

Additional File 1 for iASeq: Supplemental Methods

S.1 Data preprocessing

S.1.1 Data collection

Both ChIP-seq and RNA-seq data for GM12878 cells were downloaded from the ENCODE Project [1]. The datasets used in the analysis are summarized in Additional File 2 Supplemental Table 1. For each ChIP-seq sample, we downloaded the FASTQ file and mapped the raw sequence reads to human genome (hg18) using MAQ (Version 0.7.1) with default parameters [2]. Uniquely mapped reads with the mapping quality score above 0 were extracted.

S.1.2 Heterozygote SNP collection, bias filtering and protein binding filtering

The genotype data for GM12878 was retrieved from the [3]. A total of 1,704,166 heterozygote SNPs were obtained. As shown by [4], in addition to the reference mapping bias (i.e., reference alleles are easier to be mapped back), certain SNPs have intrinsic biases toward one of the alleles. For these SNPs, genomic DNA extracted computationally from one allele (not necessarily the reference allele) is intrinsically easier to be mapped back compared to DNA extracted from the other allele. Using the method described in [4, 5], we identified and removed these 149,996 intrinsically biased SNPs. Next, we called ChIPseq peaks for each dataset using CisGenome [6] at 1% FDR level and retained only those heterozygous SNPs without intrinsic bias located in the protein binding regions in at least one dataset. After these two steps of filtering, 94,519 SNPs were included in our subsequent analysis.

S.1.3 Collection of exonic SNPs and ASE SNPs

The exonic ASE SNPs were used as one of our gold standards to evaluate iASeq and other methods. To determine which SNPs are exonic, we downloaded the hg18 Ensemble gene annotation file Homo_sapiens.NCBI36.54.gtf from [7]. Exonic SNPs were annotated using the exonic regions from

the gene annotation file. The exonic ASE SNPs were determined using RNA-seq. For a given RNA-seq dataset, the naive Bayes statistic was calculated for each exonic SNP. The top 400 exonic SNPs (6.61% of the total 6051 exonic SNPs) ranked based on the naive Bayes statistic were called as exonic ASE SNPs. Subsequently, in the analysis of ChIP-seq data, a SNP was claimed to be a true positive if there was an exonic ASE SNP within its Xkb neighborhood. Results for X=10kb are shown in the main manuscript and results for X=1kb are shown in Additional File 7. Both criteria gave similar results.

S.2 Method of moment for estimating parameters in the Beta distribution

For a Beta distribution $Beta(\alpha, \beta)$, the mean is $\alpha/(\alpha+\beta)$, and the variance is $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. For sample (d, j) , each SNP has a p_{idj} which can be roughly estimated as $\hat{p}_{idj} = (x_{idj} + 2 * p_{dj0}^{(0)}) / (n_{idj} + 2)$. Here $p_{dj0}^{(0)} = \frac{1}{I_d^{(0)}} \sum_{i:n_{idj} \neq 0} x_{idj} / n_{idj}$, and $I_d^{(0)}$ is the number of SNPs in dataset d for which $n_{idj} \neq 0$. Let $p_{dj0} = \sum_i \hat{p}_{idj} / I$, and $v_{dj} = \frac{1}{I-1} \sum_i (p_{idj} - \hat{p}_{idj})^2$. By matching p_{dj0} and v_{dj} to the theoretical mean and variance of a Beta distribution, we obtain

$$\hat{\alpha}_{dj} = p_{dj0} * \left[\frac{p_{dj0}(1 - p_{dj0})}{v_{dj}} - 1 \right] \quad (\text{S.1})$$

$$\hat{\beta}_{dj} = (1 - p_{dj0}) * \left[\frac{p_{dj0}(1 - p_{dj0})}{v_{dj}} - 1 \right] \quad (\text{S.2})$$

In principle, one may develop a more sophisticated algorithm to estimate α and β by fitting beta-binomial distributions to $x_{idj} | n_{idj}$, but the computation will be more involved. Therefore we did not pursue this solution and instead used the simple method described above to approximately estimate α and β .

