Crunch: Completely Automated Analysis of ChIP-seq Data

SEVERIN BERGER[1], SAEED OMIDI[1,2], MIKHAIL PACHKOV[1], PHIL ARNOLD[1,3], NICHOLAS[1,3] KELLEY, SILVIA SALATINO[1,4], ERIK VAN NIMWEGEN[1,*]

[1]*Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland*

[2]*Current address: École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[3]*Current address: Novartis Institutes for Biomedical Research, Basel, Switzerland*

[4]*Current address: Oxford Genomics Center, Oxford, UK*

**Corresponding author: erik.vannimwegen@unibas.ch*

March 8 2016

ABSTRACT

Today experimental groups routinely apply ChIP-seq technology to quantitatively characterize the genome-wide binding patterns of any molecule associated with the DNA. Here we present Crunch, a completely automated procedure for ChIP-seq data analysis, starting from raw read quality control, through read mapping, peak detection and annotation, and including comprehensive DNA sequence motif analysis. Among Crunch's novel features are a Bayesian mixture model that automatically fits a noise model and infers significantly enriched genomic regions in parallel, as well as a Gaussian mixture model for decomposing enriched regions into individual binding peaks. Moreover, Crunch uses a combination of *de novo* motif finding with binding site prediction for a large collection of known regulatory motifs to model the observed ChIP-seq signal in terms of novel and known regulatory motifs, extensively characterizing the contribution of each motif to explaining the ChIP-seq signal, and annotating which combinations of motifs occur in each binding peak. To make Crunch easily available to all researchers, including those without bioinformatics expertise, Crunch has been implemented as a web server (crunch.unibas.ch) that only requires users to upload their raw sequencing data, providing all results within an interactive graphical web interface.

To demonstrate Crunch's power we apply it to a collection of 128 ChIP-seq data-sets from the ENCODE project, showing that Crunch's *de novo* motifs often outperform existing motifs in explaining the ChIP-seq signal, and that Crunch successfully identifies binding partners of the proteins that were immuno-precipitated.

INTRODUCTION

The advent of next-generation sequencing technologies, and the associated dramatic reduction of cost for sequencing, have led to a spectacular rise in the use of a variety of methods, including RNA-seq, ChIP-seq, DNaseq or CLIP-seq, that combine next-generation sequencing with other molecular biology techniques to quantitatively characterize internal states of cells on a genome-wide scale. As one of the most prominent technologies, ChIP-seq (1) combines chromatin immuno-precipitation with next-generation sequencing to quantify the genome-wide binding patterns of any molecule that associates with the DNA.

Large-scale efforts like ENCODE that systematically mapped DNA binding patterns of many TFs have led to important novel insights on their binding patterns on a genome-wide scale (2). In addition, more and more individual labs are using ChIP-seq to characterize the binding patterns of particular DNA binding proteins of interest in their specific system of interest and across particular conditions.

The result of a ChIP-seq experiment is simply a large collection of short DNA sequence reads, typically millions or tens of millions. Extracting comprehensive meaningful biological information from such data-sets is quite complex and involves a significant number of separate steps including quality control, read mapping, fragment length estimation, peak identification, peak annotation, and various downstream analyses such as the identification of sequence motifs enriched within the peak sequences. Through the efforts of many bioinformatics groups, a substantial number of tools have been developed for each of these analysis steps, e.g. for mapping of reads (3), detecting binding peaks (4), and for the discovery of sequence motifs over-represented among a set of short sequence segments (5). In addition, a number of solutions have been presented that allow combining these individuals tools into a workflow, i.e. by allowing users to manually execute one tool after another or by constructing a pipe-line that runs the tools automatically. For example, there are commercial solutions such as *Avadis NGS* (6), *Chipster* (7), *CLCbio Genomics Workbench* (8), *Genomatix Mining Station* (9), *Adtridbio GenoMiner* (10) and *Partek Genomics Suite* (11) as well as free-to-use solutions such as *HOMER* (12), *cisGenome* (13), *seqMINER* (14), *ChIPseeqer* (15), *GeneProf* (16) and *Galaxy/Cistrome* (17).

Here we present Crunch, a new integrated ChIP-seq analysis procedure, which is novel in a number of respects. First of all, Crunch is completely automated and implemented in a web server (crunch.unibas.ch) such that researchers are only required to upload raw sequencing reads and identify the model organism from which the data derive, which may be human, mouse or drosophila. That is, there is no need for choosing among tools, for tuning or setting any parameters, or for managing the movement of output from one analysis step into inputs for subsequent analysis steps. In addition, besides flat files for download, all results of Crunch's analysis are available through an easily navigable graphical web interface. In this way, Crunch provides a simple, automated, and comprehensive analysis of ChIP-seq data which will be especially attractive to experimental researchers that produce such data but lack the required bioinformatics expertise, or lack the necessary computer hardware to run the analysis tools.

An overview of the Crunch workflow, which is implemented using the Anduril workflow engine (21), is provided in Fig. 1. Besides providing complete automation of such procedures as quality control, adaptor removal, read mapping, and fragment size estimation, the methodology that Crunch employs includes several important novel features. Firstly, following our previous work on noise characteristics of next-generation sequencing data (18), we model fluctuations in read counts as a combination of multiplicative and Poisson sampling noise and developed a Bayesian procedure for, at the same time, estimating the parameters of this noise distribution and inferring the genomic regions with significantly enriched read counts. In addition, we employ automated Gaussian mixture modelling to decompose each genomic region with significant enrichment of ChIP-seq reads into individual binding peaks. Secondly and most importantly, Crunch combines *de novo* motif finding, using our previously developed PhyloGibbs algorithm (19), with comprehensive binding site prediction for a large library of known regulatory motifs, using our MotEvo algorithm (20), to model the observed ChIP-seq signal in terms of sites for novel and known

regulatory motifs. In particular, Crunch identifies a set of novel and known regulatory motifs that are complementary to each other and collectively best explain the observed ChIP-seq signal. Moreover, Crunch extensively characterizes the contribution of each motif using a number of different statistics. In this way, Crunch is able to infer not only the binding specificity of the transcription factor (TF) that was immuno-precipitated, but also identify other TFs that either directly bind or co-occur with the immuno-precipitated TF.

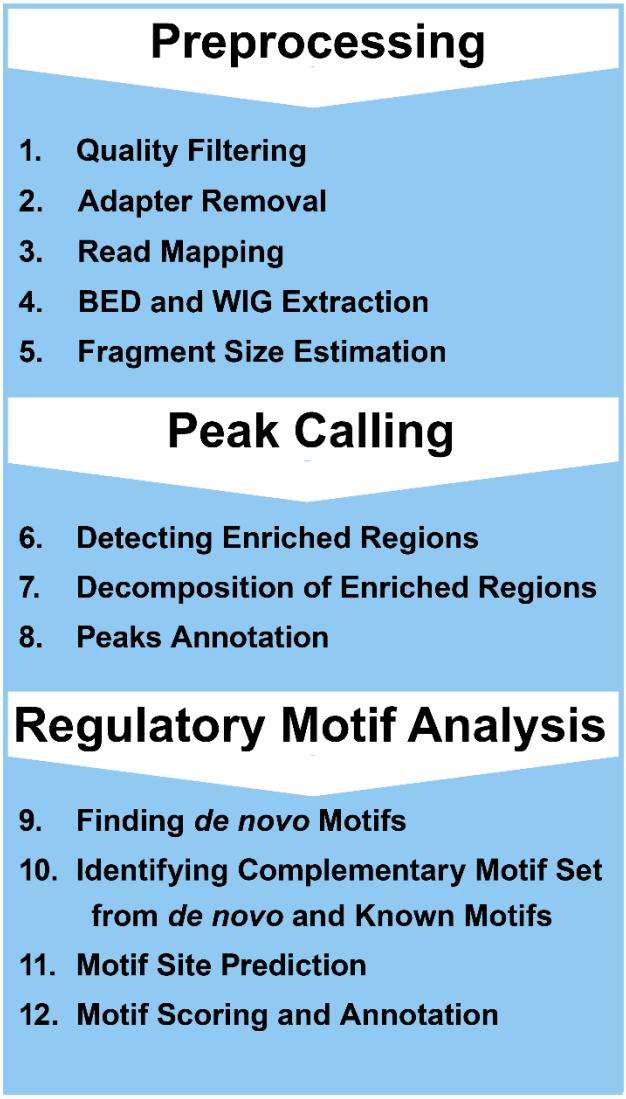


Figure 1: An overview of Crunch's ChIP-seq analysis workflow. The workflow consists roughly of three parts, i.e. pre-processing of the data, binding peak identification, and regulatory motif analysis.

To demonstrate Crunch's performance, we applied it to 128 ChIP-seq data sets for 95 different TFs from the ENCODE experiments and compared the performance of the *de novo* motifs identified by Crunch with known motifs from other resources. This analysis shows that, for the majority of data-sets, Crunch's motifs outperform known motifs.

MATERIALS AND METHODS

Quality Control and Adapter Removal

Raw sequencing reads are typically provided in FASTQ format and Crunch directly takes such FASTQ files as input. These raw reads can have diverse quality due to uncertainty in base calling and errors in the sequencing itself, and they potentially contain artefacts such as sequenced parts of sequencing adapters. To avoid contaminating dowstream analyses with low quality or erroneous sequence, we perform an initial quality filtering, followed by adaptor removal, and a final round of quality filtering as follows.

In the first round of quality filtering we discard reads that are either shorter than 25 nucleotides (nt), contain more than 2 ambiguous nucleotides (N's), or have an average Phred base calling quality score below 20 (corresponding to an error rate of 1%). As sequencing quality tends to decrease from the 5' to the 3' end of the read, we select the longest 5' prefix of the read that has an average Phred score of at least 20. The chosen prefix is maintained if it has a minimum length of 25 nt and contains at most 2 sequencing errors.

In the adapter removal step, we focus exclusively on 3' adapters, i.e. adapters that get (partially) sequenced if the sequence of interest (i.e. the fragment) is shorter than the length of the sequenced read. Crunch first aims to determine, for any given data set, which 3' adapter was used. For this a list is used that by default contains adapters from (22) and that can be extended by the user if

desired. For each adapter in the list, prefixes of length 14, 16, 18 and 20 nt are mapped to 250'000 randomly chosen reads from the data set, allowing for 2 mismatches. The adapter with the highest average number of matches is chosen as the putative adapter sequence. This adapter sequence is then trimmed stringently from the reads according to the procedure previously described in (23) and supplementary text S1.

The second round of quality filtering includes the discarding of low complexity reads and – as necessitated by the truncation of some reads introduced through adapter removal – a 14 nt threshold on read length. For complexity filtering we calculate, for every read, the normalised di-nucleotide entropy H given by $-\sum_i (f_i \log(f_i)) / \log(16)$ where i runs over all di-nucleotides and f_i is the frequency of dinucleotide *i*. All reads with H < 0.5 are discarded (23).

Mapping

After filtering, Crunch maps the remaining reads to the reference genome using bowtie version 1.1.1 (24). Bowtie's parameters are set such that for every read all mapping positions with the least number of mismatches get reported (-a --strata --best) allowing for at most three mismatches (-v 3) and skipping reads when the number of mapping positions exceeds 100 (-m 100). For further analysis we use mapped reads where each mapping position is weighted by the copy number of the read divided by the number of positions the read was reported to map to. We store these aligned reads in a BED-like format called BEDWEIGHT, which besides specifying the location of the mapping, also specifies the mapping's weight. To allow visualization of the raw data in a genome browser, we also produce WIG files of the aligned reads at a 100 nt resolution.

Fragment Size Estimation

After shearing and pulling down the DNA, ChIP-seq protocols for sequencing library preparation generally include a step that selects fragments in a certain size range. Typically this selected fragment size is significantly longer than the length of the sequencing reads such that, for each double stranded DNA fragment, reads are produced from the 3' ends of either strand of the fragment. Consequently, the read distribution in the neighbourhood of a protein binding site typically shows two peaks on opposite strands of the DNA, approximately one fragment size apart (25). The typical fragment size can thus be estimated by analysis of the correlation between read counts on opposite strands as a function of their distance. Crunch estimates the fragment size by finding the distance *d* that maximizes the correlation function

$$C(d) = \sum_{i \notin R} r_+(i) r_-(i+d) \qquad (1)$$

where $r_+(i)$and $r_-(i+d)$ are indicator functions that equal one if at least one read starts on the plus strand at position *i*, and on the minus strand at position *i*+*d*, respectively, and the sum is over all genomic positions excluding regions annotated as repeats (which we denoted by *R*). Note that we use indicator functions rather than raw read counts to avoid the correlation function to be dominated by a few pairs of positions with aberrantly large read counts. $C(d)$ is computed for a range of d between 0 and 600 nt. A typical resulting cross-correlation function is depicted in Fig. 2, left panel.

As Fig. 2 illustrates, we often observe a local maximum in *C*(*d*) at a value of *d* that equals read length. Since this local maximum is clearly an artefact (we believe it derives from reads mapping to repeat regions), we determine the fragment size by finding the local maximum in *C*(*d*) while only considering values of d that are 80nt or more.

Peak Calling

Peaks are called in a two-step procedure. First, genomic regions that are enriched for reads from the chromatin immuno-precipitation are detected, and

second, individual binding events inside these enriched regions are identified.

