ChIP-R: Assembling reproducible sets of ChIP-seq and ATAC-seq peaks from multiple replicates

Rhys Newell¹, Richard Pienaar¹, Brad Balderson¹, Michael Piper², Alexandra Essebier^{1,*}, Mikael Bodén^{1,*}

¹*School of Chemistry and Molecular Biosciences, The University of Queensland, Cooper Road, QLD 4072, Australia*

²*School of Biomedical Science, The University of Queensland, Australia*

**Joint senior authors. Address for correspondence: m.boden@uq.edu.au*

Abstract

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the primary protocol for detecting genome-wide DNA-protein interactions, and therefore a key tool for understanding transcriptional regulation. A number of factors, including low specificity of antibody and cellular heterogeneity of sample, may cause “peak” callers to output noise and experimental artefacts. Statistically combining multiple experimental replicates from the same condition could significantly enhance our ability to distinguish actual transcription factor binding events, even when peak caller accuracy and consistency of detection are compromised.

We adapted the rank-product test to statistically evaluate the reproducibility from any number of ChIP-seq experimental replicates. We demonstrate over a number of benchmarks that our adaptation “ChIP-R” (pronounced ‘chipper’) performs as well as or better than comparable approaches on recovering transcription factor binding sites in ChIP-seq peak data. We also show ChIP-R extends to evaluate ATAC-seq peaks, finding reproducible peak sets even at low sequencing depth. ChIP-R decomposes peaks across replicates into “fragments” which either

form part of a peak in a replicate, or not. We show that by re-analysing existing data sets, ChIP-R reconstructs reproducible peaks from fragments with enhanced biological enrichment relative to current strategies.

Keywords: ChIP-seq, ATAC-seq, Reproducibility, Transcription factor, Transcriptional regulation

1. Introduction

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) provides an *in vivo* snapshot of genomic sites occupied by a target protein in a biological sample (Johnson et al., 2007); the target is typically a transcription factor (TF) or a histone with a specific modification. ChIP-seq protocols are susceptible to produce false positive signals as DNA regions not bound by the target protein can be pulled down indiscriminately during immunoprecipitation. The protocol requires a large cell population, increasing the risk of sample impurities that in turn compromise both sensitivity and specificity, with targeted binding events only present in a fraction of the cell population (Furey, 2012; Bailey et al., 2013).

A *reproducible experiment* is one performed over multiple, independent replicates with consistent results (Viswanathan et al., 2007; Ioannidis et al., 2009). Measuring reproducibility of a *whole* experiment indicates the overall quality of the resulting data set and is thus vital to the scientific process (Li et al., 2011). This is particularly pertinent in the case of ChIP-seq where variability between *individual* data points, both within and across replicates is commonplace; an understandable and reportable metric enables filtering of binding sites with the aim of reducing noise and prioritising binding sites for further study.

Peak calling methods identify “peaks” or regions where the target protein is

bound based on the aggregation of sequenced reads. A large number of peak calling tools exist; the most popular include MACS2, SPP, SISSRs and GEM (Feng et al., 2012; Kharchenko et al., 2008; Jothi et al., 2008; Guo et al., 2012; Thomas et al., 2017). Peak calling tools often have stringent cutoffs to combat false positives which, if used carelessly, result in false negatives. Taking this, and the variance in ChIP-seq data into account, it is advisable to use relaxed cutoffs for peak calling and, through the application of proper reproducibility testing, filter by statistics instead (Li et al., 2011). Many peak calling methods are only capable of assessing biological replicates individually and few report on the reproducibility of the experiment or individual peaks present across replicates.

With the advent of ENCODE 3 there is a need to upgrade peak reproducibility tools to enable comparison between multiple replicates (The ENCODE Project Consortium et al., 2020; Landt et al., 2012), since the decreased cost of sequencing has resulted in a corresponding increase in the typical number of replicates used for ChIP-seq experiments (Zhang et al., 2014; Muir et al., 2016). Until the work presented herein, there has been no such method which has been designed for this purpose.

Highly variable yet reproducible peaks will often be missed during a pairwise analysis, especially if a peak signal is strong in one replicate but weak in another (Yang et al., 2014). Using a third or fourth replicate can validate signals by providing a broader context for the observed variability. A less considered aspect is that reproducibility of a peak is usually evaluated after its genomic boundaries are

42 predicted. Optimising specificity of peak boundaries through the use of additional
43 replicates ensures the inclusion of sequence content relevant to the biology of the
44 cell being studied, and exclusion of neighbouring regions that do not capture ac-
45 tual binding events.

46 The assessment of reproducible peaks representing thousands of TF binding
47 events is non-trivial; independently verified data sets are limited to a small number
48 of TFs and conditions, and they typically include few negative data points. There-
49 fore, this paper outlines several complementary analyses to measure performance
50 of reproducibility metrics, with a focus on identifying genomic regions enriched
51 with biologically relevant information and highlighting the ability to recover re-
52 producible portions of TF binding sites without loss of information.

53 Current implementations that provide metrics for ChIP-seq assay reproducibil-
54 ity are tailored for *pairwise* comparison, including irreproducible discovery rate
55 (IDR; Li et al., 2011); others use empirical methods that are computationally pro-
56 hibitive in the presence of multiple replicates (Feng et al., 2012; Nix et al., 2008;
57 Zhang et al., 2008).

58 Peak callers that themselves process biological replicates include MACS2
59 (Feng et al., 2012), PePr (Zhang et al., 2014), Sierra Platinum (Müller et al.,
60 2016) and BinQuasi (Goren et al., 2018). MACS2 can analyse multiple replicates
61 by pooling reads into a combined super sample. Reads from separate replicates
62 occurring at the same location are filtered out as a result from over-amplification
63 of DNA, which causes loss of information and resolution. PePr, Sierra Platinum
64 and BinQuasi also combine reads across multiple replicates to improve peak call-
65 ing; while quality statistics are produced, none report specifically on peak repro-
66 ducibility. While we include MACS2 as a baseline comparison, our focus is on

67 the category of tools that are run subsequent to peak calling, treating each sample
68 as originating from an independent experiment.

