

## Consistency of replicates: Analysis using IDR

As noted above, the modENCODE and ENCODE consortia generate two independent biological replicates, with each experiment passing the basic quality control filters. As another measure of experiment quality, we take advantage of the reproducibility information provided by the duplicates using the IDR (irreproducible discovery rate) statistic that has been developed for ChIP-seq (Li et al. 2011; discussed in detail in A Kundaje, Q Li, B Brown, J Rozowsky, A Harmanci, S Wilder, S Batzoglou, I Dunham, M Gerstein, E Birney, et al., in prep.).

Given a set of peak calls for a pair of replicate data sets, the peaks can be ranked based on a criterion of significance, such as the *P*-value, the q-value, the ChIP-to-input enrichment, or the read coverage for each peak (Fig. 6A–E). If two replicates measure the same underlying biology, the most significant peaks, which are likely to be genuine signals, are expected to have high consistency between replicates, whereas peaks with low significance, which are more likely to be noise, are expected to have low consistency. If the consistency between a pair of rank lists that contains both signif-

**Figure 4.** Criteria for assessing the quality of a ChIP-seq experiment. (*A*) Library complexity. Individual reads mapping to the plus (red) or minus strand (blue) are represented. (*B*) Distribution of functional regulatory elements with respect to the strength of the ChIP-seq signal. ChIP-seq was performed against myogenin, a major regulator of muscle differentiation, in differentiated mouse myocytes. While many extensively characterized muscle regulatory elements exhibit strong myogenin binding, a large number of known functional sites are at the low end of the binding strength continuum. (*C*) Number of called peaks vs. ChIP enrichment. Except in special cases, successful experiments identify thousands to tens of thousands of peaks for most TFs and, depending on the peak finder used, numbers in the hundreds or low thousands indicate a failure. Peaks were called using MACS with default thresholds. (*D*) Generation of a cross-correlation plot. Reads are shifted in the direction of the strand they map to by an increasing number of base pairs and the Pearson correlation between the per-position read count vectors for each strand is calculated. Read coverage as wigglegram is represented, not to the same scale in the *top* and *bottom* panels.) (*E*) Two cross-correlation peaks are usually observed in a ChIP experiment, one corresponding to the read length ("phantom" peak) and one to the average fragment length of the library. (*F*) Correlation between the fraction of reads within called regions and the relative cross-correlation coefficient for 1052 human ChIP-seq experiments. (*G*) The absolute and relative height of the two peaks are useful determinants of the success of a ChIP-seq experiment. A high-quality IP is characterized by a ChIP peak that is much higher than the "phantom" peak, while often very small or no such peak is seen in failed experiments.
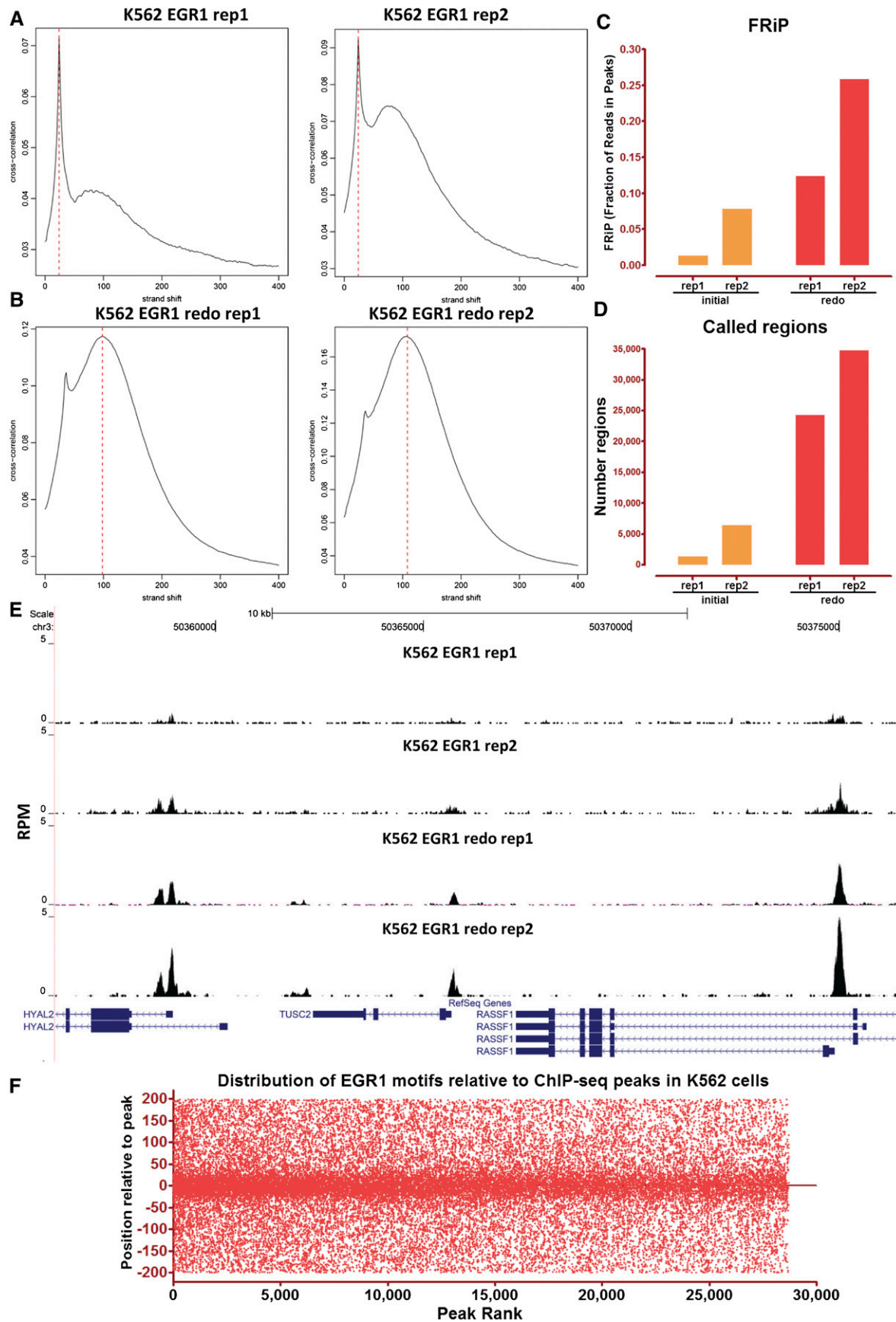
**Figure 5.** Quality control of ChIP-seq data sets in practice. EGR1 ChIP-seq was performed in K562 cells in two replicates. ChIP enriched regions were identified using MACS. However, the cross-correlation plot profiles (*A*) indicated that both IPs were suboptimal, with one being unacceptable. In agreement with this judgment, ChIP enrichment (*C*) and peak number (*D*) also indicated failure. The ChIP-seq assays were repeated (*B*), with all quality control metrics improving significantly (*B,D*), and many additional EGR1 peaks were identified as a result. (*E*) Representative browser snapshot of the four EGR1 ChIP-seq experiments, showing the much stronger peaks obtained with the second set of replicates. (*F*) Distribution of EGR1 motifs relative to the bioinformatically defined peak position of EGR1-occupied regions derived from ChIP-seq data in K562 cells. Regions are ranked by their confidence scores as called by SPP.