S.3 Parameter choice for the Dirichlet prior

Although $\eta = 1$ can specify an uniform prior and seems to be a natural choice, it will make the EM algorithm numerically unstable. This is because the EM searches for posterior mode and is implemented on log scale. The mode of a Dirichlet distribution $D(\eta_1, \dots, \eta_M)$ for the m -th component is $(\eta_m - 1) / \sum_{l=1}^M (\eta_l - 1)$. It is not defined if all η_m s are equal to one. As a result, if $\eta = 1$ is used as the prior, and when the expectation of the counts $\sum_i \delta(a_i = k)$, $\sum_i \delta(a_i = k) b_{id}$ or $\sum_i \delta(a_i = k) c_{id}$ in the E-step of the algorithm is close to zero, then the algorithm can easily lose

its numerical stability due to issues such as $\log(0)$ and ill-defined posterior mode. These issues can be avoided by using $\eta = 2$ which still imposes a relatively non-informative prior.

S.4 The EM algorithm used in iASeq

This section presents the EM algorithm used to search for posterior mode $(\hat{\pi}, \hat{V}, \hat{W})$ of the distribution $Pr(\pi, V, W|X, N) = \sum_{A, B, C} Pr(A, B, C, \pi, V, W|X, N)$. In the EM algorithm, A , B and C are the missing data. The algorithm iterates between an E-step and an M-step.

In the E-step, one evaluates the Q-function $Q(\pi, V, W|\hat{\pi}^{old}, \hat{V}^{old}, \hat{W}^{old})$ which is defined as $E_{old}[\ln Pr(A, B, C, \pi, V, W|X, N)]$. Here the expectation is taken with respect to probability distribution $Pr(A, B, C|X, N, \hat{\pi}^{old}, \hat{V}^{old}, \hat{W}^{old})$, abbreviated as $Pr_{old}(A, B, C)$, where $\hat{\pi}^{old}$, $\hat{V}^{old}$ and $\hat{W}^{old}$ are the parameter estimates obtained from the last iteration.

When we use $\eta = 2$ in the Dirichlet priors for π , V and W , we have

$$\begin{aligned}
& \ln Pr(A, B, C, \pi, V, W|X, N) \\
&= \sum_{i=1}^I \{ \delta(a_i = 0) (\ln \pi_0 + \sum_{d=1}^D \ln L_{id0}) \\
&+ \sum_{k=1}^K \delta(a_i = k) [\ln \pi_k + \sum_{d=1}^D b_{id} (\ln v_{kd} + \ln L_{id1}) + \sum_{d=1}^D c_{id} (\ln w_{kd} + \ln L_{id2}) \\
&+ \sum_{d=1}^D (1 - b_{id} - c_{id}) (\ln(1 - v_{kd} - w_{kd}) + \ln L_{id0})] \} \\
&+ \sum_{k=0}^K \ln \pi_k + \sum_{k=1}^K \sum_{d=1}^D [\ln v_{kd} + \ln w_{kd} + \ln(1 - v_{kd} - w_{kd})] \\
&+ \text{constant}
\end{aligned} \tag{S.3}$$

Therefore,

$$\begin{aligned}
& Q(\pi, V, W|\hat{\pi}^{old}, \hat{V}^{old}, \hat{W}^{old}) \\
&= E_{old}[\ln Pr(A, B, C, \pi, V, W|X, N)] \\
&= \sum_{k=0}^K \{ \sum_{i=1}^I E_{old}[\delta(a_i = k)] + 1 \} \ln \pi_k + \sum_{k=1}^K \sum_{d=1}^D \{ (\sum_{i=1}^I E_{old}[\delta(a_i = k) b_{id}] + 1) \ln v_{kd} \\
&+ (\sum_{i=1}^I E_{old}[\delta(a_i = k) c_{id}] + 1) \ln w_{kd} + (\sum_{i=1}^I E_{old}[\delta(a_i = k) (1 - b_{id} - c_{id})] + 1) \ln(1 - v_{kd} - w_{kd}) \} \\
&+ \text{constant}
\end{aligned} \tag{S.4}$$

In the M-step, one finds π , V and W that maximize the Q-function $Q(\pi, V, W|\hat{\pi}^{old}, \hat{V}^{old}, \hat{W}^{old})$. Denote them by $\hat{\pi}^{new}$, $\hat{V}^{new}$ and $\hat{W}^{new}$. These will give the new parameter estimates.