69 The current most widely used approach for assessing reproducibility from
70 replicates is IDR. IDR is suggested by the ENCODE guidelines and quantifies
71 reproducible peaks by evaluating the consistency in the assignment of ranks be-
72 tween replicates (Li et al., 2011). The standard implementation of IDR is lim-
73 ited to evaluate two replicates at a time, requires that a peak occurs in both, and
74 assigns a reproducibility metric to the interval represented by the (single, unad-
75 justed) peak favoured by the peak caller. If the goal is to merge more than two
76 replicates, a naive approach is to take the union of peaks across replicates. MSPC
77 (Jalili et al., 2015) corroborates statistical evidence in individual replicates using
78 Fisher’s method to rescue weaker peak calls; it extends naturally to greater than
79 two replicates but was only tested on a single transcription factor (TF), and reports
80 a large number of peaks compared to IDR.

81 The rank-product test is inherently suited to look at multiple replicates in tan-
82 dem, avoiding the need for pairing replicates arbitrarily for analysis (Breitling
83 et al., 2004; Breitling and Herzyk, 2005). The test is non-parametric and has
84 previously been used to evaluate reproducibility of other data types including mi-
85 croarrays (Eisinga et al., 2013). We hypothesised that this statistical approach
86 can effectively evaluate reproducibility of peaks as well as ascertain the most reli-
87 able boundaries of such, by decomposing individual peaks across replicates. Ul-
88 timately, reproducible and appropriately specific ChIP-seq peaks should display
89 greater enrichment of biological content. This paper outlines our approach, im-
90 plemented as a tool called ChIP-R; we demonstrate that an adaptation of the rank-
91 product test enhances the quality of biological information extracted from ChIP-

seq data, while typically (but not exclusively) outperforming other approaches that statistically combine multiple replicates.

2. Results

2.1. Analysis of replicates

ChIP-R adapts the rank-product analysis of replicates. It initially inspects peaks in all k replicates to extract n canonical “fragments” (Figure 1). Each fragment is defined in terms of the genomic interval it occupies, and is linked to all replicates (allowing the inspection of the values assigned by peak caller to intersecting replicates). With a total of exactly n canonical fragments, each linked to k replicates, a fragment i - replicate j tuple is assigned a rank r_{ij} , based on the signal it has *within* replicate j ; the rank ranges between 1 and n , for the highest to the lowest signal, respectively. (Here, “signal” means what the peak caller uses to indicate the strength of the peak, which can be any meaningful statistic representing the confidence of calling it.) Tied fragments receive a rank that is the average of the positions they occupy, e.g. if three fragments are tied at positions 13 to 15, all three receive a rank of 14.

From the above, each canonical fragment i is associated with a rank tuple $\{r_{i1}, r_{i2}, \dots, r_{ik}\}$. The rank-product statistic is then calculated as the product of these ranks for the i th fragment across k replicates.

$$rp_i = \prod_{j=1}^k r_{ij} \quad (1)$$

2.1.1. Peak fragments and composite peak boundary

While peaks often occupy similar, intersecting genomic locations between replicates, the base pairs of start and end points of such intervals are not always

111 identical (Figure 1a). Minor differences in location reflect the nature of data and
112 are not a sign of low reproducibility of the peak itself. The availability of multiple
113 replicates exacerbates the fragmentation but provides us with a basis to identify
114 the specific genomic interval that makes up the most reproducible peak.

115 To determine the rank-product statistic, a consistent set of entries is required
116 to compare and rank across replicates; we developed a scanning approach to de-
117 fine the start, end and support for peak fragments across the genome (Figure 1b).
118 Fragments are formed from (a) inspecting all intervals that make up peaks in a
119 genomic region *across* the replicates, and (b) decomposing the intervals into the
120 largest sub-intervals from which all the original peaks can be reconstructed by
121 concatenation. The fragments are defined by the sub-intervals and are consistent
122 across and shared between all replicates; each fragment when linked to a replicate
123 is assigned the same level of confidence or signal as the peak it originates from.
124 Each fragment can then be ranked within each replicate and labeled with a rank
125 tuple (Figure 1d).

126 There is one important exception to this. Some peak fragments do not intersect
127 with any peak interval in a specific replicate; subject to a user-specified minimum
128 number of intersecting peak intervals (m , by default 1), we assign to this “pseudo”
129 fragment the lowest rank in the corresponding replicate (Figure 1c).

130 When assessing a fragment across all replicates, with one or more being pseudo
131 fragments, the rank-product will be penalised, decreasing the likelihood of the
132 fragment being called reproducible. Pseudo fragments allow detection of actual
133 binding sites within peak regions of high variability (e.g. supporting reads occur-
134 ring in a small fraction of the input cell population).

Figure 1: Key steps of ChIP-R. An example of evaluating and composing a set of reproducible peaks from three replicates. (a) Peak caller-specific signals are considered to indicate strength of each peak in each replicate. (b) Peaks are assigned replicate-specific ranks based on peak signal values. (c) Peak fragments are identified based on the intersection of peak intervals across replicates. (d) Pseudo fragments are introduced to account for absence of intersecting peak intervals in one or more replicates, each assigned the lowest possible rank in each replicate (here n is the number of peaks called in each replicate). (e) Collect replicate tuples with ranks for each peak fragment. (f) Calculate rank products for each peak fragment. (g) Determine rank product test statistics. (h) Retain only peak fragments that satisfy test criteria. (i) Combine fragments into reproducible peaks presented to user with statistics from composite fragments.

135 2.1.2. *Reproducible peaks from reproducible fragments*

136 Designating a specific fragment as reproducible involves two steps. First,
137 a rank-product p is calculated from the null sampling distribution described in
138 Heskes et al. (2014); this efficiently reproduces the original rank-product statistic
139 (Koziol, 2010; Eisinga et al., 2013) that involves large numbers of permutations
140 of k ranked lists of n fragments.

141 As a second step, we use the cumulative distribution function of the bino-
142 mial distribution to determine a cut-off $p \leq \theta$ for fragments to be reported (see
143 Methods). Failed fragments are removed from consideration. This step avoids the
144 arbitrary scaling of p , in favour of adapting it to the data at hand.