---

**Box 3:** ChIP-seq quality assessment guidelines

Within ENCODE, a set of data quality thresholds has been established for submission of ChIP-seq data sets. These have been constructed based on the historical experiences of ENCODE ChIP-seq data production groups with the purpose of balancing data quality with practical attainability and are routinely revised. The current standards are below and the performance of ENCODE data sets against these thresholds is shown in Figure 7.

**Cross-correlation analysis**

The current ENCODE practice is to calculate and report NSC and RSC for each experiment. For experiments with NSC values below 1.05 and RSC values below 0.8, we currently recommend that an additional replicate be attempted or the experiment explained in the data submission as adequate based on additional considerations.

**Irreproducible discovery rate (IDR)**

The following guidelines have been established for mammalian cells (optimal parameter may differ for other organisms). Biological replicates are performed for each ChIP-seq data set and subjected to peak calling. IDR analysis is then performed with a 1% threshold. For submission to ENCODE, we currently require that the number of bound regions identified in an IDR comparison between replicates to be at least 50% of the number of regions identified in an IDR comparison between two "pseudoreplicates" generated by pooling and then randomly partitioning all available reads from all replicates ($N_p/N_t < 2$) (Fig. 7). To ensure similar weighting of individual replicates for identifying binding regions, we further recommend that the number of significant peaks identified using IDR on each individual replicate (obtained by partitioning reads into two equal groups for the IDR analysis) be within a factor of 2 of one another ($N_1/N_2 < 2$) (Fig. 7). Data sets which fail to meet these criteria may still be deposited by ENCODE experimenters, provided that at least three experimental replicates have been attempted and a note accompanies these data sets explaining which parameters fail to meet the standards and providing any technical information that may explain this failure. This guideline is for point source features; metrics are still being determined for broad peak analyses.

Updated information about the performance of ENCODE data sets against these quality metrics and tools for determining these metrics will be forthcoming through the ENCODE portal (http://encodeproject.org/ENCODE/).