Identifying enriched regions

For the first step we begin with counting mapped reads in genome-wide sliding windows both for the immuno-precipitated sample (which we call foreground) and a reference sample which typically consists of input DNA (which we call background). For each read, we estimate the central position of the corresponding fragment to be half a fragment size toward the 3' direction, i.e. forwards for reads on the plus strand and backwards for reads on the minus strand. We then slide a window of length 500 nt along the genome, shifting the window in steps of 250 nt, and count the number of mapped fragments whose estimated central position falls inside each window. The choice of window length 500 is a trade-off between obtaining sufficient mapped reads to measure local fragment density with reasonable accuracy, and obtaining sufficient spatial resolution to ensure that windows cover only one or a few binding peaks. If desired users can change this default window length.

To estimate the density of fragments from the background sample we count fragments using windows of length 2000 nt centred on the same position as the corresponding foreground windows. We decided to use windows of larger width for the background because the background signal typically varies more slowly along the genome, and because the background read density is typically lower than in binding peaks, larger regions are needed to suppress the Poisson sampling noise.

We then normalise the counts by the total sample read count to get read densities. In case of replicates we simply sum the counts of the replicates together.

The read densities in the background sample should ideally be approximately constant along the genome. However, as illustrated in Fig. 3A, we typically find that, while the large majority of windows has read counts within a narrow range, a very small fraction of windows shows abnormally high read counts. Although we have not extensively studied how these high read counts arise, we know that they tend to be repetitive regions that align poorly with the genomes of closely-related species (data not shown). We suspect that these repeat regions may have significantly expanded in the genome from which the sample derives compared to the reference assembly. These regions do not obey the statistics that are observed for the vast majority of the genome, and this leads to a high rate of false prediction of binding peaks in these regions. We thus developed a procedure to filter out these regions with unusually high background signal, and exclude them from further analysis.

Reasoning that normal background signals should show roughly exponential tails, we fit the tail of the reverse cumulative distribution of the background read counts to an exponential distribution and determine the point at which the observed cumulative starts deviating more than $e^{0.5}$ vertically from the exponential tail (red line in Fig. 3A). All windows with counts above the determined point are excluded from further analysis, which typically corresponds to approximately 0.1% of all windows.

To detect regions that have a significantly higher density of reads in the ChIP sample than in the background sample we fit a mixture model to the observed read counts across the genome. In particular, if *m* out of a total of *M* reads in the background sample fell in a given window, then we assume that the probability for *n* out *N* reads from the ChIP sample to fall in the same window is given by a two-component mixture:

$$P_{mix}(n, \mid N, m, M, \sigma, \mu, \rho) = \rho P_{bg}(\, n|N, m, M, \sigma, \mu) + (1-\rho) P_{fg}(\, n|N, m, M) \quad (2),$$

where $P_{bg}(\, n|N, m, M, \sigma, \mu, \rho)$ is the probability of observing *n* out of *N* reads under a background

model which assumes that there is no enrichment in the ChIP sample, and $P_{fg}(n|N,m,M)$ is the probability of observing *n* out of *N* reads under a `foreground' model that allows for an arbitrary enrichment in the ChIP-sample. The parameter *ρ* is the fraction of windows that are not enriched, while (μ,σ) are parameters of the background model that we will now explain.

As we have shown previously (18), the fluctuations in next-generation sequencing read-densities across replicate experiments can be well approximated by a combination of multiplicative noise (which may results from uncontrolled variations both in the biological state of the cells and variations in the process of preparing a sequence library from the sample) and Poisson sampling noise (from the sequencing itself), which leads to an approximately log-normal distribution of read-counts. This yields the following background distribution:

$$P_{bg}(n \mid N,m,M,\sigma,\mu) = \frac{1}{\sqrt{2\pi\left(2\sigma^2+\frac{1}{n}+\frac{1}{m}\right)}}\exp\left(-\frac{\left(\log\left(\frac{n}{N}\right)-\log\left(\frac{m}{M}\right)-\mu\right)^2}{2\left(2\sigma^2+\frac{1}{n}+\frac{1}{m}\right)}\right) \quad (3)$$

The term $2\sigma^2$ is the variance of the multiplicative noise component and the term $1/n + 1/m$ constitutes the contribution to the variance from the Poisson noise components of both the foreground and background samples (see (18) for details). As a significant fraction of the reads in the foreground sample derive from bound regions, the read density in regions without binding is systematically lower than in the background sample. The parameter μ corresponds to the resulting overall shift in log-density in the unbound regions. Note that this distribution differs from the negative binomial distribution, which is often used to model the statistics of next-generation sequencing read counts (26), and which is obtained when one assumes that biological replicate noise is approximately Gamma distributed as opposed to the log-normal distribution that our model assumes.

Finally, for the foreground distribution we assume a simple uniform distribution for the difference in log read-density. That is, we have $P_{fg}(n|N,m,M) = \frac{1}{W}$, where W is a constant that corresponds to the possible range of values for the difference in log read densities $f(n,m) = \log\left(\frac{n}{N}\right) - \log\left(\frac{m}{M}\right)$ which is set to $\max(f(n,m)) - \min(f(n,m))$. We then fit the parameters μ, σ and ρ by maximising the log-likelihood of the ChIP-seq data using expectation maximisation (for more detail consult the supplementary text S4). Finally, since the Gaussian approximation of equation (1) becomes inaccurate when the raw read counts are as low as zero or one, we only include window pairs where both the foreground and background window have a read count of 2 or more. In addition, we add a pseudo count of 0.5 to the read counts.

For every window we then compute a z-score with:

$$z = \frac{\log\left(\frac{n}{N}\right)-\log\left(\frac{m}{M}\right)-\mu}{\sqrt{2\sigma^2+\frac{1}{n}+\frac{1}{m}}} \quad (4)$$

Note that, if there was no binding at any of the windows in the genome, the statistic *z* should follow a standard normal distribution. As illustrated in Fig. 3B, for most data-sets we find that the z-scores of the large majority of windows indeed accurately follow a standard normal distribution, indicating that our statistical model successfully captures the noise distribution in the unenriched regions, which constitute the large majority of the genome. It also illustrates that a small fraction of windows shows substantially higher z-scores than expected under the background distribution.

We set the minimum z-score to call a window significantly enriched by fixing the false discovery rate to 0.1 (Fig. 3C and supplementary text S5 for more detail). All windows above the chosen z-score threshold are then used for further analysis. Since we chose the sliding windows to overlap, we merge

overlapping windows that passed the threshold into larger enriched regions.

Identifying individual binding peaks within enriched regions

Our methodology for identifying enriched genomic regions returns that are typically 500-1000 base pairs in length, which is significantly longer than the length of individual protein binding sites on the DNA. In the second step of peak calling we search for individual binding events by inspecting the ChIP-signal at a higher resolution. For this we compute for each position in each significantly enriched region the number of foreground fragments that overlap it. Here fragments are reads that were extended from their 5' end to fragment size in 3' direction. The result is a coverage profile for each significantly enriched region (Fig. 3D).

To detect individual binding events we now fit the coverage profile of each enriched region as a mixture of Gaussian peaks plus a uniform background distribution. Approximating the data as if the coverage *C*(*i*) at each position *i* in the region were an independent observation, the likelihood of the mixture model takes on the following form:

$$L(C \mid \vec{\mu}, \vec{\sigma}, \vec{\rho}, W) = \prod_{i=1}^{l} \left[\sum_j \rho_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(i-\mu_j)^2}{2\sigma^2} \right) + \left(1 - \sum_j \rho_j \right) \frac{1}{l} \right]^{C(i)} \qquad (5)$$

where i runs over all positions from 1 to l in the region, $C(i)$ is the coverage at position i, i.e. the number of fragments that are overlapping position i, j runs over all Gaussian peaks in the model, μ_j and σ_j are the central position and width of the Gaussian j, ρ_j is the fraction of all observations that belong to Gaussian j, and the last term corresponds to the uniform background distribution that accounts for the coverage not associated with the Gaussian peaks. Importantly, for individual binding sites on the genome, the width of the resulting coverage peak is a relatively well-defined function of the fragment length and we use this to constrain the widths σ_j of the Gaussian peaks to fall within a range that is consistent with these peaks corresponding to single binding sites on the genome (see supplementary text S5 for more details). We also use our knowledge of the typical width of individual binding peaks to set the numbers of Gaussians in the mixture model. In particular, the number of Gaussian components is determined by dividing the length of the region by four times the fragment size, and taking the floor of this value. In case this number is smaller than 2, we use 2 Gaussians in the mixture model. We again fit the parameters of the model by maximizing the likelihood of the coverage profile using the method of expectation maximisation (supplementary text S6).

After fitting the parameters of the mixture model we often end up with highly overlapping Gaussians. We merge overlapping Gaussians if the difference between their means is less than the sum of their widths (standard-deviations), which roughly corresponds to the condition that there is no local minimum in the coverage profile between the two peaks. The ρ-weighted averages of the means and standard-deviations of the overlapping Gaussians are then used to define a single merged Gaussian. Finally, for each Gaussian component, the binding peak is defined as the region from $\mu - \sigma$ until $\mu + \sigma$, as illustrated in Fig. 3D.

Note that, because the number of Gaussians used in the mixture is chosen to be an upper bound on the number of real binding peaks, some of the component peaks may not exhibit a significant enrichment over background. We thus calculate a z-score for each individual peak and only retain those peaks with a z-score above the threshold computed in the preceding section. The z-score is computed according to equation (4) using newly computed counts n which result from summing the contribution

of the Gaussian peak in equation, together with the uniform background. The number of renormalized number of reads associated with peak *j*, in a region of length 500, is given by

$$n_j = \frac{\sum_{i=1}^{l} C(i)}{f}\left(\rho_j + \frac{500}{l}\left(1 - \sum_j \rho_j\right)\right), \qquad (6)$$

where *f* is the fragment size. Note that the first factor corresponds to the total number of reads assigned to the region. The two term within the parentheses give the total fraction of these reads that are assigned to peak *j*. The first term, ρ_j, corresponds to the fraction directly assigned to the peak by the mixture model, and the second term corresponds to the total number of background reads in a region of length 500.

Association of binding peaks with genes and promoters

To annotate which genes may be regulated by the regulatory elements within the peak, we use our curated collection of promoters from SwissRegulon (18), (27), and record the three closest promoters up- and downstream from the peak, as well as the genes associated with these promoters. In addition, for each peak we provide a link to the SwissRegulon genome browser, displaying the peak within its genomic context, including annotations of known transcripts, transcription start clusters, known promoters, and predicted transcription factor binding sites within these promoters.

Regulatory Motif Analysis

One of Crunch's main strengths is that it performs an extensive regulatory motif analysis, with the aim of optimally explaining the observed ChIP-seq signal in terms of the occurrences of regulatory sites for both novel and known regulatory motifs. To comprehensively characterize the regulatory sites occurring within each binding peak, we use a combination of *de novo* motif finding and binding site prediction for a large collection of known regulatory motifs. Using these predictions we then find a complementary set of regulatory motifs that optimally explains the observed ChIP-seq data using a novel approach that computes alikelihood of a given motif set under an idealized model of the chromatin immuno-precipitation process. A greedy optimization algorithm is used to find the set of complementary motifs that maximizes this likelihood. In addition, we compute a number of different statistics to characterize the distribution of binding sites within peaks for all motifs within the complementary set.

To perform the motif analysis we collect the top 1000 individual binding peaks with the highest z-scores and randomly divide them into a training set $\{P_{training}\}$ of 500 peaks that we will use to find and optimize motifs, and a test set $\{P_{test}\}$ of 500 peaks that we will use to assess the performance of possible motif combinations. If there were less than 1000 significant peaks, the algorithm will use all peaks and return a warning in its output.

De Novo Motif Finding

To identify novel motifs Crunch uses PhyloGibbs (19), which implements a Bayesian model for assigning posterior probabilities to configurations of putative sites for a number of unknown regulatory motifs (with both the total number of sites and maximum number of different motifs defined by the user), and samples configurations in proportion to their likelihood using Markov Chain Monte Carlo. PhyloGibbs was specifically constructed to incorporate information from conservation patterns across orthologous genomic regions given their phylogenetic relationships. For each binding peak in the training and test sets we extract orthologous sequences from other mammalian genomes (hg19, mm9, rheMac2, canFam2, bosTau6, equCab2, monDom5) in case of data from human or mouse, or from droSim1, droYak2, droEre2, droAna3, dp4, droWil1, droVir3, droMoj3 and droGri2 in case of data from Drosophila, using UCSC's pairwise alignments and multiply align them using T-Coffee

(28) as described in (20). To enable the detection of several, potentially non-redundant, motifs we run PhyloGibbs six times with different settings: Either using phylogenetic information and multiple alignments or using only the sequence from the reference species, and searching for motifs of lengths of either 10, 15, or 20 nucleotides. In each case we are searching for two motifs simultaneously (-z 2) defining that both together are expected to have 350 binding sites within the 500 peaks of the training set. Further, we use a first order background model (–N 1). This procedure yields 12 predicted motifs, represented by position specific weight matrices (WMs).

To further optimise these WMs we use the MotEvo algorithm (20) in WM refinement mode. MotEvo uses an expectation maximization procedure to optimize a set of WMs so as to maximize the likelihood of the input sequences as a mixture of WM sites and background. Applying this procedure separately to each WM using the sequences from the training set yields 12 refined motifs.