145 ChIP-R assembles output peak intervals from adjacent peak fragments, re-
146 quiring that they all meet the statistical test of reproducibility $p \leq \theta$. ChIP-R
147 annotates each output peak with the p -value of the most significant of the com-
148 posite fragments. ChIP-R provides an optional filter for transcription factors (see
149 Methods). A complete workflow is illustrated in Supp. Figure 1.

150 2.1.3. *Benchmarks*

151 Apart from IDR, we compare ChIP-R against MSPC, and refer to m -union
152 and MACS2 pooled replicates option as performance baselines (see Methods for
153 details and settings). While ChIP-seq data for TFs are available at volume, data
154 sets strictly suited to validation of reproducibility are limited. To capture a variety
155 of conditions, we collated ChIP-seq data with four replicates across six TFs in five
156 cell lines, including REST and MAX in K562, SRF in GM12878, CTCF in MCF-
157 7, CEBPB in A549 and FOXA1 in HepG2; CTCF in A549 were available in three
158 replicates (Supp. Table 5). Rye et al. (2011) uniquely provides an independently
159 validated but incomplete set of positive and negative peak labels for REST, SRF

160 and MAX, in the aforementioned cell lines.

161 For peak calling, we used MACS2 to ensure results are not confounded by
162 choice of peak caller; to verify that the evaluation herein is not specific to tool
163 used, we also used SPP on the REST, MAX and SRF data sets. To understand
164 the influence of peak caller on reproducibility statistics, on CTFC in A549, we
165 explored SISSRs, GEM and ENCODE-curated peaks apart from those we called
166 with MACS2 and SPP. Complete details of data sets and methods including their
167 parameters are provided in Methods.

168 2.2. *ChIP-R separates validated positive from negative TF binding sites*

169 The use of the rank-product test allows for ChIP-R to make use of all avail-
170 able replicates and consistently outperform other tested tools on the data sets with
171 independently validated positives and negatives. Figure 2 displays the results of
172 the ROC curves for IDR and ChIP-R when using either MACS2 and SPP as the
173 peak caller. ChIP-R outperforms IDR when using multiple replicates and pairwise
174 combinations of replicates, suggesting that the rank product is better at assigning
175 higher reproducibility to true peaks.

176 Supp. Table 1 contains auPRC values for all methods. The auPRC values
177 for IDR when using MACS2 showed a noticeable decrease that was not reflected
178 when using any of the other methodologies. ChIP-R maintained consistently high
179 auPRC values regardless of peak caller, replicate combination or number of repli-
180 cates used. As an aside, both IDR and ChIP-R generated better auPRC values on
181 peaks called by SPP.

Figure 2: Reproducibility by rank product predicts verified binding sites better than IDR across three benchmark ChIP-seq data sets, using two peak callers. Receiver operating characteristic (ROC) curves for IDR and ChIP-R using MACS2 (a) and SPP (b) as the peak caller. The x-axis displays the False Positive Rate (FPR) and the y-axis the True Positive Rate (TPR). IDR was used on each set of replicates and then the resulting peaks were pooled in the IDR Pool peak set. ChIP-R was used on each pair of replicates and also on all four replicates at once using ‘-m 2’ and ‘-fragment’ options.

182 2.3. *ChIP-R recovers known positives with fewer predictions*

183 With an arbitrary set threshold a method divides tens of thousands of potential
 184 peaks into two groups (positive and negative predictions), but validated labels are
 185 available only for a fraction (ranging from 156 to 289 depending on TF). Stan-
 186 dard metrics such as auROC, auPRC and F1 that rely on the grouping of data into
 187 true and false positives, and true and false negatives, are therefore inadequate in
 188 isolation. Moreover, the number of peaks that are evaluated varies from method
 189 to method, arbitrarily truncating predictions and distorting metrics. We there-
 190 fore devised a complementary metric based on the recovery of positives while
 191 minimising the inclusion of negatives, allowing for (a large number of) undesig-
 192 nated predictions: area between the positive and negative curves (ABC) involves
 193 sorting output peaks by method-specific p -value; each output overlaps with a val-
 194 idated positive or negative peak, or no peak. The respective sum of validated
 195 positive and negative peaks is plotted against the number of reproducible peaks
 196 reported by each method, with all counts normalised to their respective totals (to-
 197 tal positive, total negative or total peaks). The difference of the area under the
 198 positive and negative curves using the trapezoidal method approaches 1 for opti-
 199 mal performance. A negative ABC value indicates a greater accumulation rate of
 200 validated negative peaks compared to validated positive peaks. ABC values are
 201 influenced by the number of predictions that a method produces, as the number
 202 of known positives and negatives are usually not exhausted. If two ABC values
 203 are equal, the method with better performance is the one that called fewer peaks.
 204 Some methods do not quantify reproducibility for all peaks, so to compare them
 205 fairly, we also calculated an ABC value for the Top- N peaks, where N equals the
 206 smallest number of peaks produced by *any* method for the TF.

207 The adaptation of the rank-product test identifies reproducible peaks from
208 multiple ChIP-seq samples and recovers known positives ahead of known neg-
209 atives, as illustrated by higher ABC values, while providing the user with a bio-
210 logically relevant number of potential TF binding sites; Figure 3b shows repre-
211 sentative outcomes for the TF REST. Supp. Tables 2 and 3 contain both ABC and
212 Top-*N* ABC values for all tested TFs.

213 It is essential to interpret ABC values alongside peak counts and peak width
214 to represent performance, as a method calling a greater number or overly broad
215 reproducible peaks is more likely to identify a peak validated to be true by chance
216 (see Supp. Table 2). In several instances, MSPC, *m*-union and MACS2 also
217 achieve high ABC values at default settings; these methods are permissive and
218 report a contestably large number of peaks. IDR is conservative in its peak calls
219 and has consistently low ABC values; IDR identifies *no* true peaks in SRF.

220 Each method shows key differences in the quality of returned peaks as exem-
221 plified in Figure 3a. First, MSPC and *m*-union both provide the union of overlap-
222 ping peaks for each genomic location resulting in a broader range of peak sizes
223 including large peaks and, at less stringent settings, a number of smaller peaks
224 less likely to represent actual binding events. In contrast, ChIP-R and IDR do not
225 rely on a union approach and report peaks that are more concentrated around the
226 highest scoring motif matches. IDR selects a single, original peak to represent
227 the genomic region. ChIP-R alters the reproducible peak boundary based on all
228 available information; resulting widths tend to be uniform, rather than influenced
229 by the number of replicates and in line with the size expected for a TF.

