BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 2 2014, pages 165–171
doi:10.1093/bioinformatics/btt667

Genome analysis
Advance Access publication November 18, 2013

# Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data

Sebastian M. Waszak[1,2], Helena Kilpinen[2,3,4], Andreas R. Gschwind[2,5], Andrea Orioli[5], Sunil K. Raghav[1], Robert M. Witwicki[5], Eugenia Migliavacca[2,5], Alisa Yurovsky[2,3,4], Tuuli Lappalainen[2,3,4], Nouria Hernandez[5], Alexandre Reymond[5], Emmanouil T. Dermitzakis[2,3,4] and Bart Deplancke[1,2,*]

[1]Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne, [2]Swiss Institute of Bioinformatics, 1015, Lausanne, Switzerland, [3]Department of Genetic Medicine and Development, University of Geneva Medical School, [4]Institute of Genetics and Genomics in Geneva, University of Geneva, 1211, Geneva, Switzerland and [5]Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, 1011, Lausanne, Switzerland

Associate Editor: Michael Brudno

**ABSTRACT**

**Motivation**: High-throughput sequencing technologies enable the genome-wide analysis of the impact of genetic variation on molecular phenotypes at unprecedented resolution. However, although powerful, these technologies can also introduce unexpected artifacts.

**Results**: We investigated the impact of library amplification bias on the identification of allele-specific (AS) molecular events from high-throughput sequencing data derived from chromatin immunoprecipitation assays (ChIP-seq). Putative AS DNA binding activity for RNA polymerase II was determined using ChIP-seq data derived from lymphoblastoid cell lines of two parent–daughter trios. We found that, at high-sequencing depth, many significant AS binding sites suffered from an amplification bias, as evidenced by a larger number of clonal reads representing one of the two alleles. To alleviate this bias, we devised an amplification bias detection strategy, which filters out sites with low read complexity and sites featuring a significant excess of clonal reads. This method will be useful for AS analyses involving ChIP-seq and other functional sequencing assays.

**Availability**: The R package *absfilter* for library clonality simulations and detection of amplification-biased sites is available from http://updepla1srv1.epfl.ch/waszaks/absfilter

**Contact**: sebastian.waszak@epfl.ch or bart.deplancke@epfl.ch

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on March 19, 2013; revised on September 30, 2013; accepted on October 29, 2013

## 1 INTRODUCTION

The advent of high-throughput sequencing technologies has enabled us to study the relationship between genomic variants and various molecular phenotypes at single basepair resolution. To elucidate the molecular mechanisms underlying allele-specific (AS) gene regulation, such as those affected by *cis*-regulatory variants (Montgomery *et al*., 2010; Pickrell *et al*., 2010), several studies have started to probe allele- or individual-specific changes in transcription factor (TF) binding and chromatin states (Degner *et al*., 2012; Kasowski *et al*., 2010; Kilpinen *et al*., 2013; McDaniell *et al*., 2010; Reddy *et al*., 2012), revealing that such changes are abundant and often correlate with gene expression differences.

To detect these changes, specialized computational workflows have been developed to eliminate several possible sources of bias inherently linked to the analysis of AS behavior. Although the analysis may seem straightforward, several technical hurdles need to be overcome to ensure reliable results. These include, for example, the reference allele mapping bias, i.e. the tendency of reads to map more frequently/better to the reference allele than the alternative allele (Degner *et al*., 2009; McDaniell *et al*., 2010; Rozowsky *et al*., 2011), biases linked to duplicated genomic regions (Pickrell *et al*., 2010; Rozowsky *et al*., 2011), quality of genotype calls (Rozowsky *et al*., 2011), or the statistical tests used (Montgomery and Dermitzakis, 2011; Rozowsky *et al*., 2011).

Here, using RNA polymerase II chromatin immunoprecipitation assay (ChIP) sequencing data from two parent–daughter trios from the 1000 genomes project, we studied another significant source of bias, namely, library amplification, which can introduce a large number of false-positive AS DNA binding (ASB) events. To address this bias, we developed a two-tiered amplification bias detection strategy that filters out many of these likely false-positive sites and greatly improves the overall reliability of the data. This pipeline should prove useful for many studies aiming to use ChIP-sequencing (ChIP-seq) technology to identify AS molecular changes (e.g. in protein–DNA binding or chromatin structure).

## 2 METHODS

### 2.1 RNA polymerase II ChIP-seq dataset

RNA polymerase II ChIP-seq data for two parent–daughter trios was obtained from Kilpinen *et al*. (2013). Briefly, we used $5 \times 10^7$ cross-linked lymphoblastoid cells for chromatin immunoprecipitation experiments

*To whom correspondence should be addressed.

with an anti-RPB2 antibody (sc-67318, Santa Cruz Biotechnology). ChIP-seq library preparation was performed using the Illumina ChIP-seq DNA Sample Prep Kit with 10.5 ng of purified ChIP-DNA. ChIP-DNA libraries were amplified using 18 polymerase chain reaction (PCR) cycles, according to manufacturers instructions, and sequencing was performed on an Illumina HiSeq 2000 machine (36-bp read length, single end) with one sample per lane. ChIP-seq reads were aligned against the standard hg19 build of the human reference genome with *BWA* 0.5.9 [Li and Durbin (2009); default settings]. For sequencing depth simulations, we randomly sampled 25–200 million mapped reads in 25 million steps using *samtools* (view -s; seed = 0). Reads with mapping quality <10 were discarded for AS binding analysis.