Since PhyloGibbs searches for motifs of a predefined width, one often observes a core motif flanked by uninformative columns, i.e. columns with nucleotide frequencies matching the background frequencies. We trim all 24 motifs from both ends until a column with information content of at least 0.25 bits appears. Thus, at the end of these procedures, we have at most 24 candidate *de novo* motifs that we will subject to further analysis.

Reference Library of Known and de novo Candidate Motifs

We have collected a large library of known mammalian regulatory motifs from the literature. This library consists of the motif libraries from JASPAR (29), HOCOMOCO (30), HOMER (12), UNIPROBE (31), ENCODE (32), HTSELEX (33), and SwissRegulon (27), containing a total of 2325 motifs. For each data-set, we fuse the library of known motifs with the *de novo* motifs to form a complete set of candidate motifs that we denote $\{W_{lib}\}$.

Test Set of Pooled Binding Peaks and Background Sequences

We now aim to find a set of non-redundant regulatory motifs $\{w\}$ that jointly best distinguish the observed binding peaks from a large set of DNA sequences with similar nucleotide composition. To this end we take the sequences from $\{P_{test}\}$ and augment them with a set of 5000 background sequences $\{P_{bg,test}\}$ that we obtain by repeatedly shuffling the nucleotides in the sequences from $\{P_{test}\}$. Note that this ensures that the background sequences have the same distribution of nucleotide compositions and sequence lengths as the peaks in the test set. We denote the joint set of binding peak sequences and background sequences as the pooled test set $\{P_{pool,test}\}$. Note that we have ten times as many background sequences as binding peak sequences.

Enrichment for a Set of Motifs

To assign a performance measure to a set of motifs $\{w\}$, we calculate the probability of observing the set of ChIP-seq peaks $\{P_{test}\}$ under an idealized representation of the immuno-precipitation. We image that our set of 500 peaks resulted from immuno-precipitating 500 sequences from our pool, and we assume that the probability $P_{IP}(p|\{w\})$ for a particular sequence *p* to be immuno-precipitated is proportional to

$$P_{IP}(p|\{w\}) \propto n_{p,\{w\}} + \beta l_p \qquad (7)$$

where $n_{p,\{w\}}$ is the total number of binding sites for motifs of the set $\{w\}$ within *p*, l_p is the peak's length and *β* corresponds to the amount of non-specific binding per nucleotide. We added a non-specific binding term mainly because we observed that this strongly improves the likelihood of our model (data not shown). That such non-specific binding occurs is well supported by the known electrostatic

attraction between DNA and DNA binding proteins. As the probability of immune-precipitating any particular peak from our pool must equal one, we obtain the normalized probabilities:

$$P_{IP}(p|\{w\}) = \frac{n_{p,\{w\}} + \beta l_p}{\sum_{p' \in \{P_{pool,test}\}} n_{p',\{w\}} + \beta l_{p'}} = \frac{n_{p,\{w\}} + \beta l_p}{N_{\{w\}} + \beta L} \quad (8)$$

Where $N_{\{w\}}$ is the total number of binding sites of $\{w\}$ within $\{P_{pool,test}\}$ and *L* is the total length of all sequences in $\{P_{pool,test}\}$. We can now define the log-likelihood to observe precisely the set of binding peaks $\{P_{test}\}$ when sampling with the probabilities (8), which is given by

$$L_{IP}(\{P_{test}\}|\{w\}) = \sum_{p \in \{P_{test}\}} \log\left(\frac{n_{p,\{w\}} + \beta l_p}{N_{\{w\}} + \beta L}\right) \quad (9)$$

Although the relative values of the log-likelihood (9) can be used to optimize the motif set $\{w\}$, its absolute value depends on the specifics of our setup, i.e. the total number of sequences, and the 1:10 ratio of true peak sequences to background sequences. We apply two transformations to the log-likelihood (9) in order to obtain a score with a more general and intuitive interpretation. First, we subtract the log-likelihood of sampling the peak sequences from the pool entirely by chance. We then get the log-likelihood ratio:

$$LR_{IP}(\{P_{test}\}|\{w\}) = \sum_{p \in \{P_{test}\}} \log\left(\frac{n_{p,\{w\}} + \beta l_p}{N_{\{w\}} + \beta L}\right) + \log(F + B) \quad (10)$$

Where *F* is the number of foreground peak sequences, i.e. $F = |\{P_{test}\}|$ and *B* is the number of background sequences, i.e. $B = |\{P_{bg,test}\}|$. With $\langle n_{fg,\{w\}} \rangle$ and $\langle n_{bg,\{w\}} \rangle$ denoting the average numbers of sites of *{w}* in the foreground and background, respectively, and $\langle l \rangle$ the average length of the sequences in $\{P_{pool,test}\}$ we can rewrite this as:

$$LR_{IP}(\{P_{test}\}|\{w\}) = \sum_{p \in \{P_{test}\}} \log\left(\frac{n_{p,\{w\}} + \beta l_p}{\frac{F}{F+B}\langle n_{fg,\{w\}} \rangle + \frac{B}{F+B}\langle n_{bg,\{w\}} \rangle + \beta \langle l \rangle}\right) \quad (11)$$

Defining $\rho = F/(F + B)$ as the fraction of true ChIP-seq peaks in the pool, this can also be written as

$$LR_{IP}(\{P_{test}\}|\{w\}) = \sum_{p \in \{P_{test}\}} \log\left(\frac{n_{p,\{w\}} + \beta l_p}{\rho\langle n_{fg,\{w\}} \rangle + (1-\rho)\langle n_{bg,\{w\}} \rangle + \beta \langle l \rangle}\right) = \sum_{p \in \{P_{test}\}} \log\left(\frac{n_{p,\{w\}} + \beta l_p}{\rho\left(\langle n_{fg,\{w\}} \rangle - \langle n_{bg,\{w\}} \rangle\right) + \langle n_{bg,\{w\}} \rangle + \beta \langle l \rangle}\right) \quad (12)$$

While, for computational efficiency, we created 10 times as many background sequences as true peaks, in reality the observed binding peak sequences form a very small fraction of the entire genome that the ChIP-seq experiment is samplingfrom. A more realistic score is thus obtained when taking the limit of the fraction of true peaks in the pool going to zero:

$$\lim_{\rho \to 0} LR_{IP}(\{P_{test}\}|\{w\}) = \sum_{p \in \{P_{test}\}} \log\left(\frac{n_{p,\{w\}} + \beta l_p}{\langle n_{bg,\{w\}} \rangle + \beta \langle l \rangle}\right) \quad (13)$$

This expression gives the log-likelihood ratio of immuno-precipitating the true peak sequences $\{P_{test}\}$ from a very large pool of sequences of equal composition and length, between a model in which sequences are sampled proportional to the number of sites they contain for motifs from *{w}*, and a model in which sequences are sampled randomly. Finally, we normalize by dividing by the number of peak sequences in $\{P_{test}\}$ and exponentiating:

$$E_{\{w\}} = \exp\left[\frac{1}{F} \sum_{p \in \{P_{test}\}} \log\left(\frac{n_{p,\{w\}} + \beta l_p}{\langle n_{bg,\{w\}} \rangle + \beta \langle l \rangle}\right)\right] \quad (14)$$

The resulting `enrichment' $E_{\{w\}}$ has a simple interpretation: it measures how much more likely it is (on average) to immuno-precipitate a true binding peak as opposed to a background sequence.

Obviously, this quantity will depend on how we predict binding sites for *{w}* as well as on the choice of the parameter *β*. In the next section we discuss both thoroughly. Finally, we note that we denote the enrichment of a single motif w by E_w.

Binding site prediction and accounting for non-specific binding

To calculate the number of binding sites $n_{p,\{w\}}$ in each peak (or background) sequence *p* we use the MotEvo algorithm (20). MotEvo is a Bayesian algorithm that models the input sequences as a mixture of non-overlapping sites for the motifs from $\{w\}$ and nucleotides deriving from a background model (34) and calculates posterior probabilities of binding sites to occur for each of the motifs in $\{w\}$ at each of the positions in the sequences. MotEvo's TFBS predictions depend on a set of prior probabilities $\{\pi\}$, with π_w denoting the prior probability that a randomly chosen position on the input sequences corresponds to the start of a binding site for motif $w \in \{w\}$. For any input data-set MotEvo can be run so as to optimize the parameters $\{\pi\}$, i.e. maximizing the likelihood on the input data. To optimize the parameters $\{\pi\}$ in a way that is independent of the test data $\{P_{pool,test}\}$ we create an equivalent pool of sequences consisting of the set of training peaks and a set of background sequences $\{P_{pool,training}\} = \{P_{training}\} \cup \{P_{bg,training}\}$, where $\{P_{bg,training}\}$ is produced analogously to $\{P_{bg,test}\}$. Once we have determined the optimal prior probabilities $\{\pi\}^*$ for $\{w\}$ on the set $\{P_{pool,training}\}$ we additionally optimize the non-specific binding parameter *β* by maximizing equation (14) for this training set. We then fix the optimal priors $\{\pi\}^*$ and optimal *β** and use MotEvo with these parameters on the test pool $\{P_{pool,test}\}$ to calculate the site counts $n_{p,\{w\}}$ by summing all predicted binding site posteriors within peak p.

It is worthwhile to note that the algorithm that MotEvo employs is equivalent to a thermodynamic biophysical model in which the priors $\{\pi\}$ correspond to the concentrations of the TFs associated with the motifs in $\{w\}$ and the posterior probabilities of the sites correspond to the fraction of time the sequence is bound by the TF. In this interpretation the maximization of the priors $\{\pi\}$ corresponds to maximizing the total binding free energy of the input sequences.

Another important thing to note is that, as MotEvo only considers non-overlapping configurations of binding sites, redundant motifs compete for binding and consequently will not increase free energy of binding when added to $\{w\}$. More precisely, the sum of the optimized priors of two redundant motifs will be approximately equal to the optimized prior of one of the two redundant motifs by itself. In this way, addition of redundant motifs to the set $\{w\}$ will generally leave the binding site counts $n_{p,\{w\}}$ unchanged.

Finding an Optimal Complementary Set of Motifs

Now that we are able to compute the enrichment $E_{\{w\}}$ given an arbitrary set of motifs, we can start trying to find the optimal subset $\{W\} \subseteq \{W_{lib}\}$ that maximises $E_{\{W\}}$. As an extensive search for all subsets of all sizes of $\{W_{lib}\}$ is computationally infeasible, we use a greedy algorithm that maximizes $E_{\{W\}}$ by adding one motif to the final subset at a time.

We start by calculating the enrichment E_w for each motif in our library and sort all motifs by this score. To speed up the selection procedure, we reduce $\{W_{lib}\}$ further by removing motifs that are highly similar to a motif that is higher in this sorted list. To this end we use the inner product of two matrices as a simple similarity measure. Further details on this procedure are provided in supplementary text S7.

Using the reduced set of motifs $\{W_{reduced}\}$ we construct our final set of motifs $\{W\}$ with the following algorithm:

We initialize $\{W\}$ with the motif w_{top} which had the maximal enrichment E_w of all motifs in the library. We then iterate:

1. For every motif w_i left in $\{W_{reduced}\}$ compute $\mathrm{E}_{\{W \cup w_i\}}$.
2. Denote the motif w_i with maximal $\mathrm{E}_{\{W \cup w_i\}}$ by w_{top}.
3. If $\mathrm{E}_{\{W \cup w_{top}\}}$ increases $\mathrm{E}_{\{W\}}$ by more than 5%, add w_{top} to $\{W\}$ and go to step 1. Otherwise, terminate the algorithm.

The cut-off of at least a 5 percent increase for each added motif was chosen so as to allow even motifs that add relatively little to be incorporated, while at the same time avoiding adding redundant motifs.

Assessing Motif Quality

Besides the enrichment score E_w we use a number of additional statistics to characterize the way in which each motif from the set $\{W\}$ associates with the binding peaks. For example, for each motif we report what fraction of the binding peaks contains at least one site for the motif. In addition, although the enrichment score E_w rigorously quantifies the ability of the motif to explain the observed peaks, we also provide a more standard precision-recall curve that shows how well binding peaks can be distinguished from background sequences based on the number of predicted sites. That is, by varying a cut-off on the total number of binding sites *T* we calculate what fraction of binding peaks have a number of sites larger than *T* (sensitivity) and what fraction of all sequences with more than *T* sites are true binding peaks (precision). The precision-recall curve shows the precision as a function of the sensitivity and the overall quality of the classification is quantified by the area under the curve, which is 1.0 for a perfect classifier and 0.1 for a random classification (because the true binding peaks are 10 times rarer than background sequences).

If sites for a motif correspond directly to the binding sites for the TF that was immuno-precipitated, then one would expect the strength of the ChIP signal of a peak to correlate with the number of predicted binding sites in the peak. As another statistic, we calculate the Pearson correlation between the number of binding sites per peak and the z-value of the peak. To visualize the correlation between the ChIP-signal and the number of predicted sites, we bin all peaks by their z-value and show, for each bin, a box-plot showing the distribution of binding site numbers for peaks within the bin.

Finally, if binding sites for the motif were directly responsible for the immuno-precipitation of the fragments, then we would expect the positions of the binding sites within the peak region to co-localize with the peak of the ChIP signal. To quantify this co-localization we calculate a positional enrichment

$$EABS_w = \frac{\sum_i c(i) p_w(i)}{\bar{c} \sum_i p_w(i)} \qquad (17)$$

Here *i* runs over all predicted binding sites of w with posterior probability $p_w(i) \geq 0.2$, $c(i)$ is the read coverage at the centre of site i and $\bar{c}$ is the mean read coverage of the enriched regions from the first step of peak calling. To visualize this enrichment across binding peaks, we also show the distributions of average read coverage at predicted sites and the distribution of average read coverage across the entire peak region.