230 2.4. *ChIP-R recovers peaks that are biologically meaningful*

231 To improve the comprehensiveness and biological meaning of evaluation, we
 232 devised two tests to evaluate the biological specificity of reproducible peaks as
 233 offered by different methods, and by varying numbers of replicates; both tests are
 234 based on the expectation that TFs bind to specific sequence motifs. Sequences
 235 that affect binding of a TF are easier to recover *if* peaks are centered and con-
 236 fined around all actual binding events, and exclude regions that are not supporting
 237 binding. Both tests use evidence for binding as is offered through the presence
 238 or absence of sequence motifs, or the extent of sequence reads from the ChIP-seq
 239 experiment.

240 **Test 1: Reproducible peaks contain TF-specific motifs at sites with higher** 241 **read density.**

242 We used FIMO to provide independent evidence of potential TF binding (Grant
 243 et al., 2011) based on position weight matrices (PWMs) (Mathelier et al., 2016).
 244 Significant occurrences were scored by the number of sequence reads (in the
 245 merged BAM files for the TF) that overlapped precisely with the motif match site.
 246 “Motif positive peaks” are peaks that are reported reproducible by the method and
 247 also overlap at least one motif match. Appropriately set peak boundaries ensures
 248 the inclusion of motif matches with greater sequence read support relevant to the
 249 binding of the TF, and exclusion of sequence that is not supported by reads from
 250 binding events. In our assessment, a better method identifies a higher proportion
 251 of peaks *with* motif matches, while still allowing for binding events not repre-
 252 sented by a motif match. More importantly, amongst motif positive peaks only,
 253 a better method produces more peaks with increased read depth, and fewer peaks
 254 with low read depth at the motif match site.

ChIP-R selectively identifies motif matches with greater read depth that are more likely true binding sites compared to other approaches (Figure 3b and Supp. Figures 2 and 3). Most methods recover the motif matches with greater read depth (the green line represents the mean score of motif matches overlapping known positives); IDR is the exception that discards a set of genomic regions containing highly supported motif matches. This is also supported by the Top-*N* ABC values in Supp. Tables 1 and 2, which show all tools have comparable performance when identifying the most reproducible peaks. Increasing the number of output peaks from IDR (by relaxing its threshold) includes motif matches with lower read counts.

Across all TFs, MSPC, *m*-union and MACS2 recover large numbers of peaks with scores comparable to motif matches overlapping true negative peaks, while ChIP-R and IDR tend to recover peaks with highly supported motif matches (Figure 3b and Supp. Figures 2 and 3). IDR's performance varied across TFs with mean scores for CTCF, CEBPB and FOXA1 dropping in line with or below MSPC, *m*-union and MACS2 while ChIP-R consistently maintained high scores. For REST, IDR has the highest proportion of motif positive peaks followed by ChIP-R while the other methods drop to low proportions. ChIP-R generally has equal or higher proportions of motif positive peaks across all tested TFs. It is unlikely that all biologically relevant peaks contain a motif match and this is reflected by only 78% of the (independently verified) true positive peaks containing one (Figure 3c).

Test 2: Reproducible peaks from multiple replicates enhance discovery of TF-specific motifs. *De novo* motif discovery should benefit from including precisely the peaks with bound sequences, and excluding everything else. In our

assessment, the better method (or setting of method) allows for (a) the discovery of enriched sequence motif (by MEME-ChIP motif E-value; Machanick and Bailey 2011), (b) significant differential enrichment of sequence motif in foreground compared to background sets (MEME-ChIP's Fisher E-value) and (c) optimised centering of peak around the potential site of binding.

ChIP-R identifies sets of reproducible peaks centrally enriched for a target sequence motif, from the replicates available for each tested TF. m indicates the minimum number of intersecting peak intervals across k replicates, and thus controls the degree with which pseudo-fragments fill in for missing replicates. Figure 4b shows that for REST, a full or partial sequence motif is discovered in peak sets deemed reproducible at any value of m , while in SRF, a full or partial sequence motif is discovered in all *except* the most stringent setting ($m = 4$). For REST and SRF, reproducible peaks are enriched for the target sequence motif for $m \leq 3$; for CTCF, CEBPB and FOXA1 the same is true for $m \leq 2$ (not shown). MAX is the exception; here, the target sequence motif is *only* found when $m \geq 3$.

Figure 4a shows an example for each TF of the proportion of positive and negative peaks detected as more peaks are considered, as well as the difference between the curves captured as the ABC value, with an optimal score of 1. We observed a decrease in the ABC value with m increasing across REST and SRF. For MAX, which has a highly variable set of replicates, we observed an increase (Supp. Table 2).

The ideal setting of m varies with TF and data quality. As the set of known peaks is limited, ABC values with motif information (read depth and enrichment), peak count and peak size provide a guide to setting m . Using SRF as an example, setting m to 2 returned the expected number of peaks (representative of SRF's ac-

305 tivity in K562) with a high ABC score, high Top- N ABC score, high mean motif
306 score and strong presence of the motif combined with enrichment of the motif in
307 the foreground set indicating a biologically relevant result (Figure 4 and Supp. Ta-
308 bles 1-3). 4-union had a comparable ABC and Top- N ABC score but called a large
309 number of peaks with increased breadth and lower mean motif scores.

310 **Pseudo-fragments identify variable but reproducible peaks.** Pseudo-fragments
311 permit peaks represented non-exhaustively across replicates to still be included in
312 the rank-product test, thereby alleviating consequences of peak callers failing to
313 recognise under-represented events. Technically, this effect is increased at lower
314 values of m , and the permissibility of pseudo-fragments appears to influence our
315 ability to extract biologically meaningful events from the resulting peaks. Weakly-
316 or non-reproducible, but biologically relevant peaks are ever-present; for example
317 where a binding event occurs in a sub-population of cells present in only one
318 replicate. All tested TFs (bar MAX) show a general decrease in ABC values, mo-
319 tif enrichment and peak count as m increases (Figure 4 and Supp. Tables 1 and
320 2). This suggests requiring that all replicates carry a peak inhibits the capturing of
321 biological information; to make matters worse, this restrictive setting is required
322 by MSPC and m -union to achieve higher accuracy. Through the use of pseudo-
323 fragments, the rank-product test evaluates peaks across all replicates (at $m \geq 1$)
324 and by re-composing them into reproducible peaks, our adaptation achieves con-
325 sistent performance above or on par with alternative methods (Supp. Tables 1-2).