By solving

$$\frac{\partial Q(\boldsymbol{\pi}, \mathbf{V}, \mathbf{W} | \hat{\boldsymbol{\pi}}^{old}, \hat{\mathbf{V}}^{old}, \hat{\mathbf{W}}^{old})}{\partial \pi_k} = 0 \quad (\text{S.5})$$

$$\frac{\partial Q(\boldsymbol{\pi}, \mathbf{V}, \mathbf{W} | \hat{\boldsymbol{\pi}}^{old}, \hat{\mathbf{V}}^{old}, \hat{\mathbf{W}}^{old})}{\partial v_{kd}} = 0 \quad (\text{S.6})$$

$$\frac{\partial Q(\boldsymbol{\pi}, \mathbf{V}, \mathbf{W} | \hat{\boldsymbol{\pi}}^{old}, \hat{\mathbf{V}}^{old}, \hat{\mathbf{W}}^{old})}{\partial w_{kd}} = 0 \quad (\text{S.7})$$

We have

$$\hat{\pi}_k^{new} = \frac{\sum_{i=1}^I Pr_{old}(a_i = k) + 1}{I + K + 1} \quad (\text{S.8})$$

$$\hat{v}_{kd}^{new} = \frac{\sum_{i=1}^I Pr_{old}(a_i = k, b_{id} = 1) + 1}{\sum_{i=1}^I Pr_{old}(a_i = k) + 3} \quad (\text{S.9})$$

$$\hat{w}_{kd}^{new} = \frac{\sum_{i=1}^I Pr_{old}(a_i = k, c_{id} = 1) + 1}{\sum_{i=1}^I Pr_{old}(a_i = k) + 3} \quad (\text{S.10})$$

In the formulas above, $Pr_{old}(a_i = k)$, $Pr_{old}(a_i = k, b_{id} = 1)$ and $Pr_{old}(a_i = k, c_{id} = 1)$ are computed as follows. To compute $Pr_{old}(a_i = k) = Pr(a_i = k | \mathbf{X}_i, \mathbf{N}_i, \hat{\boldsymbol{\pi}}^{old}, \hat{\mathbf{V}}^{old}, \hat{\mathbf{W}}^{old})$, recall

$$Pr(a_i | \mathbf{X}_i, \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) = Pr(\mathbf{X}_i, a_i | \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) / Pr(\mathbf{X}_i | \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) \quad (\text{S.11})$$

From Equation 5 in the main manuscript, we have

$$\begin{aligned} Pr(\mathbf{X}_i, a_i | \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) &= \sum_{\mathbf{B}_i, \mathbf{C}_i} Pr(\mathbf{X}_i, a_i, \mathbf{B}_i, \mathbf{C}_i | \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) \\ &= \{\pi_0 \prod_{d=1}^D L_{id0}\}^{\delta(a_i=0)} \prod_{k=1}^K \{\pi_k \prod_{d=1}^D [v_{kd} L_{id1} + w_{kd} L_{id2} + (1 - v_{kd} - w_{kd}) L_{id0}]\}^{\delta(a_i=k)} \end{aligned} \quad (\text{S.12})$$

and

$$\begin{aligned} Pr(\mathbf{X}_i | \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) &= \sum_{a_i} Pr(\mathbf{X}_i, a_i | \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) \\ &= \pi_0 \prod_{d=1}^D L_{id0} + \sum_{k=1}^K \{\pi_k \prod_{d=1}^D [v_{kd} L_{id1} + w_{kd} L_{id2} + (1 - v_{kd} - w_{kd}) L_{id0}]\} \end{aligned} \quad (\text{S.13})$$