Correlations in Motif Occurrence

The set of non-redundant motifs $\{W\}$ jointly explain the binding peaks but we have so far not analysed which of these motifs tend to co-occur in the same peaks, and which motifs tend to occur in different peaks. As last step of the characterisation of TF binding we assess relative motif occurrence, i.e. all pairwise correlations in occurrence of the motifs in our selected set $\{W\}$. That is, for each pair of motifs

we calculate the Pearson correlation between the site counts across all peak regions. To visualize these correlations we create a heat map of between-motif correlation over peaks.

Consistency of Motif Sets

To compute the consistency C of two sets of motifs S1 and S2, we use the following measure:

$$C(S1,S2) = \frac{\sum_i \sum_j \delta(S1_i - S2_j) \frac{1}{2^{\max(i,j)-1}}}{\sum_i \frac{1}{2^{i-1}} + \sum_j \frac{1}{2^{j-1}}} \qquad (18)$$

Here, *i* and *j* run from 1 to the number of motifs in S1 and S2, respectively. The measure runs from 0 (no matching members) to 1 (two identical sets) and is an extension of the Dice set similarity measure to ordered sets (35).

In the above two motifs are only considered to match when they are the exact same motif. However, in our library there are many redundant motifs for the same TF that are highly similar and when we compare two sets of motifs, we want to consider such highly similar motifs as matches. We thus loosen the consistency measure as follows:

$$C(S1,S2) = \frac{\sum_i \sum_j \theta(0.2 - d(S1_i, S2_j)) \frac{1}{2^{\max(i,j)-1}}}{\sum_i \frac{1}{2^{i-1}} + \sum_j \frac{1}{2^{j-1}}}, \qquad (19)$$

where Θ(x) is the Heaviside step-function which is 0 if its argument is negative and 1 otherwise, and $d(S1_i, S2_j)$ is the distance between motif *i* in S1 and motif *j* in S2 as described in supplementary text S7. If the distance between two motifs is lower than 0.2, the motifs are considered to match.

RESULTS

For testing the performance of Crunch we analysed a large set of ChIP-seq experiments performed by the ENCODE consortium (2). We chose to run all experiments performed on cell line GM12878 and to additionally run all experiments of the Michael Snyder laboratory at the Stanford University performed on the HeLaS3 cell line. In total we analysed 128 experiments in which 93 different factors were immuno-precipitated. Full reports for all ENCODE ChIP-seq data-sets are available at: crunch.unibas.ch/ENCODE_REPORTS.

Firstly, to illustrate the results that Crunch provides after submitting a data-set, we give an overview of Crunch's analysis report for one of the samples of the ENCODE ChIP-seq experiments. Secondly, to assess Crunch's performance we compare the quality of Crunch's *de novo* motifs to a large library of motifs that we collected from other resources.

Submitting data to Crunch

One of the main strengths of Crunch is its simplicity of use. A user only needs to upload ChIP-seq data, in FASTQ, FASTA or BED format, and select the corresponding organism (human, mouse or Drosophila). Although a single immuno-precipitation dataset, i.e. a `foreground', suffices, it is strongly advisable to upload `background' samples, i.e. input DNA, as well . Optionally, more advanced users can choose to edit a number of options such as the window sizes that are used to identify enriched regions, and the cut-off on false discovery rate, although these are set to defaults that we believe should work for most datasets. Users can specify an e-mail address to get a notification when the analysis has finished.

Analysis and report overview

In a typical ChIP-seq experiment the protein that is immuno-precipitated is a DNA binding protein and the main aims of the experiment are to identify the genomic loci where the protein is binding either directly or indirectly, and the genes that are potentially regulated by these binding events. For DNA binding proteins that bind DNA in a sequence-specific manner, additional aims are to characterize the sequence specificity of the protein, and also to identify other DNA binding proteins that are co-

localizing with the immuno-precipitated factor, possibly through direct interactions. Besides these major biological aims, researchers of course also want to assess the quality of their experimental results. Crunch provides answers to all of these questions and makes its analysis results accessible through an interactive graphical web interface.

To illustrate the results that Crunch provides, we chose an experiment from the ENCODE collection, conducted by the Michael Snyder laboratory at Stanford University, where the BRCA1 protein was pulled down from GM12878 cells. Two replicate foreground (immuno-precipitation) samples as well as two replicate background (input DNA) samples were produced. The raw foreground replicates were downloaded from (36), (37) and the raw background replicates from (38), (39). The complete Crunch report for this data is available at (40).

BRCA1 is a tumour suppressor that, since it was first uncovered in 1990 (41), has been subject to extensive studies. Mutated BRCA1, together with BRCA2, is reported to be responsible for approximately two thirds of all familial breast cancer cases, whereas about 10% of all breast cancer cases are familial (42). In addition, mutations within these genes increase risk for ovarian, pancreas, uterus, cervix and prostate cancers (43). BRCA1 is involved in several different cellular pathways including DNA damage repair, cell cycle check points, centrosome duplication, transcriptional regulation as well as the immune response (42), (43). Although BRCA1 is a DNA-binding factor with a preference for binding at sites of damaged DNA, it does not bind DNA specifically in that it has a clear binding sequence motif (43), (44). However, it interacts with a number of different proteins including sequence-specific DNA binding factors and RNA polymerase II. For example, in the context of DNA damage repair, BRCA1 builds a large complex called BASC (45).

The analysis performed by Crunch is structured into three steps, as schematized in Fig. 1, and the analysis report is structured accordingly. In the first section a separate report is provided for each submitted data sample, containing a quick overview of the quality of the sample, some statistics about how its reads mapped to the reference genome, as well as the estimation of the sample's fragment size. The second section concerns the peak calling and presents different information about the called peaks. The third section discusses the results of the motif analysis as well as the characteristics of the selected motifs. In a fourth and last section, files containing the analysis results are made available for download.

Quality Control, Mapping, Fragment Size Estimation (Preprocessing)

- 30,343,607 input reads (in FASTQ format).
- 27,784,934 reads after removal of low quality reads.
- 27,713,160 reads after removal of adapters and low complexity reads.
- 25,867,852 reads after mapping.
- Report PDF

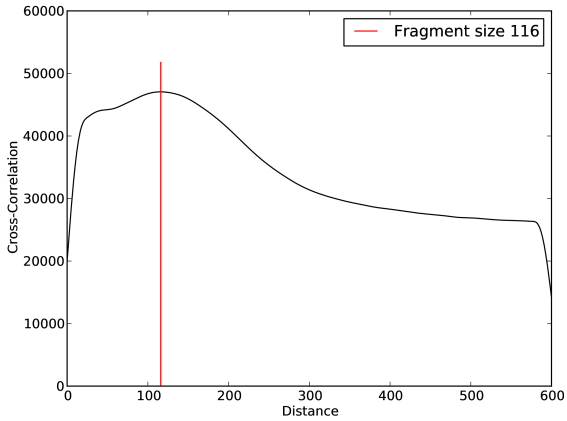


Figure 2: Summary of the quality control, mapping and fragment size estimation steps performed on the immuno-precipitation dataset (36).

Crunch produces a simple summary of the pre-processing steps of the raw ChIP-seq data that allows the user to assess the quality of the data. The bulleted list shows the initial number of reads, the number of reads remaining after the first and second rounds of quality control, as well as the final number of reads that ended up successfully mapped to the reference genome. Loss of a large number of reads

at any of these steps may indicate low data quality. A more extensive report with more detailed statistics regarding the mappings and quality control is provided as a PDF for download (see supplementary text S2).

For fragment size estimation we use a cross correlation approach (see Fig. 2 left panel). In most cases the correlation function shows a clear peak indicating that Crunch could unambiguously infer the fragment size. However, if no definite peak is visible or the estimated fragment size does not clearly match a peak in the profile, this indicates that Crunch had difficulty inferring the fragment size from the data.

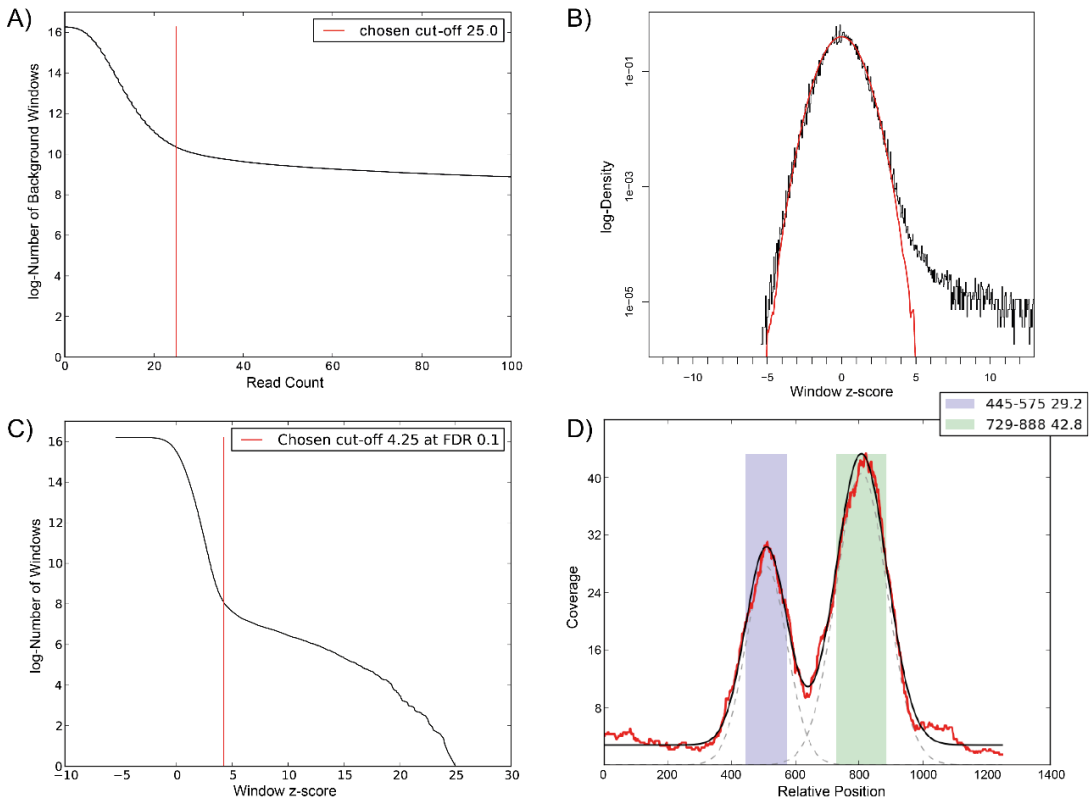
Peak Calling and Annotation

Peaks are called in a two-step procedure. First, genomic regions that are enriched for reads from the chromatin immuno-precipitation are detected, and second, individual binding events inside these enriched regions are identified. Finally, all called peaks are annotated with the closest neighbouring genes.

The main challenge in detecting enriched regions is to account properly for the noise in the data. We developed a novel noise model that models the read coverage fluctuations as a convolution of log-normal and Poisson sampling noise. We use a Bayesian mixture model to separate the genome into enriched and un-enriched regions, while fitting the parameters of the noise in parallel. According to our noise model, properly rescaled enrichment z-values should follow a standard normal distribution, and enriched regions should be evident as a tail of regions with aberrantly high z-values. As shown in Fig. 3B, we find that this pattern is indeed observed for this example dataset, confirming that our noise model correctly captures the fluctuations in read coverage across most of the genome. A bad fit of the z-value distribution to the standard normal would indicate to the user to be cautious with the interpretation of the significance levels that are reported by Crunch, and a missing right tail is an indication for a failed ChIP-seq experiment. i.e. no significantly enriched regions are detected. We set a z-value cut-off that corresponds to a default false discovery rate of 10% (Fig. 3C) and select regions with z-values over this cut-off as enriched.

For identifying individual binding peaks within the enriched regions, Crunch uses a Gaussian mixture model that decomposes each enriched region into individual binding peaks plus a uniform background (see Finding Binding Peaks within Enriched Regions, and Fig. 3D).