326 2.5. *The binomial test removes non-reproducible peaks*

327 The binomial test provides a dynamic and informative threshold $p < \theta$ for
328 separating reproducible from non-reproducible peaks. θ is consistently placed at
329 the ‘elbow’ of the positive curve where a large proportion of positives have been

Figure 4: Reproducibility of rank product recovers known binding sites and regions with TF-specific motifs. (a) The sum of positive and negative peaks was plotted against the number of peaks reported by ChIP-R (sorted by p -value) with all peak counts normalised to their respective totals (total positive, total negative or total peaks). The area between the positive and negative curves (ABC value) approaches 1 by including known positives and excluding known negatives. The threshold from the binomial test θ is marked by a dashed line. (b) MEME suite analysis of ChIP-R outputs from analysing REST and SRF. ChIP-R was run on all four replicates from REST and SRF with m set from 1 to 4. Outputs were separated into a foreground of peaks at $p < \theta$ and a background of all remaining peaks. MEME suite was used to perform motif discovery and differential enrichment. Each plot shows the central enrichment of the foreground set (solid line) and background set (dashed line) for the discovered target sequence motif represented as a logo with associated E-value (from MEME). A small value means greater enrichment of the sequence motif. A small Fisher E-value (from MEME-ChIP) implies a difference of motif occurrence between the foreground and background peaks, which in turn indicates that the background have less biologically relevant sequence content. Motif E-values and Fisher E-values for all factors and m settings are available in Supp. Table 3.

330 identified without introducing negatives for all tested TFs (Figure 4a). When used
331 to as a default threshold to distinguish positive from negative peaks, the outcomes
332 are encouraging relative to alternative methods (Supp. Tables 1-2); sequences at
333 the recovered positive sites often contain the expected motifs (Figures 3 and 4,
334 Supp. Figure 2), an ability which seems unaffected by choice of peak caller as
335 discussed below.

336 2.6. *ChIP-R evaluates reproducible peaks from a variety of peak callers*

337 We called peaks from high quality CTCF ChIP-seq experiment with three
338 replicates by MACS2, SPP, SISSRs and GEM. We also retrieved peaks as pro-
339 cessed and archived by the ENCODE pipeline. We ran ChIP-R on the output of
340 each peak caller using m set to 1, 2, and 3 to evaluate how the probability distri-
341 butions changed alongside increased peak fragmentation, at $p \leq \theta$.

342 Supp. Figure 4 shows that the distributions of rank-product p -values are con-
343 sistent across peak callers, while the number of peaks varies as m increases. Vari-
344 ation in the number of reproducible peaks may be manifested in differences in-
345 herent to each peak caller; specifically, SISSRs was previously shown to produce
346 inconsistent results compared to other peak callers (Li et al., 2011).

347 Supp. Figure 3 shows motif matches and read depth distributions when SPP
348 is used to call peaks for the REST, SRF and MAX data with reproducible peaks
349 identified using ChIP-R or IDR; the patterns are broadly consistent with obser-
350 vations for MACS2 inputs, shown in Figure 3b and Supp. Figure 2. ChIP-R has
351 equal or slightly higher proportions of motif positive peaks when SPP is used in-
352 stead of MACS2. IDR is optimised for use with SPP but ChIP-R was still able
353 to identify higher proportions of motif positive peaks, with increased read depth.
354 Finally, we emulated the use of IDR with four replicates (Yang et al., 2014) but

355 doing so did not improve the proportions seen when compared to the paired runs
356 (Supp. Figure 3).

357 2.7. *ChIP-R outperforms IDR when assessing the reproducibility of ATAC-seq* 358 *peaks in low coverage data sets*

359 Calviello et al. (2019) describes ATAC-seq data generated from the bulk through-
360 put of HEK293 cell line samples. These samples were sequenced independently
361 at three separate depths termed the low, medium and high coverages. Both the low
362 and the high coverages have four separate replicates (two biological and two tech-
363 nical) whereas the medium coverage dataset only three (two biological and one
364 technical). The intention behind the application of ChIP-R in this capacity is to
365 ascertain whether it can be used to identify which chromatin regions from the low
366 and medium depth runs are reproducibly accessible. We suggest from these runs
367 the reproducible regions are those also present in the corresponding high depth
368 data set; we designated the high coverage data sets as ground truth and compared
369 the performance of ChIP-R in capturing this data to IDR. It is also noteworthy that
370 IDR is not complete in its labelling of each data point, it only reports reproducible
371 peaks with a significance greater than a determined threshold. For this reason, the
372 ROC curve generated via the data output from IDR will never reach completion.

373 After pre-processing, read alignment and peak calling with MACS2, regions
374 were labelled at whole peak and base pair levels based on their presence in the
375 corresponding high coverage run. The capabilities of both ChIP-R and IDR to
376 discriminate each labelled data point were then assessed and displayed in the form
377 of a ROC curve (Figure 5). ChIP-R is able to more accurately define reproducible
378 peaks at the base pair level from sparse information (Figure 5b), whilst displaying
379 a similar standard of accuracy at the peak level and medium coverage data sets.

Figure 5: Reproducibility of rank product extends to ATAC-seq. ROC curves for the ATAC-seq data at the peak (left) and base pair (right) levels. Each data point was labelled based on its presence in each high coverage set. The shaded areas represent the inter-replicate variation (the overall minimum and maximum values of any replicate), and each bold line the mean of the replicates. IDR and ChIP-R perform comparably in all instances except the low coverage set where ChIP-R is superior.