Therefore, $Pr_{old}(a_i = k)$ can be computed by replacing $\boldsymbol{\pi}$, $\mathbf{V}$ and $\mathbf{W}$ by $\hat{\boldsymbol{\pi}}^{old}$, $\hat{\mathbf{V}}^{old}$ and $\hat{\mathbf{W}}^{old}$.

$Pr_{old}(a_i = k, b_{id} = 1) = Pr_{old}(a_i = k) Pr_{old}(b_{id} = 1 | a_i = k)$. $Pr_{old}(a_i = k)$ is computed as above. However, $Pr(b_{id}, c_{id} | a_i = k, \mathbf{X}_i, \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W})$ can be computed as

$$\begin{aligned} \frac{Pr(b_{id}, c_{id}, \mathbf{X}_i | a_i = k, \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W})}{Pr(\mathbf{X}_i | a_i = k, \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W})} &= \frac{Pr(b_{id}, c_{id}, \mathbf{X}_{id} | a_i = k, \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W})}{\sum_{b_{id}, c_{id}} Pr(b_{id}, c_{id}, \mathbf{X}_{id} | a_i = k, \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W})} \\ &= \frac{[v_{kd} L_{id1}]^{b_{id}} [w_{kd} L_{id2}]^{c_{id}} [(1 - v_{kd} - w_{kd}) L_{id0}]^{1-b_{id}-c_{id}}}{v_{kd} L_{id1} + w_{kd} L_{id2} + (1 - v_{kd} - w_{kd}) L_{id0}} \end{aligned} \quad (\text{S.14})$$

$Pr_{old}(b_{id} = 1 | a_i = k)$ and $Pr_{old}(c_{id} = 1 | a_i = k)$ can be obtained by plugging in $\hat{\boldsymbol{\pi}}^{old}$, $\hat{\mathbf{V}}^{old}$ and $\hat{\mathbf{W}}^{old}$ into the formula above to replace $\boldsymbol{\pi}$, $\mathbf{V}$ and $\mathbf{W}$.

S.5 Bayesian Information Criterion (BIC) for choosing K

We compute BIC as

$$\begin{aligned}
 BIC(K) &= -2 * \ln \left\{ \prod_{i=1}^I Pr(\mathbf{X}_i | \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) \right\} + (K + 2 * K * D) * \ln I \\
 &= -2 * \sum_{i=1}^I \ln \left[\pi_0 \prod_{d=1}^D L_{id0} + \sum_{k=1}^K \left\{ \pi_k \prod_{d=1}^D [v_{kd} L_{id1} + w_{kd} L_{id2} + (1 - v_{kd} - w_{kd}) L_{id0}] \right\} \right] \\
 &\quad + K(2D + 1) \ln I
 \end{aligned} \tag{S.15}$$

We calculate BIC for different values of K , and choose the K with the smallest BIC. Here $K + 1$ is the number of classes. K is also the number of parameters in $\boldsymbol{\pi}$. $2KD$ is the number of parameters involved in $\mathbf{V}$ and $\mathbf{W}$. I is the SNP number. Strictly speaking, the data likelihood also involve terms $Pr(\mathbf{N}_i | \boldsymbol{\pi}, \mathbf{V}, \mathbf{W})$. However, based on our assumption, these terms do not depend on K , $\boldsymbol{\pi}$, $\mathbf{V}$ and $\mathbf{W}$, and can be reduced to $Pr(\mathbf{N}_i)$. They can be viewed as constants for the purpose of choosing the optimal K . We do not include them in the BIC computation.