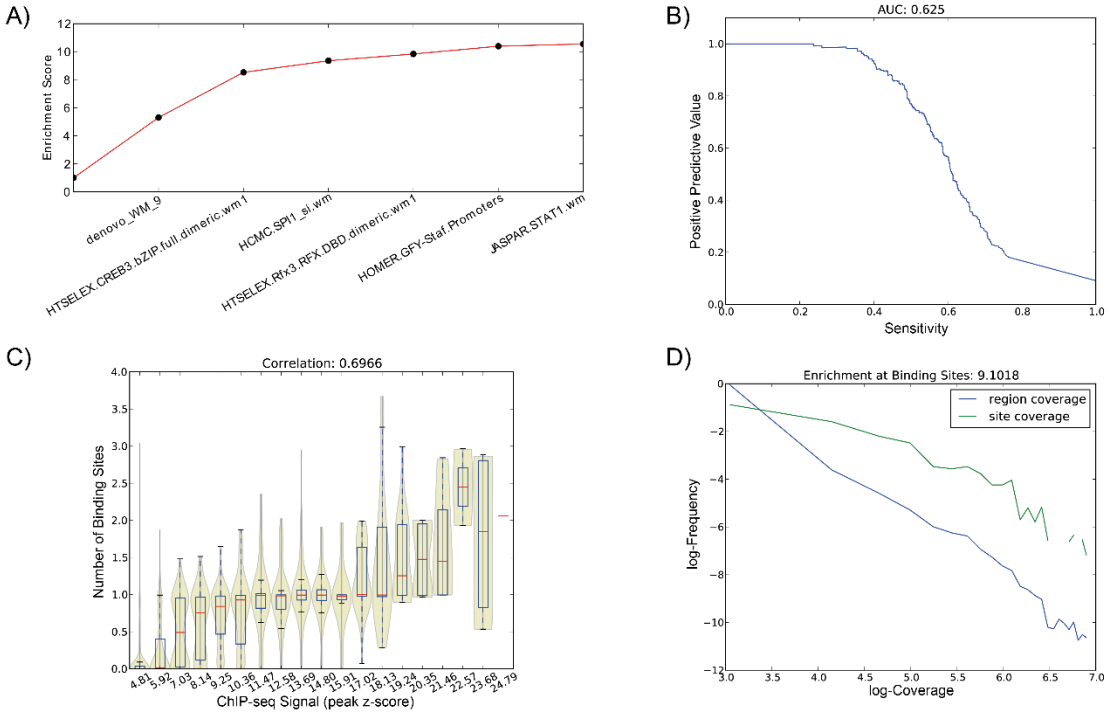
Each Gaussian of the fitted mixture model results in a candidate peak. We re-compute z-values for these individual binding peaks, and by applying the same z-value threshold as in the first peak calling step, we end up with our final set of called peaks. We annotate the peaks with the nearest genes up- and downstream and link them to the SwissRegulon genome browser with further annotations (see section Annotating Peaks to Promoters and Genes, and Fig. 3E). Within each peak we additionally predict binding sites for every motif of the final complementary set of motifs (see below).



E)

Coordinates	Z-score	Nearest Genes on the Left	Offset of Nearest TSS on the Left (Strand)	Nearest Genes on the Right	Offset of Nearest TSS on the Right (Strand)
chr12:123237288..123237405	25.350	HCAR1|hydroxycarboxylic acid receptor 1	-22341 (-)	DENR|density-regulated protein	45 (+)
chr19:49999397..49999550	24.750	RPL13A|ribosomal protein L13a	-8590 (+)	RPS11|ribosomal protein S11	174 (+)
chr19:4124121..4124230	23.907	MAP2K2|mitogen-activated protein kinase kinase 2	-42 (-)	CREB3L3|cAMP responsive element binding protein 3-like 3	29492 (+)
chr17:37617567..37617708	23.901	MED1|mediator complex subunit 1	-10216 (-)	CDK12|cyclin-dependent kinase 12	115 (+)
chr15:22833360..22833493	23.199	GOLGA6L1|golgin A6 family-like 1	-97181 (+)	TUBGCP5|tubulin, gamma complex associated protein 5	44 (+)
chr14:102414455..102414575	23.152	PPP2R5C|protein phosphatase 2, regulatory subunit B', gamma	-138206 (+)	DYNC1H1|dynein, cytoplasmic 1, heavy chain 1	16313 (+)
chr17:4843503..4843635	22.744	SLC25A11|solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11	-237 (-)	RNF167|ring finger protein 167	43 (+)
chr3:49131487..49131599	22.125	QRICH1|glutamine-rich 1	-34 (-)	QARS|glutaminyl-tRNA synthetase	10637 (-)
chr9:98637814..98637949	21.625	LINC00476|long intergenic non-protein coding RNA 476	-28 (-)	C9orf102|chromosome 9 open reading frame 102	60 (+)
chr9:6413101..6413214	21.588	UHRF2|ubiquitin-like with PHD and ring finger domains 2	-13 (+)	UHRF2|ubiquitin-like with PHD and ring finger domains 2	40 (+)

Figure 3: A) Reverse cumulative distribution of the summed read counts from the background samples (38) and (39) in genome-wide sliding windows of 2000 nt. Regions with aberrantly high coverage in the background sample are removed from further analysis. B) Distribution of the z-scores from all genome-wide sliding widows (in black) and a reference standard normal distribution (in red). C) Reverse cumulative distribution of the same z-scores (in black) as well as the z-score threshold used in both steps of peak calling (in red), which is 4.25 for this example, and corresponds to an FDR of 10%. Note that the vertical axes in both B and C are in log-scale. D) High resolution ChIP-seq coverage profile of an individual enriched region (red), together with the fitted mixture model (black). The two dashed lines in light grey show the individual Gaussians used in the mixture model of this particular coverage profile. The lightly coloured bars highlight the two individual binding peaks for which the relative start and end coordinates (first two numbers) as well as the amplitudes (last number) are listed in the legend. E) Table with the top 10 called peaks of the BRCA1 experiment. Each peak is annotated with its coordinates on the genome, its z-value, its nearest genes and the offsets to its nearest transcription start sites, both toward higher and lower genomic coordinates. Through the “Select motif” button the user can add columns containing predicted binding sites for each of the selected motifs (see next section).



E)

Motif Name	Sequence Logo	Motif Ensemble Enrichment (Motif Ensemble log-Likelihood Ratio)	Enrichment (log-Likelihood Ratio)	Precision and Recall	Prediction - Observation Correlation	Enrichment at Binding Sites	Number of Positively Predicted Peaks
denovo_WM_9		5.302 (834.062)	5.302 (834.062)	0.6250	0.6966	9.1018	652/1348
HTSELEX.CREB3.bZIP.full.dimeric.wm1		8.526 (1071.539)	1.491 (199.847)	0.2132	-0.0513	2.3533	129/1348
HCMC.SPI1_si.wm		9.355 (1117.973)	1.065 (31.691)	0.1508	0.0393	1.5312	15/1348
HTSELEX.Rfx3.RFX.DBD.dimeric.wm1		9.837 (1143.069)	1.051 (24.91)	0.1041	-0.0236	0.9579	3/1348
HOMER.GFY-Staf.Promoters		10.393 (1170.567)	1.226 (101.989)	0.1891	-0.021	1.2423	1334/1348
JASPAR.STAT1.wm		10.547 (1177.931)	1.039 (19.247)	0.1327	-0.0009	1.4645	18/1348

F)

Motif Name	Distance to Query Motif	Logo
HOMER.GFX(?).Promoter	0.033	
ENCODE.UA1	0.034	
HOMER.ZBTB33.GM12878-ZBTB33-ChIP-Seq	0.063	

G)

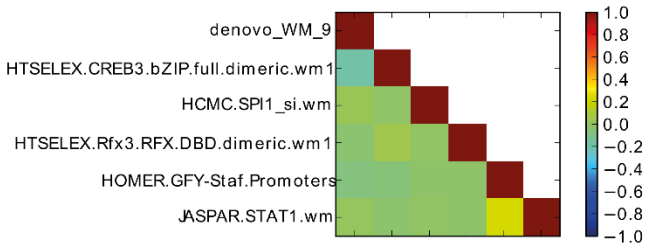


Figure 4: A) Increase of the enrichment score as motifs are added to the complementary set of motifs, showing the contribution of each of the selected motifs to explaining the Chip-seq peaks. B)-D) Performance statistics of the top motif `denovo_WM_9', i.e. a motif that was identified de novo in this case. B) Precision and recall curve. The area under the

precision-recall curve (AUC) is 0.625. For an optimal classifier AUC equals to 1. The curve indicates that occurrence of denovo_WM_9 accurately identifies roughly 30-40% of the true binding peaks, but does accurately identify the remaining 60%. Given that the BRCA1 is not itself binding to the denovo_WM_9 motif, but a different associated protein instead, this might not be too surprising as BRCA1 is likely to have different binding partners at different peaks. C) Correlation between the ChIP-signal (peak z-values) and the number of predicted binding sites for denovo_WM_9 sites within every peak. The (binned) correlation is quantified by the Pearson correlation coefficient (0.6966). The curve shows that low z-value peaks tend to have no predicted binding sites, whereas peaks with z-values of 8 and higher nearly always have a predicted binding site. Peaks with very high z-values even tend to have multiple binding sites. D) Illustration of how the coverage of ChIP-signal at predicted denovo_WM_9 sites compares to the overall coverage at all significantly enriched regions. The green curve is the histogram of the coverages i.e. the number of overlapping fragments, at the centers of all predicted denovo_WM_9 sites. The blue curve is a histogram of the coverage at every position of all significantly enriched regions. The coverage at the predicted denovo_WM_9 sites relative to the overall coverage is quantified as Enrichment at Binding Sites which here is 9.1018. E) Summary statistics for all motifs in the selected set of complementary motifs, sorted in the order that motifs were added to the set. The columns show motif name, sequence logo, the enrichment of the set up to and including this motif, the motif's enrichment in isolation, its precision-recall, its correlation with peak z-value, enrichment of ChIP-seq coverage at its sites, and the fraction of peaks in which the motif occurs. F) The top three motifs in the library of known motifs with most similarity to denovo_WM_9. G) Heatmap of pairwise correlations between the occurrence of binding sites for all selected motifs across all binding peaks.

A Complementary Set of Motifs jointly explaining the observed peak sequences

A novel feature of Crunch is that it performs an extensive regulatory motif analysis. To comprehensively characterize the regulatory sites occurring within each binding peak, we use a combination of *de novo* motif finding and binding site prediction for a large collection of known regulatory motifs. Using these predictions we then find a complementary set of regulatory motifs that optimally explains the observed ChIP-seq data using a novel Bayesian approach that computes an enrichment which models the process of immuno-precipitation itself.

If the protein that was immuno-precipitated is a sequence-specific DNA-binding protein, then we might expect that he motif describing the binding specificity of this protein should suffice to explain the observed binding peaks. Indeed, we find that for some factors one motif is enough to explain all the data (CTCF, for example). However, for the majority of datasets we find that a set of complementary motifs can explain the data better. These additional motifs can be hypothesized to describe co-factors of the TF that was immuno-precipitated.

To characterize the contribution of the complementary motifs, we compute three performance measures, in addition to the enrichment measure (see Material and Methods). First, the precision and recall analysis measures how well a motif can classify sequences as binding peaks based on the number of occurring sites (Fig. 4B). The correlation analysis between predicted (motif occurrence) and observed (ChIP-seq z-score) signal measures to what extent a motif can predict the strength of the binding peak (Fig. 4C). Finally, we measure the extent to which the occurrence of the binding sites within the peak regions corresponds to the areas of highest ChIP-seq signal (Fig. 4D).

Fig. 4E summarizes all information collected about the complementary motif set, including motif name, sequence logo, enrichment score of the whole set and each motif separately, precision-recall score, prediction-observation correlation, enrichment at binding sites, as well as the fraction of all binding peaks that have predicted sites for the motif.

To help annotate motifs, every motif is compared to all known motifs in our library. In Fig. 4F we see the 3 motifs from our library that are most similar to the top motif of our complementary motif set, i.e. the top motif in Fig. 4E. These annotations can be viewed on the motif page that is linked through the motif name in Fig. 4E. This motif page contains all information collected on the motif, including plots of the motif statistics (Fig. 4B-D), the weight matrix file as well as a PDF containing additional statistics.

Although he statistics in Fig. 4E suggest that the denovo_WM_9 motif best represents the binding preferences of BRCA1, it is known that BRCA1 does not bind DNA sequence-specifically and, therefore, denovo_WM_9 is thus most probably not the motif describing the binding of BRCA1 itself. Previous attempts to identify the protein that is binding to sites for this motif were unsuccessful, but it was found that a protein complex is binding the motif (46). Furthermore, this motif was found to occur primarily in TATA-less promoters and to be involved in the regulation of the expression of about 5% of all human genes, in particular genes involved in protein synthesis and cell cycle regulation, especially in the G1- to S-phase transition (47). Interestingly, BRCA1's action in cell cycle regulation has also been reported to be involved in the same transition (48).

Crunch finds that the closest known motif is HOMER.ZBTB33.GM12878-ZBTB33-ChIP-Seq from the Homer library (Fig. 4F). ZBTB33 is the gene encoding the KAISO protein which is a sequence-specific TF that has been associated with breast cancer and also especially with BRCA1-related breast cancer (49). We thus hypothesize that KAISO is an interaction partner of BRCA1 and that the denovo_WM_9 motif describes the binding specificity of the KAISO TF. The same hypothesis was also put forward in (32).

Besides denovo_WM_9, Crunch finds five additional motifs that substantially increase the enrichment score and have binding sites within the binding peaks (Fig. 4A and 4E). The most significant of these motifs is a motif associated with the TF CREB3, HTSELEX.CREB3.bZIP.full.dimeric.wm1 (33). Crunch further annotates this motif with ATF and JUN family proteins. BRCA1 is well known to directly bind to CREB-binding protein (CBP) (50) as well as to directly bind to ATF1 (43) and JUN proteins (51). We thus hypothesize that a complex of BRCA1-CBP-CREB, a complex of BRCA1-ATF1 or a complex of BRCA1-JUN is binding to a subset of our BRCA1 binding peaks.

For the next two motifs, SPI1/PU1 and RFX3, we could not find any support in the literature. Although both together bind only 18 of our peaks, it might still be interesting to further investigate these two motifs in relation to BRCA1. GFY-staf was found in chromatin regions bound by H3K4me3 which are specifically open in breast cancer cells (52). And finally, we find the STAT1 motif, a TF reported to directly bind BRCA1 (43).