380 These results suggest that ChIP-R is a suitable tool for the statistical identi-
 381 fication of genomic regions such as those found in ATAC-seq data, in particular
 382 of lower coverage data sets; we propose that the inevitable loss of precision that
 383 accompanies the low coverage data sets is reduced by the stricter peak boundary
 384 definitions that result from the peak fragmentation in ChIP-R.

385 3. Discussion

386 The rank-product test is a non-parametric statistical test based on differences
 387 from independently conducted and ranked experimental measurements; it extends
 388 naturally and efficiently to more than two replicates. This study shows when

389 adapted to transcription factor ChIP-seq data this test provides a robust repro-
390 ducibility metric across peaks originating from separate biological samples.

391 A unique feature of ChIP-R is that peaks are first decomposed into fragments
392 that are statistically evaluated. This peak decomposition allows all replicates to
393 be considered when identifying the genomic region most likely to contain a bind-
394 ing site; it is used in deciding whether to retain or discard individual fragments
395 rather than complete peaks. The flexibility of the approach is extended by pseudo-
396 fragments which can identify a variety of fragment combinations as reproducible
397 across replicates rather than strictly ignoring regions where a fraction of repli-
398 cates lack a peak. The flip-side of this flexibility is that ChIP-R is not suited to
399 fragment (and de-fragment) truncated peaks that do not present the full region as-
400 sociated with the target TF, as produced by some peak callers, including GEM and
401 SISRrs.

402 *3.1. Biological relevance of reproducible peaks*

403 We demonstrate that ChIP-R enhances our ability to extract biologically mean-
404 ingful events from the output of existing peak calling methods, for two or more
405 replicates (Section 2.4). ChIP-R identifies and restricts the boundaries of peaks by
406 first decomposing them to recover TF motifs robustly supported by ChIP-seq read
407 depth; it aptly discards low read depth regions of the genome and overall provides
408 a biologically relevant peak set, all without requiring direct access to raw reads.

409 The ChIP-R work-flow contains a number of parameters that can be varied to
410 enable the user to expand their search to low reproducibility peaks or select for
411 regions with stronger statistical support. Ultimately, this results in peak bound-
412 aries that are specific to the target TF binding event. Strikingly, using a strict m
413 value (i.e. minimal number of peaks required to overlap for evaluation to occur)

414 does not guarantee optimal performance; we show that lower m values often pro-
415 vide more biologically relevant sets of reproducible peaks. We hypothesise that
416 a larger m captures either bias in the initial ChIP-seq data or a different class of
417 binding event, either of which could compromise the signal in the peak. TF bind-
418 ing is rarely independent; highly reproducible peaks across all replicates could
419 represent specific co-occurring events that implicate genome regions not relevant
420 to the binding of the TF itself (e.g. co-factors or associations with transcriptional
421 processes).

422 3.2. Improvements compared to existing methods

423 The rank-product reproducibility p -value prioritises known independently val-
424 idated positive peaks ahead of negatives and the binomial test identifies a moderate
425 threshold that accurately distinguishes a realistic number of supported sites allow-
426 ing ChIP-R to perform as well as or better than alternative methods (Section 2.3
427 and Supp. Tables 1-2).

428 m -union and MSPC outputs are constructed by set theoretic mergers of indi-
429 vidual replicate peaks around putative binding sites; they share highly similar out-
430 comes despite MSPC being informed by additional statistics. Both identify large
431 numbers of peaks with low read depth and broad peak boundaries more likely
432 associated with noise than true binding events. IDR peaks are generally broader
433 than ChIP-R peaks and in our tests, IDR cautiously reports fewer peaks than other
434 methods; working with only two replicates, motif matches are sometimes missed
435 even if supported by greater read depths. ChIP-R's advantages over MACS2 to
436 call peaks when using a set of pooled reads are also clear (Figure 3, Supp. Figure
437 2 and Supp. Tables 1-2). Pooling the reads from all replicates results in MACS2
438 calling broader peak regions with lower scoring motif matches when compared to

both ChIP-R and IDR; instead we suggest the practice of treating peak calling as a first step before evaluating reproducibility of peak fragments with ChIP-R, which in turn helps to detect boundaries of relevant sequence content.

Author contributions

RN, AE and MB designed and RN implemented ChIP-R. RN, AE, RP and BB designed and performed tests. MB, AE and MP supervised the project. RN, MB, AE, MP and RP wrote the paper.

Acknowledgments

RN, BB and AE were supported by Australian Government Research Training Program (RTP) Scholarships.

References

- Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., Zhang, J., 2013. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol 9, e1003326.
- Bailey, T.L., Bodén, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research , gkp335.
- Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P., 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573, 83–92.

- 459 Breitling, R., Herzyk, P., 2005. Rank-based methods as a non-parametric alterna-
460 tive of the T-statistic for the analysis of biological microarray data. *J Bioinform*
461 *Comput Biol* 3, 1171–1189.
- 462 Calviello, A.K., Hirsekorn, A., Wurmus, R., Yusuf, D., Ohler, U., 2019. Repro-
463 ducible inference of transcription factor footprints in atac-seq and dnase-seq
464 datasets using protocol-specific bias modeling. *Genome Biol* 42.
- 465 Eisinga, R., Breitling, R., Heskes, T., 2013. The exact probability distribution of
466 the rank product statistics for replicated experiments. *FEBS Lett* 587, 677–682.
- 467 Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X.S., 2012. Identifying ChIP-seq en-
468 richment using MACS. *Nature Protocols* 7, 1728–1740.
- 469 Furey, T.S., 2012. ChIP-seq and beyond: new and improved methodologies to
470 detect and characterize protein–DNA interactions. *Nature Reviews Genetics*
471 13, 840–852.
- 472 Goren, E., Liu, P., Wang, C., Wang, C., 2018. BinQuasi: a peak detection method
473 for ChIP-sequencing data with biological replicates. *Bioinformatics* 34, 2909–
474 2917.
- 475 Grant, C.E., Bailey, T.L., Noble, W.S., 2011. FIMO: scanning for occurrences of
476 a given motif. *Bioinformatics* 27, 1017–1018.
- 477 Guo, Y., Mahony, S., Gifford, D.K., 2012. High resolution genome wide binding
478 event finding and motif discovery reveals transcription factor spatial binding
479 constraints. *PLoS Comput Biol* 8, e1002638.