**Historical note**

A simpler heuristic for establishing reproducibility was previously used as a standard for depositing ENCODE data and was in effect when much of the currently available data was submitted. According to this standard, either 80% of the top 40% of the targets identified from one replicate using an acceptable scoring method should overlap the list of targets from the other replicate, or target lists scored using all available reads from each replicate should share more than 75% of targets in common. As with the current standards, this was developed based on experience with accumulated ENCODE ChIP-seq data, albeit with a much smaller sample size.

---

icant and insignificant findings is plotted, a transition in consistency is expected (Fig. 6C,F). This consistency transition provides an internal indicator of the change from signal to noise and suggests how many peaks have been reliably detected.

The IDR statistic quantifies the above expectations of consistent and inconsistent groups by modeling all pairs of peaks present in both replicates as belonging to one of two groups: a reproducible group, and an irreproducible group (Li et al. 2011). In general, the signals in the reproducible group are more consistent (i.e., have a larger correlation coefficient) and are ranked higher than the irreproducible group. The proportion of identifications that belong to the "noise" component and the correlation of the significant component are estimated adaptively from the data. The IDR provides a score for each peak, which reflects the posterior probability that the peak belongs to the irreproducible group.

A major advantage of IDR is that it can be used to establish a stable threshold for called peaks that is more consistent across laboratories, antibodies, and analysis protocols (e.g., peak callers) than are FDR measures (A Kundaje, Q Li, B Brown, J Rozowsky, A Harmanci, S Wilder, S Batzoglou, I Dunham, M Gerstein, E Birney, et al., in prep.). Increased consistency comes from the fact that IDR uses information from replicates, whereas the FDR is computed on each replicate independently. The application of IDR to real-life data is shown in Figure 6. A pair of high-quality RAD21 ChIP-seq replicates display good consistency between IDR ranks for a large number (~28,000) of highly reproducible peaks (Figs. 6A,B), with a clear inflection between the signal and noise populations near the 1% IDR value (Fig. 6C). In contrast, a pair of SPT20 replicates, which had already been flagged as low-quality based on the individual FRiP and NSC/RSC metrics, display very low IDR reproducibility, with very few significant peaks, and no visible inflection in the IDR curve (Fig. 6F).

It is important that the peak-calling threshold used prior to IDR analysis not be so stringent that the noise component is entirely unrepresented in the data, because the algorithm requires sampling of both signal and noise distributions to separate the peaks into two groups; thus relaxing the default stringency settings when running a given peak caller is advised if IDR analysis will follow.

A caution in applying IDR is that it is dominated by the weakest replicate (A Kundaje, Q Li, B Brown, J Rozowsky, A Harmanci, S Wilder, S Batzoglou, I Dunham, M Gerstein, E Birney, et al., in prep.). That is, if one replicate is quite poor, many "good" peaks from the higher quality replicate will be rejected by IDR analysis, because they are not reproducible in the weak replicate. To ensure similar weighting of individual replicates, the number of significant binding regions identified using IDR on each individual replicate (obtained by partitioning reads into two equal groups to allow the IDR analysis) is recommended to be within a factor of 2 for data sets to be submitted to UCSC by ENCODE (Box 3).

ENCODE has begun applying IDR analysis to all ChIP experiments. For all submitted ENCODE ChIP-seq data sets, the number of bound regions identified in an IDR comparison between replicates is at least 50% of the number of regions identified in an IDR comparison between two "pseudoreplicates" generated by randomly partitioning available reads from all replicates (Box 3).

## Guidelines for reporting ChIP-seq data

To facilitate data sharing among laboratories, both within and outside the Consortium, and to ensure that results can be reproduced, ENCODE has established guidelines for data sharing in public repositories. Raw data can be submitted to the Short Read Archive (SRA) and ChIP results are submitted to GEO. Through