## 2.2 Read clonality simulations

We simulated 36-bp reads around heterozygous single nucleotide polymorphism (SNP) sites to obtain a better sense about clonal signals. In total, 200 datasets with 1000 artificial sites each were created, whereby 1–2000 reads were simulated per site. Each simulated dataset contained between 0 and 99.5% clonal reads in 0.05% steps (thus resulting in 200 simulation sets). We define a clonal read as being the product of the PCR library amplification step during sequencing library preparation. For example, a simulation with 100 reads/SNP site and a defined clonal population of 10% indicates that the site is covered by 90 independent reads and 10 clonal reads produced during the PCR amplification step. To obtain an empirical distribution of read alignment start sites around SNPs, we randomly placed all independent reads (out of all simulated reads) on the plus or minus strand around a SNP site. The maximum number of unique alignment read start sites for 36-bp reads is 72 (i.e. 2 × 36 bp). We naively assumed that each alignment position around a SNP site has an equal probability of being covered, i.e. we did not consider mappability biases, short-read alignment artifacts, or positional effects. Within each simulated dataset and for each SNP read depth value, we counted the number of unique read alignment start sites (URSS).

## 2.3 Estimation of ChIP-seq library clonality based on clonality simulations

For each simulated dataset and SNP read depth bin, we calculated the mean number of URSSs for all 1000 simulated SNP sites. To estimate the clonality (i.e. proportion of clonal reads at SNP sites) of an actual ChIP-seq library, we first obtained the distribution of URSSs for all tested heterozygous SNP sites [i.e. heterozygous SNPs covered by at least 20 reads and filtered for other artifacts; see Kilpinen *et al.* (2013)]. From the observed URSS distribution, we calculated for each SNP read depth bin the mean number of URSSs. SNP read depth bins with <5 SNPs were discarded from further analysis. We then selected the best clonality simulation model by comparing all simulated mean URSS distributions against the observed average URSS distribution using the Kolmogorov–Smirnov (KS) test implemented in the R statistical software (*stats* package, *ks.test* function; http://www.r-project.org). The simulation model with the lowest $D$ statistic was selected as the best clonality model.

## 2.4 Alternative estimation of ChIP-seq library clonality

Within each ChIP-seq library we counted the genome-wide number of URSSs ($N_{URSS}$) and the total number of mapped reads ($N_{reads}$). The overall clonality of a ChIP-seq library was approximated with $1 - N_{URSS}/N_{reads}$.

## 2.5 Global amplification bias SNP filter

The global filter is sequencing-depth independent and flags any site that is covered by reads with less than $N$ URSS, whereby $N$ ranges between 1 to (2 × read length). It is *a priori* unknown which cutoff yields the best trade-off between removing putative amplification-biased sites and retaining real sites. Our clonality simulation results show that for highly clonal libraries (>95% clonality), the number of URSSs is on average below five at sites covered by up to 100 reads. Therefore, we required that at least five URSSs should be covered by reads irrespective of read depth and library clonality.

## 2.6 Local amplification bias SNP filter

Based on the best clonality simulation model (see Section 2.3), we have tested for each accessible SNP (minimum 20 reads/site) whether the number of observed URSSs over both alleles deviates significantly from our expectations of having at least a certain number of URSSs. Using the clonality simulations, we drew for each site the expected distribution of URSSs and flagged SNPs as amplification-biased if the observed URSS value was equal or lower than the 50th element (out of 1000) of the sorted simulated URSS distribution (thus corresponding to an one-tailed $P$-value cutoff of 0.05). In addition, we performed power calculations by testing whether the proposed local (joint-allele) filter is able to identify instances where only one allele experiences an amplification bias and the second allele behaves as expected. We ran simulations for scenarios where a significant allelic imbalance (allele1/allele2) is >1.5-fold (and >2-fold) at sites covered by 20–400 reads ($P < 0.01$). These parameters represent the observed range of read depth at SNP sites (i.e. 90% of accessible sites were covered by <400 reads) as well as allelic imbalances (i.e. 90% of all significant sites had an allelic imbalance of >1.5-fold). The simulations were performed for libraries with low (10%), intermediate (50%) and high fractions of clonal reads (90%). We only considered the worst-case scenarios in which the alternative allele behaves without any bias and that reads mapping to the reference allele cover URSSs that are not shared with URSSs from the alternative allele, thus decreasing the likelihood that the joint-allele filter can detect a biased site. The latter assumption is often negated at sites that are covered by many unique reads (i.e. #URSS-ref-allele + #URSS-alt-allele >72). For each simulated site, we calculated the number of required URSSs at which the joint-allele filter flags a site as amplification biased. The resulting fold difference in the number of URSSs required to flag a site (i.e. #URSS-allele-specific-filter/#URSS-joint-allele-filter) indicates to what extent an AS bias needs to be more pronounced to be detectable with the joint-allele filter.