480 Heskes, T., Eisinga, R., Breitling, R., 2014. A fast algorithm for determining
481 bounds and accurate approximate p-values of the rank product statistic for repli-
482 cate experiments. *BMC Bioinformatics* 15, 367.

483 Ioannidis, J.P.A., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., Culhane, A.C.,
484 Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T.,
485 Nitzberg, M., Page, G.P., Petretto, E., van Noort, V., 2009. Repeatability of
486 published microarray gene expression analyses. *Nat. Genet.* 41, 149–155.

487 Jalili, V., Matteucci, M., Masseroli, M., Morelli, M.J., 2015. Using combined
488 evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics* 31, 2761–
489 2769.

490 Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-Wide Map-
491 ping of in vivo Protein-DNA Interactions. *Science* 316, 1497–1502.

492 Jothi, R., Cuddapah, S., Barski, A., Cui, K., Zhao, K., 2008. Genome-wide iden-
493 tification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic
494 Acids Res.* 36, 5221–5231.

495 Kharchenko, P.V., Tolstorukov, M.Y., Park, P.J., 2008. Design and analysis of
496 ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* 26,
497 1351–1359.

498 Koziol, J.A., 2010. Comments on the rank product method for analyzing repli-
499 cated experiments. *FEBS Lett* 584, 941–944.

500 Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou,
501 S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al., 2012. ChIP-

502 seq guidelines and practices of the ENCODE and modENCODE consortia.
503 Genome Research 22, 1813–1831.

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

506 Machanick, P., Bailey, T.L., 2011. MEME-ChIP: motif analysis of large DNA
507 datasets. Bioinformatics 27, 1696–1697.

508 Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.y., Denay, G., Lee, J., Shi, W.,
509 Shyr, C., Tan, G., Worsley-Hunt, R., et al., 2016. JASPAR 2016: a major ex-
510 pansion and update of the open-access database of transcription factor binding
511 profiles. Nucleic Acids Research 44, D110–D115.

512 Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J.,
513 Weinstock, G.M., Isaacs, F., Rozowsky, J., Gerstein, M., 2016. The real cost of
514 sequencing: scaling computation to keep pace with data generation. Genome
515 Biol 17.

516 Müller, L., Gerighausen, D., Farman, M., Zeckzer, D., 2016. Sierra Platinum:
517 a fast and robust peak-caller for replicated ChIP-seq experiments with visual
518 quality-control and-steering. BMC Bioinformatics 17, 377.

519 Nix, D.A., Courdy, S.J., Boucher, K.M., 2008. Empirical methods for controlling
520 false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinfor-
521 matics 9, 523.

522 Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for com-
523 paring genomic features. Bioinformatics 26, 841–842.

524 Rye, M.B., Saetrom, P., Drablos, F., 2011. A manually curated ChIP-seq bench-
525 mark demonstrates room for improvement in current peak-finder programs. Nu-
526 cleic Acids Res. 39, e25.

527 The ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Ep-
528 stein, C.B., Shores, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R.,
529 Halow, J., Van Nostrand, E.L., Freese, P., Gorkin, D.U., Shen, Y., He, Y., Mack-
530 iewicz, M., Pauli-Behn, F., Williams, B.A., Mortazavi, A., Keller, C.A., Zhang,
531 X.O., Elhajjajy, S.I., Huey, J., Dickel, D.E., Snetkova, V., Wei, X., Wang, X.,
532 Rivera-Mulia, J.C., Rozowsky, J., Zhang, J., Chhetri, S.B., Zhang, J., Victorsen,
533 A., White, K.P., Visel, A., Yeo, G.W., Burge, C.B., Lécuyer, E., Gilbert, D.M.,
534 Dekker, J., Rinn, J., Mendenhall, E.M., Ecker, J.R., Kellis, M., Klein, R.J., No-
535 ble, W.S., Kundaje, A., Guigó, R., Farnham, P.J., Cherry, J.M., Myers, R.M.,
536 Ren, B., Graveley, B.R., Gerstein, M.B., Pennacchio, L.A., Snyder, M.P., Bern-
537 stein, B.E., Wold, B., Hardison, R.C., Gingeras, T.R., Stamatoyannopoulos,
538 J.A., Weng, Z., 2020. Expanded encyclopaedias of DNA elements in the hu-
539 man and mouse genomes. Nature 583, 699–710.

540 Thomas, R., Thomas, S., Holloway, A.K., Pollard, K.S., 2017. Features that define
541 the best ChIP-seq peak calling algorithms. Brief Bioinform 18, 441–450.

542 Viswanathan, C.T., Bansal, S., Booth, B., DeStefano, A.J., Rose, M.J., Sailstad,
543 J., Shah, V.P., Skelly, J.P., Swann, P.G., Weiner, R., 2007. Quantitative Bioan-
544 alytical Methods Validation and Implementation: Best Practices for Chromato-
545 graphic and Ligand Binding Assays. Pharm Res 24, 1962–1973.

546 Yang, Y., Fear, J., Hu, J., Haecker, I., Zhou, L., Renne, R., Bloom, D., McIntyre,

547 L.M., 2014. Leveraging biological replicates to improve analysis in ChIP-seq
548 experiments. *Comput Struct Biotechnol J* 9.

549 Zhang, Y., Lin, Y.H., Johnson, T.D., Rozek, L.S., Sartor, M.A., 2014. PePr: a
550 peak-calling prioritization pipeline to identify consistent or differential peaks
551 from replicated ChIP-Seq data. *Bioinformatics* 30, 2568–2575.

552 Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E.,
553 Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based
554 Analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137.

555 4. Methods

556 Further information and requests for resources and reagents should be directed
557 to and will be fulfilled by the Lead Contact, Mikael Bodén (m.boden@uq.edu.au).

558 The code for ChIP-R is implemented in Python, available at [https://github.](https://github.com/rhysnewell/ChIP-R/)
559 [com/rhysnewell/ChIP-R/](https://github.com/rhysnewell/ChIP-R/) and can be installed through pip. There are no re-
560 strictions on its use.

561 *Method details*

562 *Binomial test thresholds rank-product p-values*

563 Preliminary tests showed that using rank-product $p < 0.05$ results in a compar-
564 atively small number of peaks being called reproducible, relative to IDR $p < 0.05$
565 used by Li et al. (2011).