## RAD21 Replicates (high reproducibility)

## SPT20 Replicates (low reproducibility)
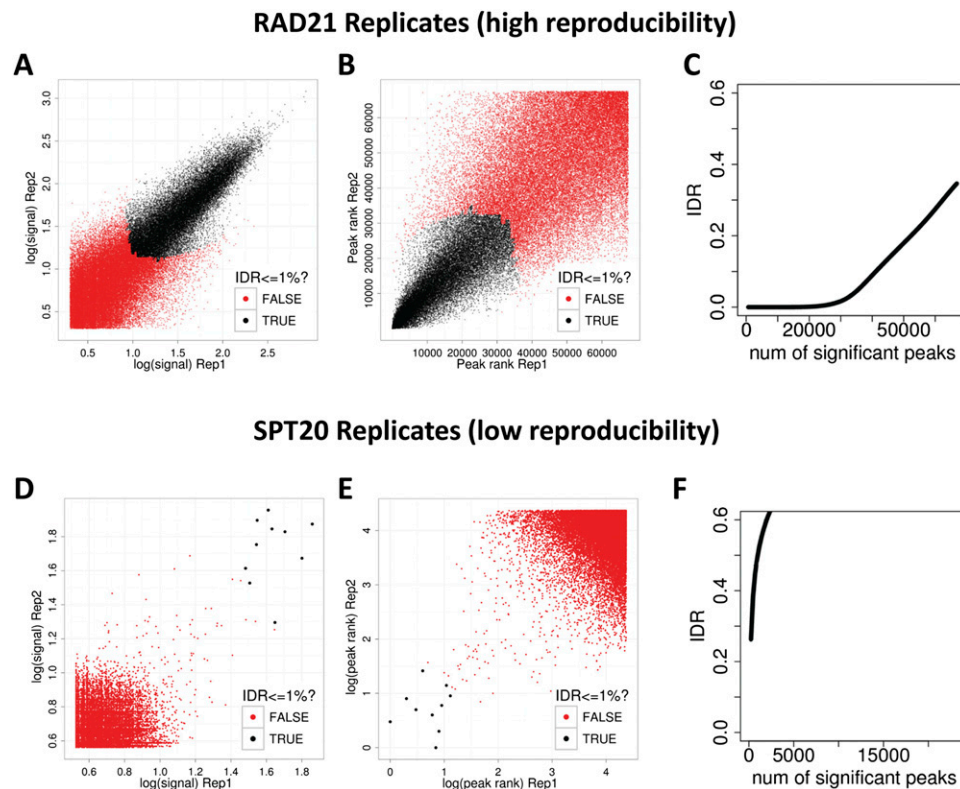


**Figure 6.** The irreproducible discovery rate (IDR) framework for assessing reproducibility of ChIP-seq data sets. (*A–C*) Reproducibility analysis for a pair of high-quality RAD21 ChIP-seq replicates. (*D,E*) The same analysis for a pair of low quality SPT20 ChIP-seq replicates. (*A,D*) Scatter plots of signal scores of peaks that overlap in each pair of replicates. (*B,E*) Scatter plots of ranks of peaks that overlap in each pair of replicates. Note that low ranks correspond to high signal and vice versa. (*C,F*) The estimated IDR as a function of different rank thresholds. (*A,B,D,E*) Black data points represent pairs of peaks that pass an IDR threshold of 1%, whereas the red data points represent pairs of peaks that do not pass the IDR threshold of 1%. The RAD21 replicates show high reproducibility with ∼30,000 peaks passing an IDR threshold of 1%, whereas the SPT20 replicates show poor reproducibility with only six peaks passing the 1% IDR threshold.

April 2012, 478 ChIP-seq data sets had been submitted to GEO at accession ID PRJNA63441, with submission of all current ENCODE data to be completed by June 2012. UCSC houses the ENCODE data (Rosenbloom et al. 2011) and modMine houses the modENCODE data (Contrino et al. 2011).

Box 4 provides a detailed description of the data and experimental and analytical details to be shared so that others can reproduce both experiments and analyses. Shared information includes the experimental procedures for performing the ChIP, antibody information and validation data, as well as relevant DNA sequencing, peak calling, and analysis details. For ENCODE experiments that do not meet the guidelines described above, data and results may be reported, with a note indicating that the criteria have not been met and explaining why the data are nevertheless released.

## Discussion

The ENCODE and ModENCODE standards and practices presented here will be further revised as the protocols, technologies, and our understanding of the assays change. Updated versions will be released and made available at http://encodeproject.org/ENCODE/experiment_guidelines.html. We have begun to address the central but vexing issue of immune reagent specificity and performance by establishing a menu of primary and secondary methods for antibody characterization, including performance-reporting practices. We also developed and applied global metrics to assess the quality of several aspects of an individual ChIP-seq experiment: Library complexity can be measured by the nonredundant fraction (NRF); immunoenrichment can be measured by the fraction of reads in called peaks (FRiP) and by cross-correlation analysis (NSC/RSC); and replicate significance can be measured by IDR. We related these global quality measures to more traditional inspection of ChIP-seq browser tracks (Fig. 5) and discuss below how different aspects of data quality interact with specific uses of ChIP-seq data.