## 2.7 Genomic distribution of AS binding sites

We used the GENCODE transcript annotation version 8 (Harrow *et al.*, 2012) to identify SNPs located within promoters of protein-coding transcripts. We defined the promoter region of transcripts as ±1 kb around transcription start sites.

# 3 RESULTS

In the context of a large-scale project aiming to study allelic variability of gene regulatory processes (Kilpinen *et al.*, 2013), we generated genome-wide RNA polymerase II DNA binding profiles from two parent–daughter trios sequenced as part of the 1000 genomes project (pilot 2) using ChIP-seq and targeting the second largest RNA polymerase II subunit RPB2 (POLR2B). We obtained on average 227 million reads/individual (36 bp, single end; Section 2.1), of which 73% were usable for AS analysis. To identify ASB activity within individuals, i.e. preferential binding of POLR2B to the maternal or paternal allele, we focused on 57 241 regions (peaks) that showed significant POLR2B enrichment (Kilpinen *et al.*, 2013).

We tested for ASB at heterozygous SNP sites (1000 Genomes Project Consortium, 2010) that were located within POLR2B peaks. Only SNPs covered by at least 20 reads/individual were considered. This initial set of putative ASB sites was further filtered for sites susceptible to various sources of artifacts, such as incorrect or poor mapping of reads containing SNPs (Kilpinen *et al.*, 2013; Montgomery *et al.*, 2010). For each SNP site, we counted the number of overlapping reads carrying either the reference or alternative allele and applied a modified binomial test that corrects for the reference allele mapping bias (Kilpinen *et al.*, 2013; Montgomery *et al.*, 2010) on the read count data to identify sites with a significant deviation from the null hypothesis of equal DNA binding. Among 2082–4446 accessible SNPs/individual, we identified 34–53% significant ASB events ($P<0.01$). These results suggest that variability in POLR2B binding is extensive and consistent with earlier observations (Kasowski *et al.*, 2010; Reddy *et al.*, 2012).

However, after plotting the reference allele ratio, i.e. the fraction of reads carrying the reference as opposed to the alternative allele at SNP sites, we observed a surprising and substantial bias toward monoallelic (MA) or near-MA DNA binding with varying levels across individuals (Fig. 1, Supplementary Fig. S1), a finding that to our knowledge has not been reported in ASB studies so far. More specifically, 8–35% of the statistically significant ASB sites demonstrated MA DNA binding, and the fraction of MA ASB sites increased using more stringent *P*-value cutoffs (e.g. 12–42% at $P=0.001$; 17–49% at $P=0.0001$) (Fig. 2A). The varying levels of MA sites across the six individuals might either point to an important, yet poorly characterized biological effect (Gimelbrant *et al.*, 2007), or might simply reflect a technical artifact, as we sequenced each library on a single lane of Illumina HiSeq2000 machine, resulting in high sequencing depth.

Because many known biases were already accounted for in Kilpinen *et al.* (2013), we looked for alternative causes that might lead to highly significant MA DNA binding activity. First, we noticed that only 10% of all significant MA sites were located on the X chromosome in females and MA binding occurred in males as well, thus refuting a simple hypothesis that the majority of MA sites are due to X chromosome inactivation mechanisms (McDaniell *et al.*, 2010). We observed that total read depth at autosomal significant MA ASB sites ($P<0.01$) was 3-fold lower than at biallelic sites, suggesting that sequencing depth may affect the detection of MA binding events (Supplementary Fig. S2). To explore this topic further, we randomly sampled 25–200 million mapped reads (in incremental 25 million steps) from two POLR2B libraries that showed the strongest MA bias and observed that the proportion of MA sites among all tested sites increased with sequencing depth (Supplementary Fig. S3). For example, the proportion of MA sites was up to 10-fold lower in a low (25 million) compared with a high (200 million) sequencing depth library. These results illustrate that the observed MA bias is at least in part related to sequencing depth. However, we noted that the proportion of MA sites was sample-specific because 2–20% of significant MA sites exhibited high depth (>50 reads), which means that this bias cannot be avoided by simply increasing the read depth threshold at which sites are considered for ASB analysis.