To give further insight into the binding of the inferred complementary motifs, we report the pairwise correlation between site predictions of all motifs across the binding peaks (Fig. 4G). As we were selecting a non-redundant set of motifs we do not expect any highly positive correlations, which is true for our example case. Nevertheless, these correlations can reveal interesting features of relative TF binding, e.g. whether motifs co-occur or whether they bind exclusively.

Downloads

In the last section of the analysis report several files with analysis results are provided for download. These include, for each submitted data sample, a BEDWEIGHT and a WIG file containing mapped reads that can be used for visualization and further processing. In addition, a file with all binding peaks and all annotation information provided by Crunch for each peak is also provided for download. In particular, this file contains for each peak, its genomic coordinates, its z-value, the three nearest flanking promoters on both sides, the genes associated with these promoters, as well as predicted binding sites for all selected motifs within the peak.

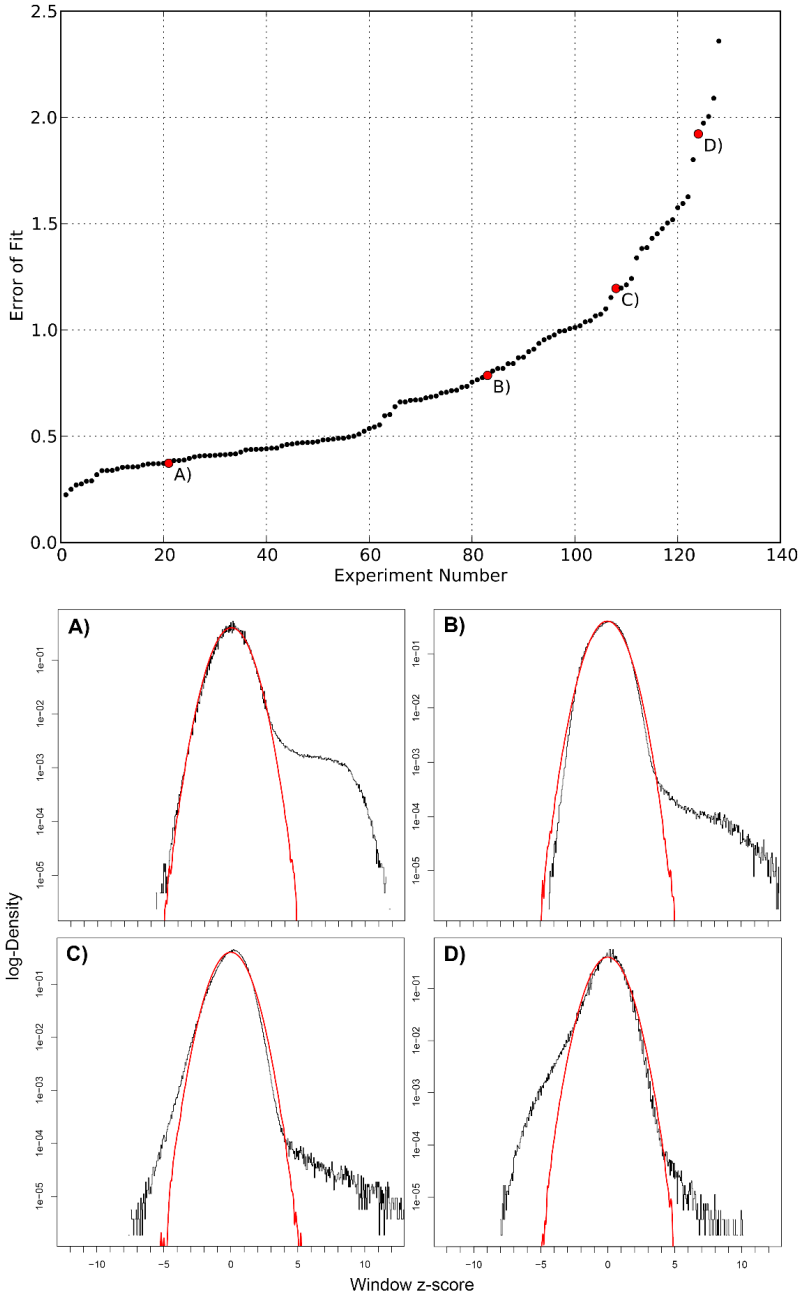


Figure 5: Genome-wide fluctuations in ChIP-seq signal fit Crunch's noise model. The top panel shows a cumulative distribution of the error of our noise model fits, which is defined as the root of the mean squared deviation of the z-value distribution, from the expected normal distribution, in the range of z-values between -5 and 3. For four experiments (red dots A, B, C and D) we show the actual z-value distributions (black) and the expected standard normals (red) in the panels below.

Crunch's noise model accurately reflects noise levels

Assessing the quality of ChIP-seq analysis tools is a non-obvious task. First, one could ask to what extent the identified binding peaks are `correct'. In the absence of any independent ground truth, this question boils down to an assessment of the correctness of the statistical model used to detect significant enrichment of the ChIP-seq signal relative the background. In the BRCA1 example above, we saw that the distribution of z-values inferred by our noise model accurately tracked the expected standard normal distribution, supporting that our noise model correctly captures the statistics of ChIP-seq coverage fluctuations across the genome. We find that this is the rule. That is, when going through Crunch's reports of all 128 analysed ENCODE experiments, we find that, with the exception of a few datasets, the distribution of z-values matches the standard normal predicted by our noise model (Fig. 5). This match between the assumed noise model and the observed fluctuations across

the majority of the genome supports the statistical significance assigned to the enriched regions. As far as we are aware, Crunch is the only tool for which the noise model is explicitly supported by the data.

Crunch's de novo motifs often outperform literature motifs

A second deliverable of ChIP-seq analysis is the analysis of sequence motifs occurring within the peaks. As discussed in the Regulatory Motif Analysis section, Crunch determines a complementary set of motifs that is enriched in the ChIP-seq peaks. If the immuno-precipitated protein binds the DNA in a sequence specific manner, this set of motifs likely contains the motif that describes the sequence binding specificity of the immuno-precipitated protein, as well as some motifs for factors that interact with this protein. As we usually do not know the set of motifs that is interacting with the immuno-precipitated protein, it is difficult to assess Crunch's performance based on the entire set of motifs.

For quite a number of TFs, there are several motifs from different resources that are supposed to describe the sequence-specificity of the TF. Since we are including motifs from these resources in the motif selection procedure, they are in direct competition for explaining the ChIP-seq binding data with Crunch's *de novo* motifs. We can thus assess Crunch's ability to generate high quality *de novo* motifs for a TF from its ChIP-seq data, by comparing the performance of these de novo motifs with those of known motifs from existing resources.

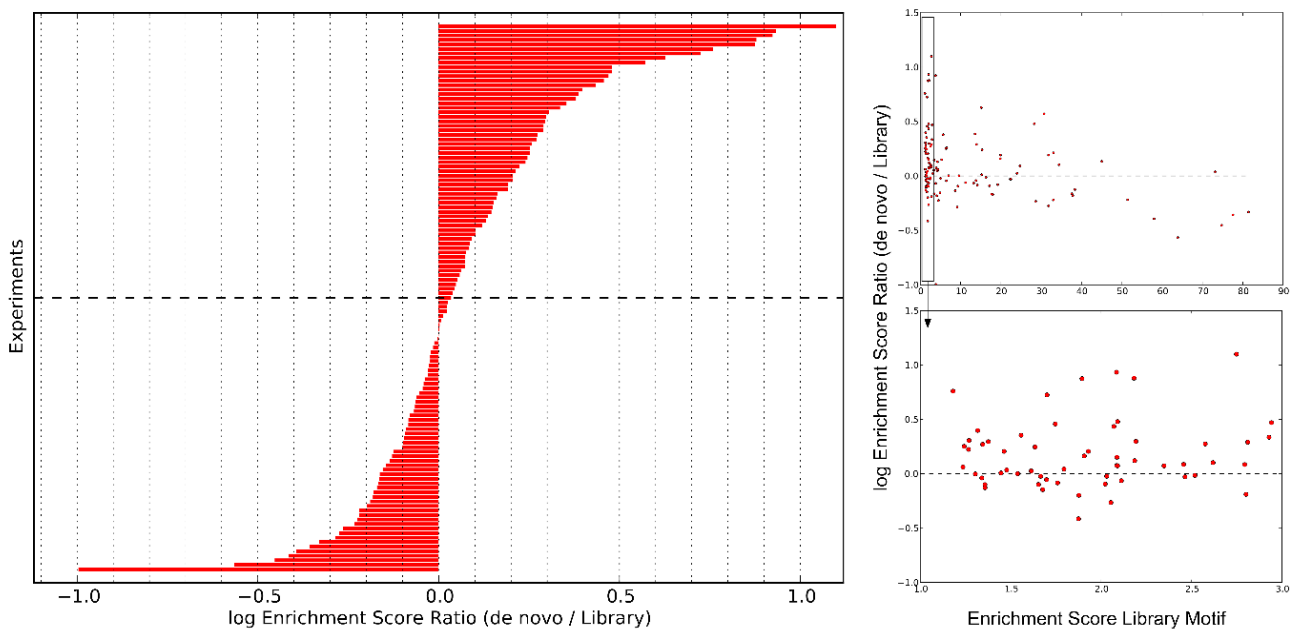


Figure 6: The left panel shows a bar plot of the log-transformed ratios between the enrichment score of the top *de novo* motif and the enrichment score of the top literature motif, for all 121 experiments. A positive log-ratio means that Crunch's *de novo* motif outperformed all library motifs. The panels on the right show the logarithm of the ratio of the enrichment score of the best *de novo* motif and the enrichment score of the best library motif, as a function of the enrichment score of the best library motif. The bottom panel shows a zoom in of the boxed region on the left in the top panel.

To this end we compared, for each of our set of ENCODE experiments, the enrichment score for the top *de novo* motif with the enrichment score of the top motif from our set of libraries. From the initial

128 ENCODE experiments we threw out 7 experiments where less than 200 peaks were called. From the remaining 121 experiments we found that for 69 experiments Crunch's *de novo* motifs performed better, while for 52 experiments a motif from our library performed better (Fig. 6). Crunch not only provides a superior motif for the majority of the experiments, it also often provides a motif that performs substantially better than any existing motif. That is, because the enrichment score is calculated per sequence, and there are 500 sequences in our test set, a log-ratio of 0.2 in enrichment score corresponds to a likelihood ratio of e^{100}. For 35 *de novo* motifs the log-ratio of the enrichment score is larger than 0.2.

In principle, as long as we can identify a motif that is capable of explaining our ChIP-seq data well, we do not care whether it is coming from a library or from Crunch. In cases where there is no known motif that explains the data well, though, we would like Crunch to find a *de novo* motif that is capable of doing that. We thus looked at the performance of Crunch's *de novo* motif relative to the best library motif as a function of the performance of the library motif itself (Fig. 6, right panel). The bottom right panel of Fig. 6 shows the 57 cases for which the best library motif has a enrichment score of 3 or lower. In 39 of these cases, Crunch finds a better motif, and in 24 of these cases there is a substantial improvement of 0.2 or more.

We observe two types of DNA binding factors: solitary binders and co-binders

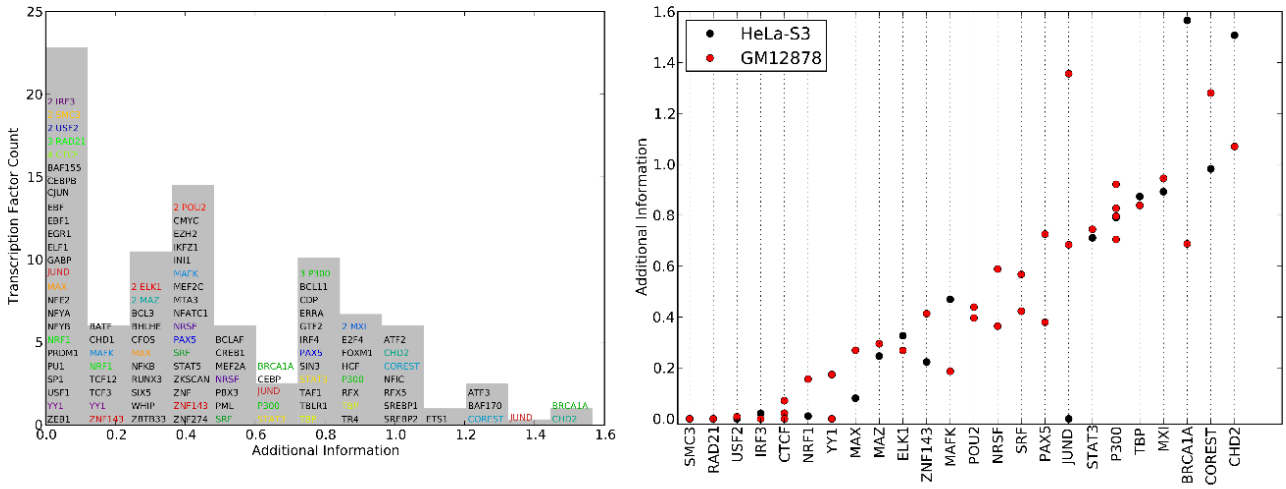


Figure 7: Left panel: Histogram of the additional information for all 121 complementary motif sets from the ENCODE data. The vertical axis shows the number of TFs in a certain additional information bin. For the TFs with names in black, there is only a single data, while for TFs with colored font there are multiple datasets. In case there is a number in front of the TF name it indicates the number of experiments for that factor within the same bin. Right panel: Additional information computed from separate experiments of 24 TFs. The horizontal axis lists the 24 factors for which we have multiple experiments. The vertical axis shows the additional information from each separate experiment. TFs are sorted by their mean additional information.