## How good can a ChIP-seq experiment be?

Thus far, the most successful point-source factor experiments for ENCODE have FRiP values of 0.2–0.5 (factors such as REST, GABP, and CTCF) (Fig. 4C) and NSC/RSC values of 5–12. Although these quality scores and characteristics were routinely obtained for the best-performing factor/antibody combinations, they are not the rule; for most transcription factors, the ChIP quality metrics were substantially lower and more variable (Fig. 7). We believe that multiple issues contribute to the variability; the quality of antibody (affinity and specificity) is surely important, but epitope availability within fixed chromatin, sensitivity of the antibody to post-translational modifications of the antigen, how long and how often the protein is bound to DNA, and other physical characteristics of the protein–DNA interaction likely also contribute. Further work with epitope-tagged factors, for which the antibody is not a variable, should begin to sort among the possibilities.

---

**Box 4.** Data reporting guidelines

Data should be submitted to public repositories. The following information is currently used by ENCODE/modENCODE to submit data to public repositories.

**Metadata**

For submission of basic experimental data by ENCODE, the following information is minimally included:

- Investigator, organism, or cell line, experimental protocol (or reference to a known protocol).
- Indication as to whether an experiment is a technical or biological replicate.
- Catalog and lot number for any antibody used. If not a commercial antibody, indicate the precise source of the antibody.
- Information used to characterize the antibody, including summary of results (images of immunoblots, immunofluorescence, list of proteins identified by mass spec, etc.).
- Peak calling algorithm[29] and parameters used, including threshold and reference genome used to map peaks.
- A summary of the number of reads and number of targets for each replicate and for the merged data set.
- Criteria that were used to validate the quality of the resultant ChIP-seq data (i.e., overlap results or IDR[29]).
- Experimental validation results (e.g., qPCR).
- Link to the control track that was used.
- An explanation if the experiment fails to meet any of the standards.

**High-throughput sequencing data**

- Image files from sequencing experiments do not need to be stored.
- Raw data (FASTQ files) should be submitted to both GEO and SRA.
- Each replicate should be submitted independently.
- Target region and peak calling results.

**Point source peaks**

For point source peaks (e.g. experiments with antibodies to sequence-specific transcription factors), common features that are reported by ENCODE researchers include:

- Peak position, defined as a single base pair.
- Start and end positions, defined as specific base pairs.
- Signal value (e.g., fold enrichment) using an algorithm chosen by the submitter.
- Significance/accuracy measures:

  ☐ P-value determined using a method chosen by the submitter.
  ☐ Q-value (false discovery rate correction) determined using a method chosen by the submitter.

- Metadata, including peak caller approach and genome reference used, plus methods for determining signal values, P-values, and Q-values, as applicable.

**Broad regions**

- Start and end positions, defined as specific base pairs.
- Signal value (e.g., fold enrichment) using an algorithm chosen by the submitter.
- Significance/accuracy measures:

  ☐ P-value determined using a method chosen by the submitter.
  ☐ Q-value (false discovery rate correction) determined using a method chosen by the submitter.

- Metadata, including peak caller approach and genome reference used, plus methods for determining signal values, P-values, and Q-values, as applicable.
- Point-source peaks can be called in addition to broad regions (i.e., one can have "peaks" and potentially "valleys" within "regions").

The investigator should determine whether their data best fits the broad region/point source peak data or both.

---

When measurements differ in quality, the higher-quality replicate often identifies thousands more sites than the lower. Do sites present only in the superior ChIP experiment reflect true occupancy? Motif analysis suggests that many do. In Figure 5F, the position of EGR1 motifs relative to EGR1 ChIP-seq peaks is shown.

The known binding motif is prominent and concentrated centrally under the ChIP peaks, as expected if the motif mediates occupancy; importantly, the central location of the motif is observed, even in the low-ranking peaks. The trend continues below the peak-calling cut-offs, suggesting additional true occupancy sites. Depending on the goals of an analysis, users may want to be more or less conservative in defining the threshold for inclusion. Motif presence could be used as one criterion for "rescuing" candidate sites identified in only one experiment.