Therefore, we further inspected the local read alignment distribution around putative ASB sites, which revealed that many sites are covered by an unexpectedly high number of clonal reads, i.e. reads mapping to the same strand and having an identical alignment start position (Fig. 2B). This might substantially bias the read count toward one allele. More specifically, 10–84% and 4–20% of all significant MA and biallelic sites were covered

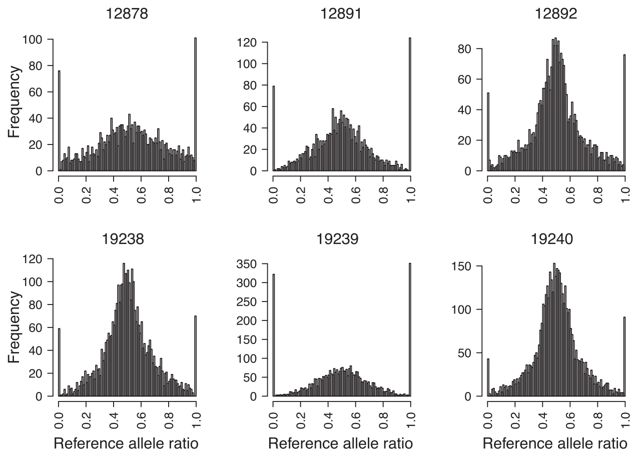

**Fig. 1.** POLR2B reference allele ratio distribution. Distribution of the reference allele ratio across all tested heterozygous SNP sites. Numbers above each panel refer to the identifier of the individual

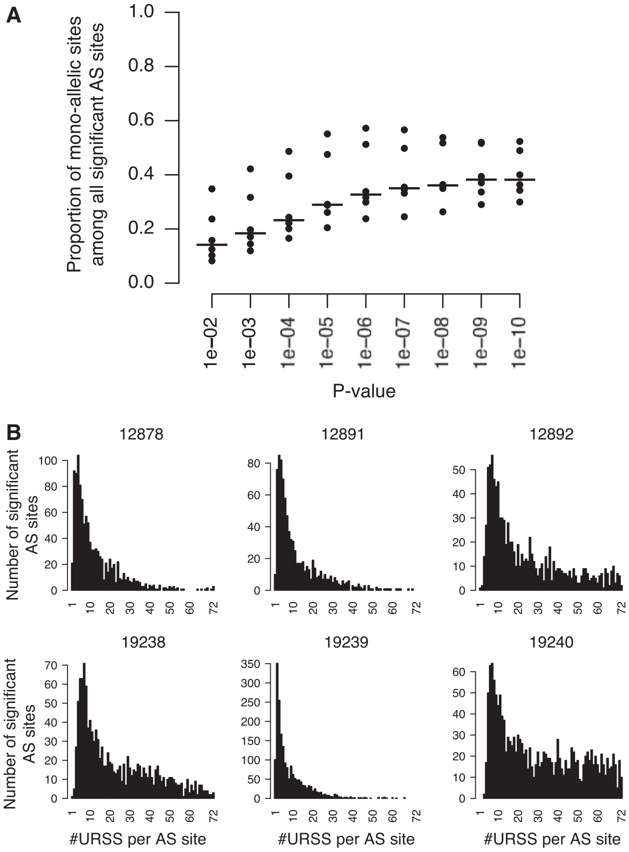

**Fig. 2.** Monoallelic POLR2B sites. (**A**) Proportion of MA sites among significant ASB sites using different ASB test *P*-value thresholds. Dots show data from all six individuals. Horizontal lines indicate the median. (**B**) Histogram of URSSs for significant ASB sites ($P<0.01$)

by only three or fewer unique reads, respectively. Grouping all testable sites by their number of URSSs and calculating for each URSS bin the proportion of significant ASB events further illustrated this trend (Supplementary Fig. S4). Sites with few URSSs correspond to highly biased ASB sites, as 94, 60, 50 and 48% of those with only 1, 2, 3 or 4 URSSs were predicted as significant, respectively.

To better understand and to account for this bias, we assessed the *clonality* of each of the six ChIP-seq libraries, which is a direct measure of the library complexity. For this purpose, we calculated how many URSSs could be expected within a region spanning twice the read length centered on an SNP site, given a certain assumed percentage of clonal (duplicate) reads at each site, and a specific SNP read depth (Fig. 3A). Specifically, we simulated sets of SNP sites with a broad range of read depth values (1–2000 reads) and various degrees of clonal signals (0–99.5% per site), thus creating an empirical distribution of URSSs (Section 2.2). Figure 3B shows the distribution of mean URSSs profiles across all 1000 simulated sites for different read depth values and clonality simulation models. For example, at an SNP site covered by exactly 20 reads, we would expect 17 URSSs when all reads are independent, but only two URSSs when 90% of the read population is clonal. These results indicate that low library complexity might lead to the presence of many sites with high depth, albeit with little positional variability around the SNP site, and as such constituting a likely source of strong bias in ASB analysis.

To further explore this phenomenon across real ChIP-seq libraries, we plotted for each individual the actual total read depth and number of URSSs per heterozygous SNP site. To our surprise, we observed considerable variation across libraries in the number of URSS per SNP (Fig. 3C), despite being sequenced to approximately similar depth. For example, SNPs accessible for ASB analysis and covered by exactly 50 reads are on average composed of 4–13 non-clonal reads/individual. This variability in independent ChIP-DNA fragments at putative ASB sites leads most likely to different levels of support for the actual ASB event, as many significant ASB sites are covered by few non-clonal reads or exhibit an unexpectedly low number of non-clonal reads even at high read depth levels. Importantly, this bias cannot be addressed by simply choosing a higher minimum read depth cutoff per site, as the degree of clonality varies widely from library to library (Supplementary Fig. S5; discussed in more detail later in the text). Therefore, this bias may have strong influence on any downstream ASB analysis, as the overall confidence in those ASB sites is rather low.