For TFs that specifically bind the DNA, one might expect that the binding peaks could be explained by a single motif. Indeed, for about one quarter of all ENCODE datasets, the complementary motif set contains only a single motif. It is conceivable that, for the remaining cases, the additional motifs also

contribute only marginally to explaining the binding peaks. To investigate this we calculated, for each dataset with multiple motifs in the complementary set, the difference in log enrichment of the entire set, and the log enrichment of the top motif. The resulting quantity can be interpreted as the amount of information the complementary motif set contains about the ChIP-seq data, excluding the top motif. We call this quantity additional information.

When looking at the histogram of the additional information of all ENCODE experiments we observe a wide spread, from experiments having no additional information to experiments where the additional information goes as high as 1.5, meaning the additional motifs increase the likelihood of explaining an individual peak by almost 4.5 times (Fig. 7). Furthermore, we observe that for a given TF, the additional information measures from different experiments tend to group, i.e. names of the same colour in the left panel of Fig. 7 tend to co-localize. This suggest that, for many TFs, there is substantial information in the additional motifs and that, moreover,the extent to which a TF acts in cooperation with other factors seems to be conserved across experiments and cell lines (Fig. 7, right panel).

The main exception to this observation is JUND, where, across three different experiments, we observed no additional information, moderate additional information, and high additional information. Notably, the experiment showing no additional information was performed on HeLa S3 cells, whereas the other two were performed on the GM12878 cell line.

Inferred complementary motif sets are consistent over replicate experiments

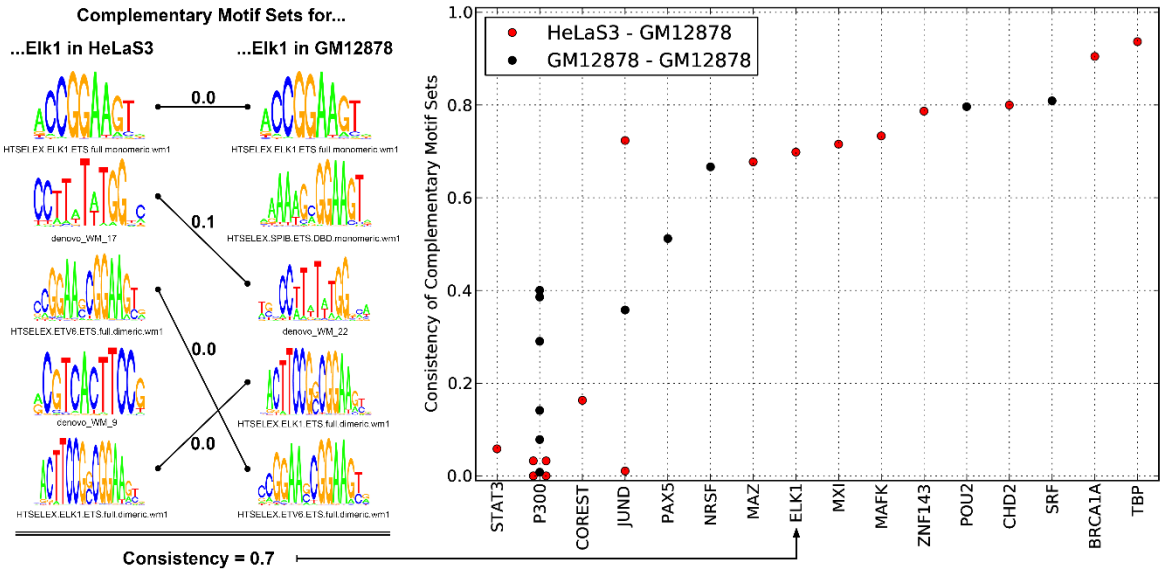


Figure 8: Left panel: The complementary motif sets for two different experiments with Elk1, one performed in HeLaS3 and the other in GM12878 cell line. Both sets contain another unpaired motif at position six which is not shown here. Similar motifs are connected by lines, and the distance between motifs is indicated. For this example we find a consistency of 0.7 in the motif sets (see Materials and Methods). The right panel shows the consistencies of all 16 co-binding TFs, i.e. TFs with an average additional information of 0.2 or greater. The colouring of the dots indicate the cell line in which the experiment was performend.

Next, we investigated to what extent the complementary motif set is consistent across different experiments with the same TF. For this we considered all TFs with an average additional information of 0.2 and greater. We call these TFs co-binding TFs, as opposed to TFs that bind

predominantly by themselves. We deviced a consistency score for a pair of motif sets that quantifies to what extent highly similar motifs occur in the ame order in both sets (Fig. 8 and Materials and Methods). Strinkingly, the majority of the 16 tested TFs shows a substantial consistency between the inferred motif sets from two separate experiments. This suggests that the complementary motif set is a inherent characteristic of the TF that is immuno-precipitated that is independent of the exact data set used. Moreover, it suggests that Crunch can successfully identify such co-binding motif sets for a TF.

Weight matrices are realistic models of transcription factor binding

Finally, for solitary binders one motif is enough to explain the ChIP-seq data. Put differently, the explanatory power of additional motifs is very low. If the single motif for a solitary binder captures an intrinsic property of the TF, then one would expect the same or highly similar motifs to be inferred across different experiments for the same TF. To investigate this, we checked whether the motifs from ourlibrary rank consistently across pairs of experiments for the same TF. Briefly, we took from every experiment the five library motifs (excluding Crunch's *de novo* motifs) with highest enrichment score and computed the consistency between these two sets of five motifs (Fig. 9 and Materials and Methods). We find that, of the eight TFs tested, only the experiments on USF2 disagreed on the order of ranking of motifs. All others showed very high to perfect consistency, even preserving the relative order of highly similar motifs. This suggests that motifs may be accurately capturing the binding specificity of a TF that is independent of the details of the experiment. This is especially attractive because it suggests that we can unambiguously determine which motif best describes the sequence specificity of a solitary binding TF.

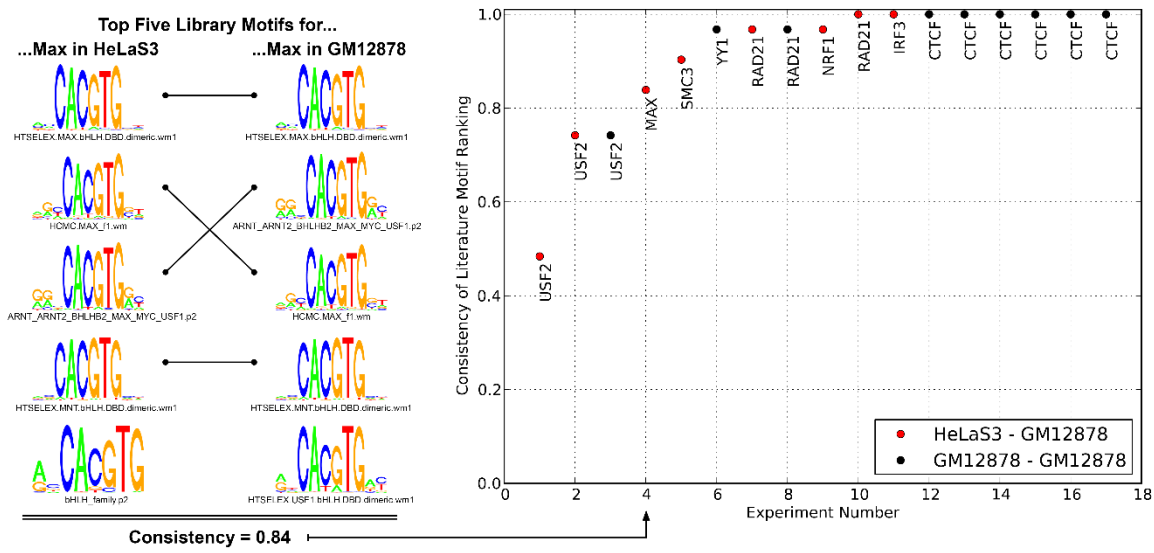


Figure 9: Left panel: The top five library motifs for two Max experiments, one performed in HeLaS3 and the other in GM12878 cell line. Identical motifs are connected. For this example we find a consistency of 0.84 (see Materials and Methods). Right panel: Consistencies between the top five library motifs for all TFs with an average additional information below 0.2.

DISCUSSION

Crunch is a new integrated ChIP-seq analysis procedure, introducing various novel features. First, running Crunch is extremely easy. As Crunch is implemented as a web server (at crunch.unibas.ch) users are only required to upload their data files and to specify the model organism from which their data derive. In contrast to most existing ChIP-seq analysis software, there is no understanding required about the function of the single analysis steps. or how they need to be connected to one another. Furthermore, no bioinformatics expertise is assumed. allowing also biological researchers an eay avenue for analyzing their own ChIP-seq data. Finally, no high performance computational hardware is needed for the processing of these large ChIP-seq data-sets.

Second, Crunch presents its analysis results in an easy-to-interpret manner with the help of an interactive graphical web interface. Thanks to its slim design, the user quickly gains an overview of the main analysis results including a quality assessment of the data, the predicted binding peaks, the peaks' annotations to nearest genes and transcription start sites, and the complementary motifs enriched within these peaks. The extensive statistics and quality reports that Crunch provides allow users to obtain comprehensive insight into the quality and trustworthiness of the results. For studying the results in more depth, Crunch also provides links to the SwissRegulon genome browser with annotations for the called peaks, including predicted transcription factor binding sites, transcripts, transcription start clusters and promoters (27). Next, Crunch provides for every enriched motif a separate web page with a comprehensive discussion of the motif's binding characteristics with respect to the submitted ChIP-seq data as well as an identification of proteins that are putatively binding the motif. For further user-specific analysis, flat files are provided with all data inferred and collected by Crunch. Besides Crunch's ease of use, user-friendly, and comprehensive reports, Crunch also incorporates several novel ChIP-seq analysis procedures. In particular, Crunch introduces a novel peak calling procedure as well as a novel procedure for determining complementary sets of enriched motifs, i.e. motifs that jointly represent the binding specificity of the immuno-precipitated protein, and that provide evidence for possible interaction partners of the immuno-precipitated protein. The peak calling method employs a novel noise model for finding ChIP-seq signal enriched genomic regions that is based on our previous work on the noise characteristics of next-generation sequencing data (18). It models the fluctuations in read counts as a combination of multiplicative and Poisson sampling noise and its parameters are fitted to the data by a Bayesian procedure. In addition, we decompose the ChIP-signal within the significantly enriched genomic regions into individual binding peaks using Gaussian mixture modelling. In contrast to other peak finders, our data provide direct evidence in support of the noise model that Crunch employs.

Another major novel feature of Crunch is its comprehensive regulatory motif analysis. Using our previously developed PhyloGibbs (19) and MotEvo (20) algorithms, we perform *de novo* motif finding followed by binding site predictions for these *de novo* motifs plus motifs from a large library consisting of motifs from JASPAR (29), HOCOMOCO (30), HOMER (12), UNIPROBE (31), ENCODE (32), HTSELEX (33) and SwissRegulon (27). Subsequently, Crunch determines the set of motifs whose predicted sites model the ChIP-seq signal best. Crunch also assesses the performance and binding characteristics of each motif of this selected set by computing a number of different statistics. We showed that, for the majority of datasets, Crunch's *de novo* motifs are superior in explaining the ChIP-seq data as compared to the motifs from the other seven libraries. We also

demonstrated that the inferred set of motifs is consistent across datasets for the same TF. Finally, for solitary binding TFs we find that the top motif is highly reproducible across different datasets, suggesting that an optimal motif representing the binding specificity of the TF can be unambiguously determined.

Acknowledgments

This research was supported by SystemsX.ch through the CellPlasticity project grant.

SUPPLEMENTARY MATERIALS

S1: 3' adaptor trimming

Since already a short adapter subsequence at the 3' end of a read can corrupt subsequent mapping to a reference genome we stringently trim adapters from reads using the following three steps: First, reads are scanned for full length adapter sequence matches allowing for 2 mismatches. All matched reads then get truncated starting at the beginning of the adapter sequence read match. Second, adapter sequence prefixes get matched to read suffixes allowing for 1 mismatch for matches longer than 6 nt and 2 mismatches for matches longer than 9 nt. All matched read suffixes are then removed.

S2: Read mapping statistics

Crunch provides 3 figures with information about the read mappings.

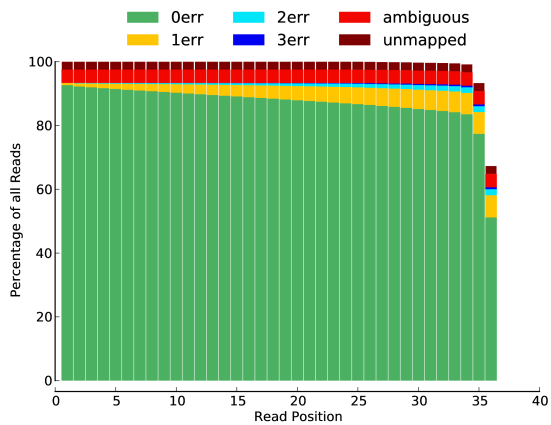


Figure S1: Distribution of mapping errors as a function of considered read length. As an example, when considering reads up to read position 20, 85-90 % of the reads had no mapping errors (green bar), about 5% had one error (yellow bar), and less than 1% had two or three errors (light and deep blue bars). About 4% were mapped to more than 100 locations (light red bar) and about 2% were not mapped at all (deep red bar).