---

[29]For uniform peak calling within ENCODE, the MACS peak caller, version 1.4.2 was used. Scripts used for IDR analysis are at https://sites.google.com/site/anshulkundaje/projects/idr.
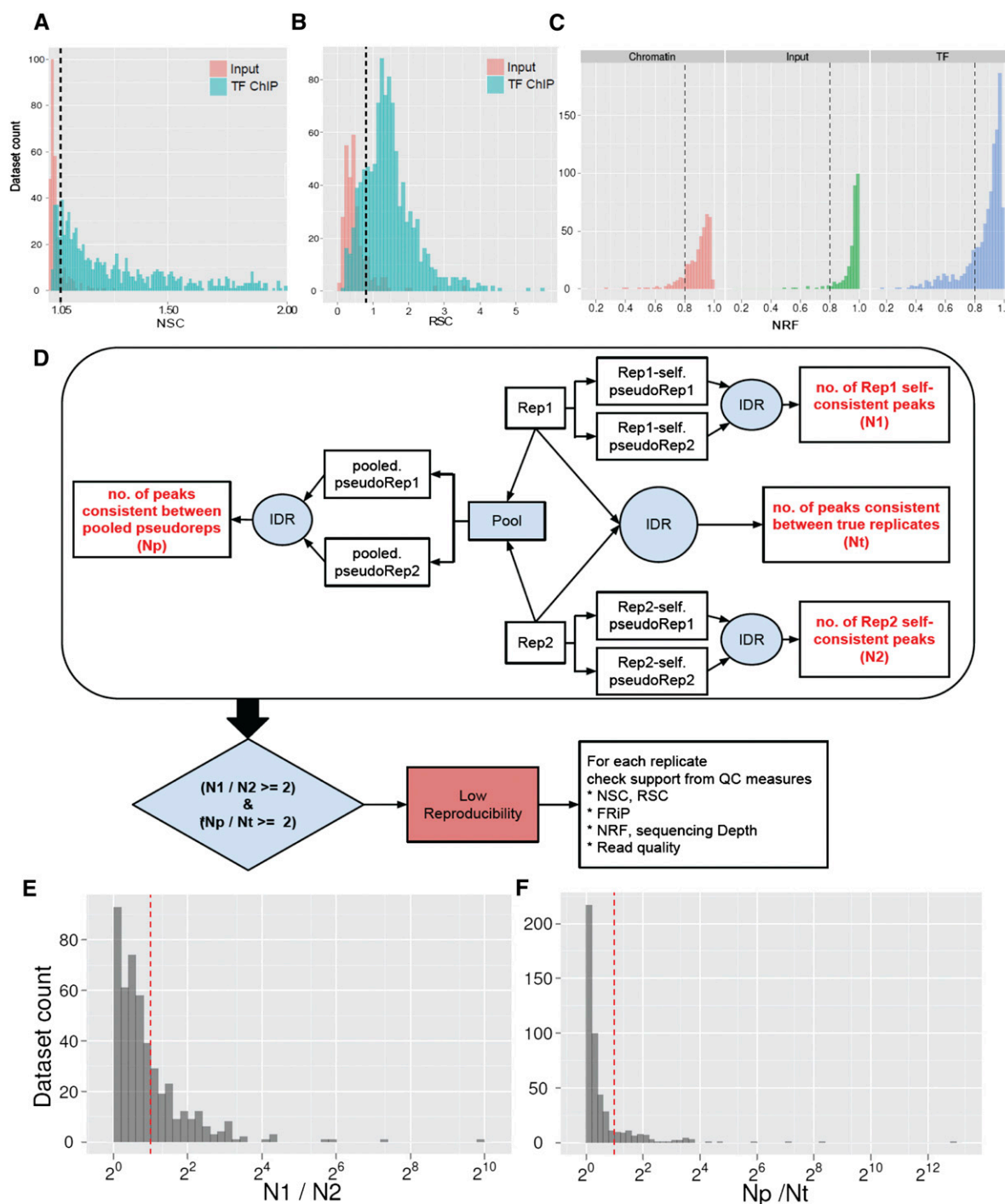
**Figure 7.** Analysis of ENCODE data sets using the quality control guidelines. (*A–C*) Thresholds and distribution of quality control metric values in human ENCODE transcription-factor ChIP-seq data sets. (*A*) NSC, (*B*) RSC, (*C*) NRF. (*D*) IDR pipeline for assessing ChIP-seq quality using replicate data sets. (*E,F*) Thresholds and distribution of IDR pipeline quality control metrics in human ENCODE transcription factor ChIP-seq data sets. (Dashed lines) Current ENCODE thresholds for the given metric, which are NSC > 1.05 (*A*); RSC > 0.8 (*B*); NRF > 0.8, N1/N2 ≥ 2 (where N1 refers to the replicate with higher N) (*E*); Np/Nt ≥ 2 (*F*).

## How good does a ChIP-seq experiment need to be?

We have observed that some biologically important sites can have modest ChIP-seq signals (Fig. 4B), while some sites with very high enrichment fail to give positive functional readouts in follow-up experiments. Given this, the best practical guidance for setting thresholds of sensitivity, specificity, and reproducibility will depend on how the data are to be used. Below, we outline four different common ChIP uses, ranging from more relaxed to stringent in their requirements toward data quality and site-calling sensitivity.