One potential and simple strategy to deal with this problem is to remove all clonal reads. However, such a strategy would lead to a situation where, for 36-bp reads, the maximal read depth per site would be 72 and the corresponding smallest statistically detectable allelic imbalance would be 1.96-fold at $P<0.01$ (binomial test). As such, removing all duplicate reads would result in 70 (median) significant ASB sites per individual ($P<0.01$), which would decrease the total number of ASB sites by >90% compared with the original and filtered set of sites. Thus, removing clonal reads substantially reduces the statistical power to detect moderate effect ASB events.

We next studied the degree of clonality of the actual libraries and determined a best fit between our simulated URSS and the

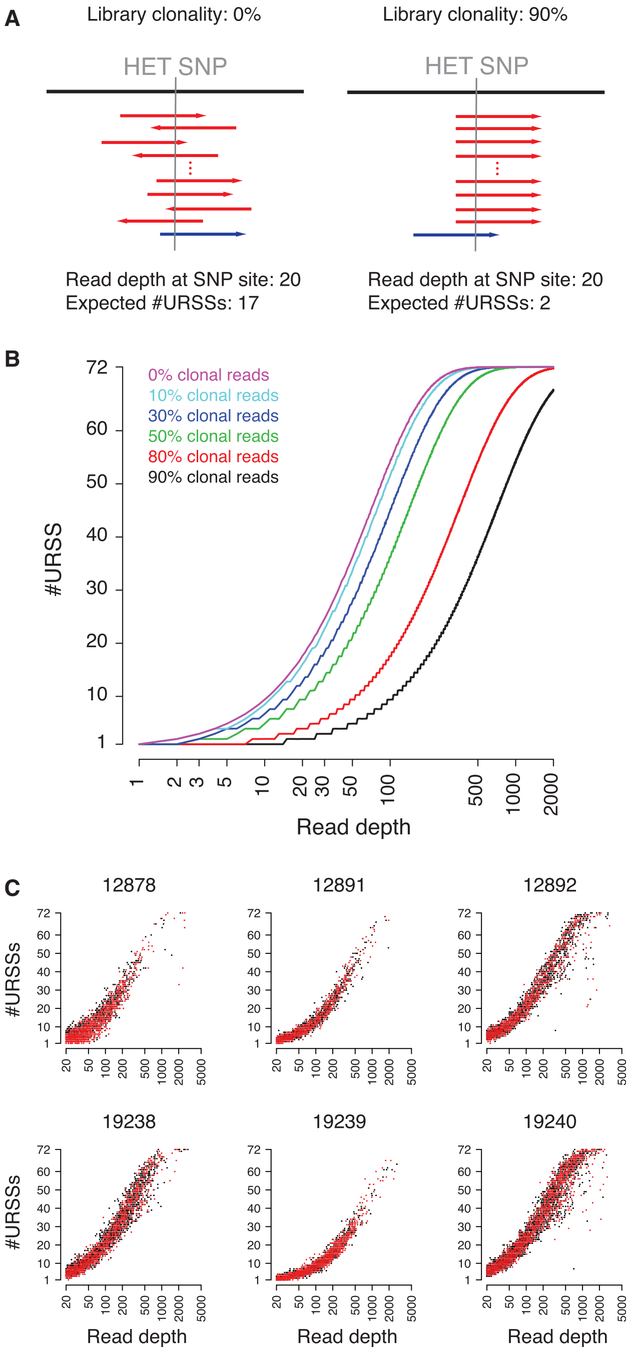

**Fig. 3.** Unique read start alignment start site simulation. (**A**) Schematic of the effect of low (0%) and high (90%) library clonality on the distribution of URSSs around a heterozygous (HET) SNP covered by 20 reads. The site represents a case of strong allelic bias whereby 19 reads support the reference allele (depicted in red) and one read the alternative allele (depicted in blue). Arrowheads indicate the strand. (**B**) Simulation of 1000 SNP sites/read depth value (*x*-axis) and sampling of a theoretical distribution of URSSs for each value from 1 to 2000 reads (in single read steps) given a defined percentage of clonal reads. Lines show the mean number of URSSs for each read depth value and for different clonality simulation models. (**C**) POLR2B URSS profiles. Total read depth (*x*-axis) at heterozygous SNP sites (dots) is plotted against the observed number of URSSs (*y*-axis). Red dots show significant ASB sites ($P<0.01$)