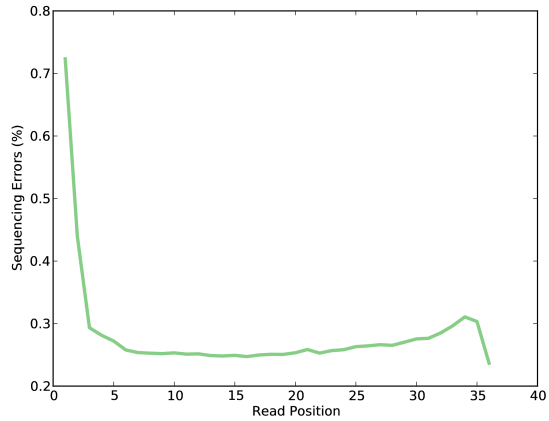


Figure S2: Percentage of all reads that have a mapping (or sequencing) error as a function of read position.

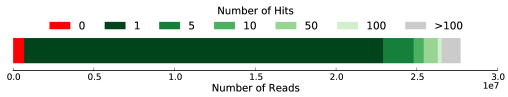


Figure S3: Cumulative bar plot showing the number of reads with a certain number of mapping locations (hits).

S3: Binding peak widths

Theoretically, if all fragments were exactly the same length, if the breaking of the DNA during library preparation were completely random, and if the probability to immuno-precipitate a fragment were independent of where in the fragment the protein was bound, then each single binding site would result in an isosceles triangular coverage distribution with a base of two times fragment size around the actual binding site. However, fragment sizes will fluctuate, the propensity for the DNA to break will vary along the genome, and the probability to

immuno-precipitate a fragment may depend on the relative position at which the protein is bound. This results in approximately Gaussian shaped distributions of read coverage (Figure 5).

As discussed above, all individual binding peaks in the coverage profiles should have the similar width as it only depends on the fragment size used in the experiment. By constraining the widths of the fitted Gaussians we can therefore help the mixture model to detect true binding peaks. For this we examined the widths, i.e. the σ's, of the Gaussian shaped distributions in all significantly enriched regions from the first step of peak calling from all our 123 ENCODE ChIP-seq data sets (see Results section) where fragment sizes range from 82 to 198 nucleotides. We observed that peak widths on average indeed scale linearly with fragment size (Fig. S4).

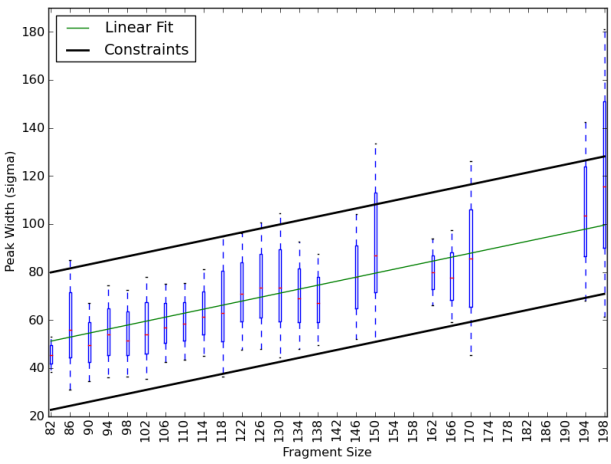


Figure S4: The boxplots in blue and red show the distribution of peak widths from our 123 ENCODE ChIP-seq experiments stratified by the experiment's estimated fragment size. The green line shows a linear fit to the data that was determined by linear regression. The bold black lines indicate the boundaries used to constrain peak widths during mixture modelling.

By performing simple least squares linear regression on these data we determined that the average peak width scales as 0.416 * fragment size + 17.1 with a residual variance of 0.887 and a standard deviation of the residuals of 19. Using this we conservatively constrain peak widths by adding and subtracting 1.5*19 for upper and lower constraints, respectively. This way we capture more than 90% of the peak widths from the 123 experiments. Moreover, note that the scaling is according to our theoretical expectations. Assuming that the hypothesized isosceles triangular distribution and the observed Gaussian distribution share the same width at half height, we expect the σ of a Gaussian to scale as $\left(2\sqrt{2\log(2)}\right)^{-1}$* fragment size, i.e. 0.425 * fragment size, which is very close to the 0.416 * fragment size that we fit above. We hypothesize that the nonzero vertical offset might result from the fact that the binding of proteins to the binding site may suppress DNA breakage at these points, thereby slightly increasing the fragment size around binding sites relative to the average fragment size genome-wide.

S4: Motif similarity measure

The similarity $S(w_1, w_2)$ between motifs $w_1$and w_2 is the inner product at their optimal alignment, where the optimal alignment is defined by a shift s that maximizes the inner product $I(s, w_1, w_2)$ and by the orientation of w_2, i.e. also considering the reverse complement of w_2 $w_{2,rc}$.

$$I(s, w_1, w_2) = \sum_i w_1(i) w_2(i - s)$$

$$I(s, w_1, w_{2,rc}) = \sum_i w_1(i) w_{2,rc}(i - s)$$

$$S(w_1, w_2) = \max\left[\max_s I(s, w_1, w_2), \max_s I(s, w_1, w_{2,rc})\right]$$

Here *i* runs over all overlapping motif columns. We further normalize $S(w_1, w_2)$ to a number between 0 and 1 and subtract it from 1 to get a dissimilarity measure

$$D(w_1, w_2) = 1 - \frac{2S(w_1, w_2)}{S(w_1, w_1) + S(w_2, w_2)}$$

To reduce $\{W_{lib}\}$ we proceed as follows: We move the highest quality motif w_{top} of $\{W_{lib}\}$, i.e. $w_{top} = max\ (E_w\ \forall\ w \in \{W_{lib}\})$, to a new set $\{W_{reduced}\}$ and remove all motifs w from $\{W_{lib}\}$ with $D(w_{top}, w) < 0.2$. The distance threshold of 0.2 is chosen such that only very close motifs get removed. This procedure is then repeated until no motif is left in $\{W_{lib}\}$.

REFERENCES

1. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–502.

2. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

3. Schbath,S., Martin,V., Zytnicki,M., Fayolle,J., Loux,V. and Gibrat,J.-F. (2012) Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *J. Comput. Biol.*, **19**, 796–813.

4. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.

5. Das,M.K. and Dai,H.-K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8 Suppl 7**, S21.

6. Strand Life Sciences Pvt. Ltd. (2012) Avadis® NGS.

7. Kallio,M.A., Tuimala,J.T., Hupponen,T., Klemelä,P., Gentile,M., Scheinin,I., Koski,M., Käki,J. and Korpelainen,E.I. (2011) Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, **12**, 507.

8. CLC Bio Genomics Workbench.

9. Genomatix Mining Station.

10. Astridbio GenoMiner.

11. Partek Inc. (2008) Partek® Genomics Suite.

12. Heinz,S., Benner,C., Spann,N. and Bertolino,E. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.

13. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotech*, **26**, 1293–1300.

14. Ye,T., Krebs,A.R., Choukrallah,M.-A., Keime,C., Plewniak,F., Davidson,I. and Tora,L. (2011)

seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.*, **39**, e35.

15. Giannopoulou,E.G. and Elemento,O. (2011) An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics*, **12**, 277.

16. Halbritter,F., Kousa,A.I. and Tomlinson,S.R. (2013) GeneProf data: a resource of curated, integrated and reusable high-throughput genomics experiments. *Nucleic Acids Res.*, 10.1093/nar/gkt966.

17. Liu,T., Ortiz,J. a, Taing,L., Meyer,C. a, Lee,B., Zhang,Y., Shin,H., Wong,S.S., Ma,J., Lei,Y., *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.

18. Balwierz,P.J., Carninci,P., Daub,C.O., Kawai,J., Hayashizaki,Y., Van Belle,W., Beisel,C. and van Nimwegen,E. (2009) Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.*, **10**, R79.

19. Siddharthan,R., Siggia,E.D. and van Nimwegen,E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.

20. Arnold,P., Erb,I., Pachkov,M., Molina,N. and van Nimwegen,E. (2012) MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, **28**, 487–94.

21. Ovaska,K., Laakso,M., Haapa-Paananen,S., Louhimo,R., Chen,P., Aittomäki,V., Valo,E., Núñez-Fontarnau,J., Rantanen,V., Karinen,S., *et al.* (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med.*, **2**, 65.

22. Http://supportres.illumina.com/documents/myillumina/6378de81-c0cc-47d0-9281-724878bb1c30/2012-09-18_illuminacustomersequenceletter.pdf GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG, ACACTCTTTCCCTACACGACGCTCTTCCGATCT, GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG, TGGAATTCTCGGGTGCCAAGG, GATCGGAAGAGCACACGTCTG, TCGTATGCCGTCTTCTGCTTG.

23. Stadler,M.B., Murr,R., Burger,L., Ivanek,R., Lienert,F., Schöler,A., van Nimwegen,E., Wirbelauer,C., Oakeley,E.J., Gaidatzis,D., *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–5.

24. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

25. Schmid,C.D. and Bucher,P. (2007) ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell*, **131**, 831–2; author reply 832–3.

26. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

27. Pachkov,M., Balwierz,P.J., Arnold,P., Ozonov,E. and van Nimwegen,E. (2013) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.*, **41**, D214–20.

28. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–17.

29. Bryne,J.C., Valen,E., Tang,M.-H.E., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–6.

30. Kulakovskiy,I. V, Medvedeva,Y. a, Schaefer,U., Kasianov,A.S., Vorontsov,I.E., Bajic,V.B. and

Makeev,V.J. (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, **41**, D195–202.

31. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–82.

32. Wang,J., Zhuang,J., Iyer,S. and Lin,X. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome …*, 10.1101/gr.139105.112.

33. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., *et al.* (2013) DNA-Binding Specificities of Human Transcription Factors. *Cell*, **152**, 327–339.

34. Van Nimwegen,E. (2007) Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, **8 Suppl 6**, S4.

35. Egghe,L. and Michel,C. (2003) Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. *Inf. Process. Manag.*, **39**, 771–807.

36. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsGm12878Brca1a300IggmusRawDataRep1.fastq.gz.

37. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsGm12878Brca1a300IggmusRawDataRep2.fastq.gz.

38. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsGm12878IggmusRawDataRep1.fastq.gz.

39. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsGm12878IggmusRawDataRep2.fastq.gz.

40. http://crunch.unibas.ch/ENCODE_REPORTS/Snyder_Stanford/IggMus/report_BRCA1A/index.html.

41. Hall,J., Lee,M., Newman,B. and Morrow,J. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science (80-.).*, **250**, 17–22.

42. DengÃ,C. and Brodie,S. (2000) Roles of BRCA1 and its interacting proteins. *Bioessays*.

43. Rosen,E., Fan,S. and Ma,Y. (2006) BRCA1 regulation of transcription. *Cancer Lett.*, **236**, 175–185.

44. Starita,L. (2003) The multiple nuclear functions of BRCA1: transcription, ubiquitination and DNA repair. *Curr. Opin. Cell Biol.*, **15**, 345–350.

45. Wang,Y., Cortez,D. and Yazdi,P. (2000) BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes …*, 10.1101/gad.14.8.927.

46. Mikula,M., Gaj,P., Dzwonek,K., Rubel,T., Karczmarski,J., Paziewska,A., Dzwonek,A., Bragoszewski,P., Dadlez,M. and Ostrowski,J. (2010) Comprehensive analysis of the palindromic motif TCTCGCGAGA: a regulatory element of the HNRNPK promoter. *DNA Res.*, **17**, 245–60.

47. Wyrwicz,L.S., Gaj,P., Hoffmann,M., Rychlewski,L. and Ostrowski,J. (2007) A common cis-element in promoters of protein synthesis and cell cycle genes. *Acta Biochim. Pol.*, **54**, 89–98.

48. Fabbro,M., Savage,K., Hobson,K., Deans,A.J., Powell,S.N., McArthur,G. a and Khanna,K.K. (2004) BRCA1-BARD1 complexes are required for p53Ser-15 phosphorylation and a G1/S arrest following ionizing radiation-induced DNA damage. *J. Biol. Chem.*, **279**, 31251–8.

49. Vermeulen,J.F., van de Ven,R. a H., Ercan,C., van der Groep,P., van der Wall,E., Bult,P., Christgen,M., Lehmann,U., Daniel,J., van Diest,P.J., *et al.* (2012)

Nuclear Kaiso expression is associated with high grade and triple-negative invasive breast cancer. *PLoS One*, **7**, e37864.

50. Pao,G. and Janknecht,R. (2000) CBP/p300 interact with and function as transcriptional coactivators of BRCA1. *Proc. ….*

51. Hu,Y. and Li,R. (2002) JunB potentiates function of BRCA1 activation domain 1 (AD1) through a coiled-coil-mediated interaction. *Genes Dev.*, **1**, 1509–1517.

52. Hong,C.P., Choe,M.K. and Roh,T.-Y. (2012) Characterization of Chromatin Structure-associated Histone Modifications in Breast Cancer Cells. *Genomics Inform.*, **10**, 145–52.