## Motif analysis

Deriving DNA sequence motifs for a ChIP-assayed factor is relatively simple and has been performed successfully for most ENCODE ChIP-seq data sets (Fig. 2E). Experiments that pass the thresholds we use for NRF, FRiP, and NSC/RSC typically produce thousands to tens of thousands of regions, a sub-sample of which can be readily used to deduce the recognition motif, although more than one motif subfamily is sometimes found by additional analysis (Johnson et al. 2007). Causal motifs are typically centrally positioned and this can be used as a confirming diagnostic (Fig. 6F). Notably, motif derivation can also be successful from marginal quality data that fall below recommended quality metric thresholds (especially if only the top-ranked peaks are used). However, the risk of artifacts increases, and results from such analyses should be cautiously interpreted and stringently validated.

## Discovering regions to test for biological function such as transcriptional enhancement, silencing, or insulation

Biologists often use ChIP-seq data to identify candidate regulatory regions at loci of interest. When the goal is to find a few examples of regulatory domains bound by a factor, data of modest quality can still be useful if combined with close inspection of ChIP signals and the corresponding controls before investing in functional and/or mutagenesis studies. However, if the aspiration is to identify a comprehensive collection of all candidate regulatory regions bound by a factor, very high-quality and deeply sequenced data sets are required.

## Deducing and mapping combinatoric occupancy

Typical cis-acting regulatory modules (CRM) are occupied by multiple factors (Ghisletti et al. 2010; Lin et al. 2010; Wilson et al. 2010; A He et al. 2011; Q He et al. 2011; Tijssen et al. 2011) and associated with multiple histone modifications (Barski et al. 2007; Mikkelsen et al. 2007; Wang et al. 2008). A frequent goal of ChIP-seq studies is to deduce a combination of factors that mediate a common regulatory action at multiple sites in the genome. This is a very quality-sensitive use of ChIP data since the presence of one or more weak data sets that fail to identify significant fractions of the true occupancy sites can seriously confound the analysis; therefore we recommend only the highest quality data sets be used for such analyses.

## Integrative analysis

A new frontier of whole-genome analysis is the integration of data from many (hundreds or thousands) experiments with the goal of uncovering complex relationships. These endeavors typically use sophisticated machine learning methods (Ernst and Kellis 2010; Ernst et al. 2011; A Mortazavi, S Pepke, G Marinov, and B Wold, in prep.) with complex and varying sensitivity to ChIP strength; and such efforts can be very sensitive to data quality.

## Conclusion

Our goal in developing these current working guidelines for ChIP-seq experiments, now applied over a large number of factors, was to provide information about experimental quality for users of modENCODE and ENCODE data. The strongest ChIP-seq data-sets that readily meet all quality specifications should be especially useful for regulatory network inference and for diverse integrative

analyses, including the effects of genetic variation on human traits and disease. The metrics, methods, and thresholds might also be useful to the wider community, although our intention in outlining our approaches was not to imply that ENCODE criteria must be applied rigidly to all studies. As discussed above, some ChIP data and antibodies can and do fall outside these guidelines for varied reasons, yet are highly valuable. In such cases it is critical to try to understand why a data set looks unusual, and to assess the implications for specific uses of those data or reagents. Similar guidelines exist in ENCODE for RNA-seq, DNase-seq, FAIRE-seq, ChIA-PET, and other related assays; the working standards and protocols for these techniques can be found at the ENCODE and modENCODE websites (http://encodeproject.org/ENCODE/experiment_guidelines.html).

## Data access

All data sets used in the analysis have been deposited for public viewing and download at the ENCODE (http://encodeproject.org/ENCODE/) and modENCODE (http://www.modencode.org/) portals.

## List of affiliations

[1]Department of Genetics, Stanford University, Stanford, California 94305, USA; [2]Division of Biology, California Institute of Technology, Pasadena, California 92116, USA; [3]Department of Computer Science, Stanford University, Stanford, California 94305, USA; [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [5]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; [6]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [7]Department of Statistics, University of California, Berkeley, California 94720, USA; [8]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, Massachusetts 02215, USA; [9]Computational Biology & Bioinformatics Program, Yale University, New Haven, Connecticut 06511, USA; [10]Department of Computer Science and Center for Systems Biology, Duke University, Durham, North Carolina 27708, USA; [11]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [12]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas 78701, USA; [13]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA; [14]Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Boston, Massachusetts 02115, USA; [15]Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA; [16]Department of Statistics, Penn State University, University Park, Pennsylvania 16802, USA; [17]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA; [18]National Human Genome Research Institute/National Institutes of Health, Rockville, Maryland 20852, USA; [19]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada; [20]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06511, USA; [21]Department of Pathology, Stanford University, Stanford, California 94305, USA; [22]Department of Medicine, University of Washington, Seattle, Washington 98195, USA; [23]University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA; [24]Department of Biochemistry & Molecular Biology,

Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California 90089, USA; [25]Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.