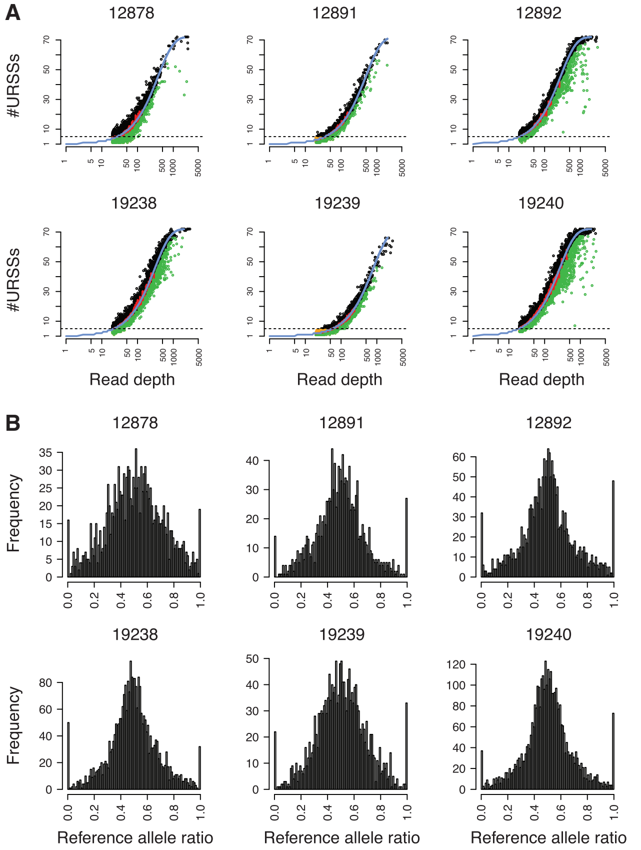

**Fig. 4.** Low-complexity site filtering. (**A**) Theoretical URSS profiles (blue line) fitted into POLR2B URSS data. The red line shows the mean URSS profile for each read depth value. SNPs that did not pass the low-complexity filtering approach are color coded: orange, global filter; green, local filter. Horizontal dashed line indicates the global filtering cutoff. (**B**) POLR2B reference allele ratio distribution after filtering for low-complexity SNP sites

real URSS distributions for each individual (Fig. 4A). We used the KS test for equality of the simulated versus the observed mean URSS distributions. The simulated distribution that yields the lowest KS statistic indicates the clonality of the actual ChIP-seq library (Section 2.3). Using this workflow, we determined that the genome-wide clonality of the six ChIP-seq libraries ranged between 73 and 91% (Supplementary Fig. S5). Thus, these libraries feature a high number of clonal reads due to deep sequencing, further emphasizing the need to account for this substantial bias. To validate our simulations, we determined an independent measure of the extent of clonality (Section 2.4). The Spearmans rank correlation coefficient between both estimates of library complexity revealed a high consistency ($\rho = 0.989$, $P = 1.9\text{e-}4$). We then identified ASB sites that feature a significantly greater number of mapped clonal reads than expected based on the determined library-specific clonality model. Biased sites, as indicated earlier in the text, have a high tendency to survive the significance test for AS binding.

We devised a two-tiered, *global* and *local* filtering strategy to account for the amplification bias. The global filter is sequencing depth independent and flags any site that is covered by reads with less than a specified number of URSS (Section 2.5). For the POLR2B datasets, requiring at least five URSSs per SNP site removed the majority of highly biased ASB events, i.e. on average 1–43% of all accessible ASB sites and 2–92% of all MA sites per individual. The local filter is sequencing depth dependent and flags sites that show an unexpected low number of URSSs. Given the library-specific clonality simulations, we can model the expected URSS distribution for each read depth bin (1–2000) and test whether the observed number of URSSs per SNP site is within the expected range of URSSs. This approach flags sites as amplification-biased that are within the left tail of the empirical URSS distribution (i.e. having significantly fewer URSSs than expected). Because the local filter models URSSs over both alleles jointly, we have performed additional simulations to test if the joint-allele filter is able to identify sites when only a single allele experiences an amplification bias (Section 2.6). Dependent on the read depth per site and library clonality, we observed that the joint-allele strategy detects a pure AS bias if the amplification bias is only borderline above the threshold of an AS filter (Supplementary Fig. S6). After applying the local filter on the POLR2B libraries, we identified that an additional 7–22% of all tested sites exhibit an excess of clonal reads and thus likely constitute biased sites. Together, both filters (local and global) flagged 19–51% of all initially tested SNPs per individual as showing low-complexity read alignment start site patterns around SNP sites (Fig. 4A), thus revealing that low-complexity biases are common in the tested libraries. After removing these sites, we observed that the reference allele ratio distribution exhibited only few sites with MA binding behavior per individual (2.4–3.9%; Fig. 4B). The filtered set of SNPs showed at 38, 27 and 20% allelic effects (median across individuals) at $P$-value cutoffs of 0.01, 0.001 and 0.0001, respectively (binomial test). In contrast to the original set of sites, MA sites accounted for only 5–10% of all significant ASB events at $P < 0.01$. Importantly, not all MA sites should be considered as false positives. MA gene expression is a known biological process and can be caused, for example, by parent-of-origin silencing of autosomal genes (imprinting), random gene silencing, or X-inactivation in females (Gimelbrant *et al.*, 2007; Li *et al.*, 2012), although it is clear that such events should only constitute a small portion of all ASB events and not the majority as observed before filtering. In this regard, it is worth noting that in comparison with significant post-filtered ASB sites, pre-filtered MA sites are depleted from promoters (54 and 16%). This suggests that biased MA binding events for POLR2B might reflect stochastic transcription events that tend to be covered by few DNA fragments and become visible as MA at high sequencing depth.