For each fragment $l \in \{1, 2, \dots, n\}$ enumerated in ascending order of rank-
product p , we determine the probability of seeing l or more successes from n

attempts as shown in Equation 2.

$$F(l; n, p) = \sum_{i=l}^n \binom{n}{i} p^i (1-p)^{n-i} \quad (2)$$

F is close to 1 almost regardless of rank-product p for small l , and gradually decreases to 0 as l approaches n ; in words, designating at least one fragment as a “success” is very probable, but for every single one to be a success is highly improbable. Again by reference to $l \in \{1, 2, \dots, n\}$, the point of intersection between the binomial cumulative function and p recovers a sweet spot, where p is sufficiently permissive to include exactly l fragments and no more; this p (referred to a ChIP-R’s threshold θ) identifies the point l after which we will no longer be recovering a reproducible fragment.

Filtering fragments before assembly

ChIP-R can apply a criterion that filters fragments to ensure retained fragments form peaks that reasonably indicate DNA binding events. We apply this optional step for transcription factor ChIP-seq but not for ATAC-seq; here, ChIP-R determines the median peak interval width for each fragment, based on the width of the peak in each replicate to which it belongs, or if a pseudo fragment, the closest peak in that replicate. ChIP-R removes fragments whose size is less than a user supplied width, if it has no capacity by adjacency to other supported fragments to form a composite interval longer than that width; adjacent fragments need to have similar or stronger statistical support (for statistical significance to be “similar” they need to be within one order of magnitude). To keep it simple, we adjusted these rules from inspecting sites in data sets separate from those on which we evaluated the approach.

587 *BED union*

588 Each called peak identifies a genomic interval, allowing peaks from multiple
 589 replicates to be merged by set theoretic operators. For instance, BEDTools is com-
 590 monly used to identify a set of peaks with support from a collection of BED files
 591 (Quinlan and Hall, 2010). We define an m -union to be the union of intersecting
 592 intervals defined by peaks from a minimum of m replicates, where $m \leq k$ (where
 593 k is the total number of replicates). The union of a pair of two genomic intervals
 594 (which must *intersect*) is a single interval defined by the widest combination of
 595 the two start and end points. For example, with $k = 3$ and m set to 2 (styled $m2$),
 596 peaks from at least two replicates must have intervals that intersect for a union to
 597 be performed across all three replicates and an output peak to be reported. We
 598 use m -union as a baseline to combine peaks from k replicates. In this approach,
 599 an output peak is considered reproducible and assigned the p -value of the peak
 600 caller’s most significant intersecting input peak.

601 *IDR*

602 IDR analysis was run using default parameters on all combinations of pairs
 603 from each set of replicates, for each TF. Results are based on the “global IDR
 604 value” assigned to each peak; reproducible peaks are those with a value ≤ 0.05 .
 605 IDR implicitly selects sites where a peak is represented in *both* replicates, equiv-
 606 alent to $m2$ which is the maximum number of replicates IDR processes. For com-
 607 parison, we include results for peaks sourced directly from ENCODE, which use
 608 SPP as peak caller and IDR to evaluate reproducibility, with settings determined
 609 independently by the ENCODE consortium. In lieu of an implementation for
 610 more than two replicates, we include results from pooling peaks from two runs
 611 with IDR, each on two replicates, to emulate a four-replicate version of IDR (Yang

et al., 2014).

MSPC

MSPC was run using using the recommended parameters of a stringent threshold of 1E-8 and a permissive threshold of 1E-4 on all replicates as joint inputs. MSPC produces consensus peaks that were used to evaluate its performance. Each output peak is assigned a p -value, determined by Fisher’s method from the p -values of each overlapping input peak, and corrected for multiple testing. Reproducible peaks are defined as having a corrected $p \leq 0.05$.

MACS2

MACS2 analysis was run using default parameters on all replicates as “multiple treatment files”, essentially combining all reads into a single sample. All results are based on the q -value determined by MACS2. While reproducibility is not defined by MACS2, we used $q \leq 0.05$ to separate positives from negatives.

Data sets

We obtained BAM files containing ChIP-seq data for six transcription factors in five cell lines, including REST and MAX in K562, SRF in GM12878, CTCF in MCF-7, CEBPB in A549 and FOXA1 in HepG2 (Supp. Table 5). Each TF was assayed using two biological replicates in two separate, but consistent experiments, resulting in four biological replicates per TF. The REST, SRF and MAX data sets from Rye et al. (2011) are uniquely suited to evaluate differences in reproducibility performance across multiple replicates as independent validation was available.

For peak calling, MACS2 provides a p -value, q -value (statistical estimators) and signal value (read depth at peak location) for each reported peak (Feng et al., 2012; Thomas et al., 2017). Most peak calling tools output a p -value, so we used

636 this (when applicable) as the score by which peaks are ranked. We called peaks
637 for each individual TF replicate using both replicates of the cell type specific input
638 files (Supp. Table 5).

639 *Benchmark design*

640 We designed two tests for assessing biological enrichment in reproducible
641 peaks. In Test 1, we used FIMO from MEME suite to identify sites with sig-
642 nificant motif matches across the union of TF specific replicates, supplying in-
643 dependent evidence of potential binding (Bailey et al., 2009; Grant et al., 2011).
644 We represented each TF by a position weight matrix (PWM), sourced from JAS-
645 PAR (Mathelier et al., 2016). FIMO returned a set of motif occurrences that were
646 scored by the number of sequence reads (in the merged BAM files for the TF) that
647 overlapped precisely with the motif match site.

648 In Test 2, we used MEME-ChIP version 5.0.2 (Bailey et al., 2009; Machanick
649 and Bailey, 2011) to perform motif discovery and tested for differential enrich-
650 ment of sequence motif comparing a foreground set of peaks $p < \theta$ against all
651 remaining peaks ($p > \theta$). To avoid conflating the inclusion of positives with the
652 effects of refining the boundaries of them (as in the motif matching test), we gen-
653 erated uniform 500 base pair peaks using a ± 250 base pair window around peak
654 centres deemed reproducible by all methods.