## Acknowledgments

## References

Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrançois P, Struhl K, Gerstein M, Snyder M. 2009. Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci* **106:** 14926–14931.

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

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459:** 927–930.

Contrino S, Smith RN, Butano D, Carr A, Hu F, Lyne R, Rutherford K, Kalderimis A, Sullivan J, Carbon S, et al. 2011. modMine: Flexible access to modENCODE data. *Nucleic Acids Res* **40:** D1082–D1088.

DeKoter RP, Singh H. 2000. Regulation of B lymphocyte and macrophage development by graded expression of PU.1. *Science* **288:** 1439–1441.

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

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

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

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

Ernst J, 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 TS, Shoresh N, Ward LD, Epstein CB, 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.

Fernandez PC, Frank SR, Wang L, Schroeder M, Liu S, Greene J, Cocito A, Amati B. 2003. Genomic targets of the human c-Myc protein. *Genes Dev* **17:** 1115–1129.

Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ. 2010. ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS ONE* **5:** e15082. doi: 10.1371/journal.pone.0015082.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330:** 1775–1787.

Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, Gregory L, Lonie L, Chew A, Wei C, et al. 2010. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* **32:** 317–328.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510.

He A, Kong SW, Ma Q, Pu WT. 2011. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci* **108:** 5632–5637.

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat Genet* **43:** 414–420.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589.

Horak CE, Snyder M. 2002. ChIP-chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol* **350:** 469–483.

Hua S, Kittler R, White KP. 2009. Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell* **137:** 1259–1271.

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533–538.

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26:** 1293–1300.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26:** 1351–1359.

Krig SR, Jin VX, Bieda MC, O'Geen H, Yaswen P, Green R, Farnham PJ. 2007. Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays. *J Biol Chem* **282:** 9703–9712.

Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, et al. 2010. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol* **6:** 377. doi: 10.1038/msb.2010.31.

Li Q, Brown J, Huang H, Bickel P. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5:** 1752–1779.

Lieb JD, Liu X, Botstein D, Brown PO. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28:** 327–334.

Lin YC, Jhunjhunwala S, Benner C, Heinz S, Welinder E, Mansson R, Sigvardsson M, Hagman J, Espinoza CA, Dutkowski J, et al. 2010. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat Immunol* **11:** 635–643.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.

Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, et al. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.

Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, Zeng L, Ogawa N, Wold BJ, Stathopoulos A. 2011. High resolution mapping of Twist to DNA in *Drosophila* embryos: Efficient functional analysis and evolutionary conservation. *Genome Res* **21:** 566–577.

Park PJ. 2009. ChIP-seq: Advantages and challenges of a maturing technology. *Nat Rev Genet* **10:** 669–680.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6:** S22–S32.

Poser I, Sarov M, Hutchins JRA, Hériché J, Toyoda Y, Pozniakovsky A, Weigl D, Nitzsche A, Hegemann B, Bird AW, et al. 2008. BAC TransgeneOmics: A high-throughput method for exploration of protein function in mammals. *Nat Methods* **5:** 409–415.

Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M. 2010. Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc Natl Acad Sci* **107:** 3639–3644.

Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol* **12:** R67. doi: 10.1186/gb-2011-12-7-r67.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4:** 651–657.

Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, et al. 2011. ENCODE whole-genome data in the UCSC Genome Browser: Update 2012. *Nucleic Acids Res* **40:** D912–D917.

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330:** 1787–1797.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27:** 66–75.

Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, Jang S, Margueron R, Reinberg D, Green R, Farnham PJ. 2006. Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* **16:** 890–900.

Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, Bellissimo DC, Oram SH, Smethurst PA, Wilson NK, et al. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* **20:** 597–609.

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5:** 829–834.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40:** 897–903.

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16:** 235–244.

Wilson NK, Foster SD, Wang X, Knezevic K, Schütte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, et al. 2010. Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7:** 532–544.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137. doi: 10.1186/gb-2008-9-9-r137.

Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HYK, Preston E, et al. 2010. Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet* **6:** e1000848. doi: 10.1371/journal.pgen.1000848.

# ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, et al.

| | |
|---|---|
| **Supplemental Material** | http://genome.cshlp.org/content/suppl/2012/08/22/22.9.1813.DC1 |

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

To subscribe to *Genome Research* go to:
**http://genome.cshlp.org/subscriptions**

**References**
This article cites 51 articles, 14 of which can be accessed free at:
**http://genome.cshlp.org/content/22/9/1813.full.html#ref-list-1**

Articles cited in:
**http://genome.cshlp.org/content/22/9/1813.full.html#related-urls**

**Open Access**
Freely available online through the *Genome Research* Open Access option.

To subscribe to *Genome Research* go to:
**http://genome.cshlp.org/subscriptions**