Finally, we set out to provide validation for our proposed filtering approach. First, we tested if allelic directions are consistent at shared significant ASB sites between two unrelated individuals (McDaniell *et al.*, 2010) and whether the consistency improves after removing biased sites (Fig. 5A). The analysis shows that, although the consistency in allelic direction is highly significant for all tested sites ($\rho = 0.28$, $P = 1.2\text{e-}11$; Spearman correlation test), it substantially improves after removing biased sites ($\rho = 0.46$, $P = 9.6\text{e-}16$), consistent with it being lower for biased sites alone ($\rho = 0.14$, $P = 0.02$). Second, we tested whether two independent, phased and significant SNP sites located within the same 200 bp POLR2B peak show

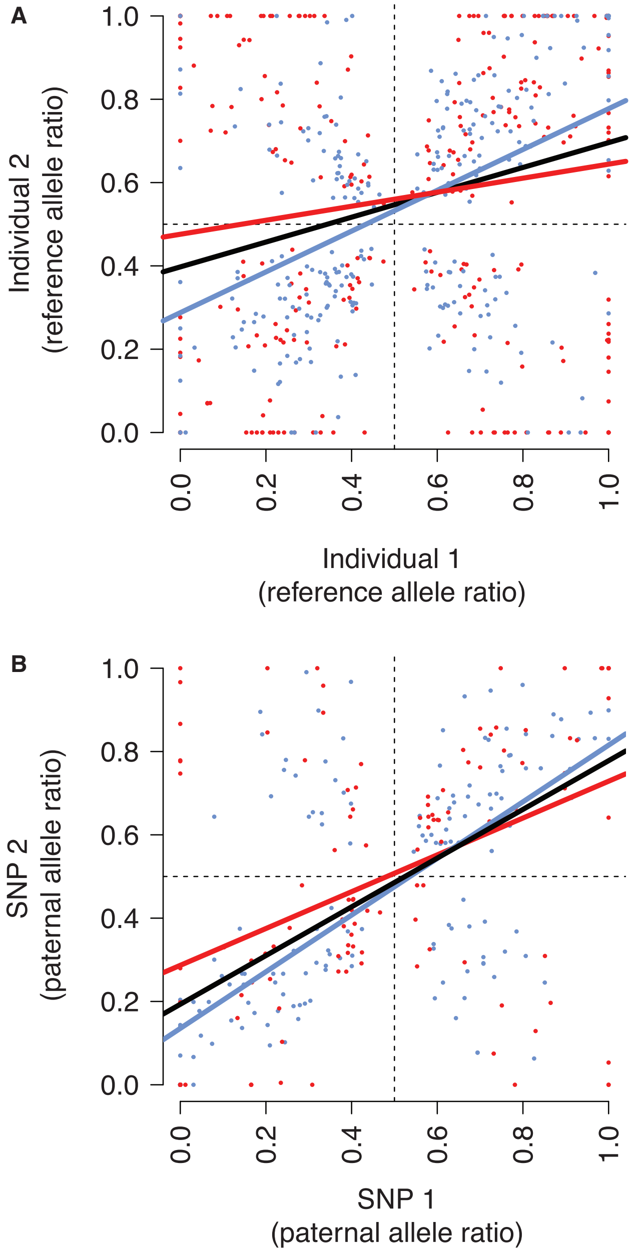

**Fig. 5.** Consistency of allelic direction before and after filtering for amplification-biased sites. (**A**) Reference allele ratios at shared ASB sites between pairs of unrelated individuals. Sites were accessible and significant in both individuals. Data from all paired individuals were pooled. Blue and red dots correspond to unbiased and biased sites, respectively. Colored lines show linear regression fits using all testable sites before filtering (black), unbiased sites (blue) and sites biased in one or both individuals (red), respectively. (**B**) Paternal allele ratios of significant ASB site pairs that are located within POLR2B peaks. Data from all individuals were pooled. Blue and read dots correspond to unbiased and biased site pairs, respectively. Colored lines show linear regression fits using all testable site pairs (black), unbiased site pairs (blue) and site pairs with bias in one or both sites(s) (red), respectively

consistent behavior in allelic direction. To avoid artificial correlations, we only considered SNP pairs separated by 37–200 bp. We observed that the allelic direction of sites located within the same peak is strongly correlated ($\rho = 0.58$, $P = 2.2e\text{-}16$). In agreement with results from the previous analysis, we found that this correlation is greater for unbiased SNP pairs than for pairs in which one or both sites are biased ($\rho = 0.71$ and $\rho = 0.42$, $P = 2.2e\text{-}16$ and 1.4e-6, respectively) (Fig. 5B). Both analyses remained valid even after excluding sites that reached significance without duplicate reads (data not shown). Therefore, these results support the validity of ASB analysis using clonal reads as well as the effectiveness in identifying biased sites with our filtering approach.

## 4 DISCUSSION

It is now well recognized that high-throughput sequencing assays are susceptible to significant, often non-intuitive artifacts when used to detect AS molecular events (Degner *et al.*, 2009; Pickrell *et al.*, 2011; Rozowsky *et al.*, 2011). Here we report another such artifact intrinsic to deeply sequenced ChIP-seq libraries and likely other sequencing assays subject to PCR amplification of a small amount of starting material. Specifically, we demonstrate that on identifying putative sites with AS POLR2B occupancy, we retain a clear excess number of sites that exhibit monoallelic DNA binding behavior. We show that many of these sites suffer from PCR amplification biases, likely introduced during standard Illumina library preparation. In a systematic study of base-composition biases in Illumina sequencing libraries, PCR during library preparation was identified as the principal bias source (Aird *et al.*, 2011). Although experimental strategies have been proposed to reduce this bias, it is widely accepted that it will be difficult to completely eliminate it, unless PCR amplification is avoided altogether. However, PCR-free libraries require a substantial amount of input material, rendering this rather unfeasible for approaches such as ChIP-seq where the sample material is often limiting (Deplancke, 2009). Consequently, PCR amplification bias should be accounted for when using ChIP-seq to detect AS molecular events involving transcription factors, cofactors, and histone modifications. As observed in Kilpinen *et al.* (2013), all ChIP-seq libraries suffered from this bias, although the extent differed substantially between factors and libraries with 8–26% (first and third quartile) of all accessible heterozygous SNP sites exhibiting low read alignment complexity across nine additional assays profiled in the same individuals as POLR2B. The later analysis included a previously published dataset on CCCTC-binding factor (CTCF) binding (McDaniell *et al.*, 2010) in which 11% of all accessible sites for AS analysis were flagged as biased. Here we show that one of the sources of this bias is sequencing depth given that the proportion of MA binding events increases with increasing sequencing depth. One simple strategy to deal with this issue is to simply remove all clonal reads from the dataset (Chen *et al.*, 2012; Heap *et al.*, 2010). However, previous and our own studies demonstrate that many clonal reads likely constitute true signals (Chen *et al.*, 2012). Removing all duplicate reads from the POLR2B libraries leads to a strong reduction in the overall number of significant ASB sites as well as to a depletion in sites with moderate ASB signals. Irrespective of the strategy, removing all clonal reads seems a rather overly cautious way of bypassing the PCR amplification bias and is in fact counterproductive as it lowers the overall statistical power to detect ASB sites and unnecessarily discards much of the sequencing data.

## 5 CONCLUSION

We have presented a method to identify and reduce the bias of clonal amplification in AS analysis of ChIP-seq data. This approach analyzes the read alignment distribution around heterozygous SNP sites and removes highly clonal, low-complexity sites based on expectations about the distribution of reads covering SNP sites. Given the current interest in understanding the interplay between DNA variants and gene regulation, this approach should be of general interest.

## ACKNOWLEDGEMENT

The authors would like to thank Andreas Massouras, Petra Schwalie, Zoltán Kutalik, Frédéric Schütz and the three anonymous reviewers for their insightful comments on the manuscript.

*Funding*: Swiss National Science Foundation [CRSI33_130326 to E.T.D., B.D., A.R. and N.H.; 31003A_132958 to N.H; 31003A_130342 to E.T.D.; 31003A_129835 to A.R.]; the European Research Council [to E.T.D.]; the Louis-Jeantet Foundation [to E.T.D.]; the European Molecular Biology Organization [EMBO ALTF 2010-337 fellowship to H.K.]; the doctoral school of the Faculty of Biology and Medicine, University of Lausanne [fellowship to R.M.W.]; the NCCR Frontiers in Genetics Program [to B.D., E.T.D.]; the ETHZ Japanese-Swiss Science and Technology Cooperation Program [to B.D.]; and institutional support from the École Polytechnique Féderale de Lausanne [to B.D.]. The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

*Conflict of Interest*: none declared

## REFERENCES

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

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Chen,Y. *et al.* (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods*, **9**, 609–614.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Degner,J.F. *et al.* (2012) DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–394.

Deplancke,B. (2009) Experimental advances in the characterization of metazoan gene regulatory networks. *Brief. Func. Genomic Proteomics*, **8**, 12–27.

Gimelbrant,A. *et al.* (2007) Widespread monoallelic expression on human autosomes. *Science*, **318**, 1136–1140.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

Heap,G.A. *et al.* (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.*, **19**, 122–134.

Kasowski,M. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

Kilpinen,H. *et al.* (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, 744–747.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,S.M. *et al.* (2012) Transcriptome-wide survey of mouse CNS-derived cells reveals monoallelic expression within novel gene families. *PLoS One*, **7**, e31751.

McDaniell,R. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

Montgomery,S.B. and Dermitzakis,E.T. (2011) From expression QTLs to personalized transcriptomics. *Nat. Genet.*, **12**, 277–282.

Montgomery,S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.

Pickrell,J.K. *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

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

Reddy,T.E. *et al.* (2012) Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.*, **22**, 860–869.

Rozowsky,J. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.