S.6 Data generation in simulation studies

To simulate a ASB SNP i , we first sampled a SNP from the 8166 non-background SNPs in the real GM12878 data. Here the non-background SNPs in the real data were determined by iASeq using $Pr(a_i = 0 | \mathbf{X}_i, \mathbf{N}_i, \boldsymbol{\pi}, \mathbf{V}, \mathbf{W}) < 0.5$ as cutoff. Additionally, we also sampled a SNP from the 86,353 background SNPs in the real GM12878 data. Next, with these two real SNPs in hand, we went through each dataset d to generate the read counts for the simulated SNP. If $[b_{id}, c_{id}] = [0, 0]$, then we used the background SNP's read count data (x_{idj}, n_{idj}) in sample (d, j) to serve as the data of the simulated SNP in dataset d sample j . If $[b_{id}, c_{id}] = [1, 0]$, then we used the non-background SNP's read count data to simulate read counts as follows. For each replicate sample j in dataset d , we obtained the observed total read count n_{idj} for the non-background SNP. We then randomly drew a number p_{idj} from $U[p_{dj0}, 1]$, where p_{dj0} is the mean of $\frac{x_{idj}}{n_{idj}}$ over all background SNPs in the same sample (d, j) . Subsequently, we simulated x_{idj} from a binomial distribution $Bin(n_{idj}, p_{idj})$ to serve as the simulated SNP's data in dataset d and sample j . If $[b_{id}, c_{id}] = [0, 1]$, we applied a similar procedure but the p_{idj} was drawn from $U[0, p_{dj0}]$.

S.7 The single dataset based EM analysis

This approach analyzes each dataset separately. Let $\mathbf{X}^d = (\mathbf{X}_{1d}, \dots, \mathbf{X}_{Id})$ and $\mathbf{N}^d = (\mathbf{N}_{1d}, \dots, \mathbf{N}_{Id})$ be the data from dataset d . We assumed that in each dataset d , a SNP i can be SR ($b_{id} = 1$), SN ($c_{id} = 1$) or NS ($b_{id} = 0$ and $c_{id} = 0$) with probability $(v_d, w_d, 1 - v_d - w_d)$. Let $\mathbf{B}^d = (b_{1d}, \dots, b_{Id})$ and $\mathbf{C}^d = (c_{1d}, \dots, c_{Id})$ be the ensemble of all ASB indicators in dataset d . Adopting the same distributional assumption as in Equations 1-3 in the main manuscript, the complete data likelihood can be derived as:

$$\begin{aligned} Pr(\mathbf{X}^d, \mathbf{N}^d, \mathbf{B}^d, \mathbf{C}^d | v_d, w_d) &= Pr(\mathbf{N}^d) Pr(\mathbf{X}^d, \mathbf{B}^d, \mathbf{C}^d | \mathbf{N}^d, v_d, w_d) \\ &= Pr(\mathbf{N}^d) \prod_{i=1}^I \{ [v_d L_{id1}]^{b_{id}} [w_d L_{id2}]^{c_{id}} [(1 - v_d - w_d) L_{id0}]^{(1 - b_{id} - c_{id})} \} \end{aligned} \quad (\text{S.16})$$

By imposing a Dirichlet prior $D(2, 2, 2)$ on $(v_d, w_d, 1 - v_d - w_d)$, we obtain the posterior distribution of the unknown parameters and missing indicators:

$$Pr(\mathbf{B}^d, \mathbf{C}^d, v_d, w_d | \mathbf{X}^d, \mathbf{N}^d) \propto \prod_{i=1}^I \{ [v_d L_{id1}]^{b_{id}} [w_d L_{id2}]^{c_{id}} [(1 - v_d - w_d) L_{id0}]^{(1 - b_{id} - c_{id})} \} v_d w_d (1 - v_d - w_d) \quad (\text{S.17})$$

An EM algorithm can be similarly derived as in iASeq to estimate the parameters v_d and w_d by searching for the posterior mode of $Pr(v_d, w_d | \mathbf{X}^d, \mathbf{N}^d)$.

In the E-step, we compute the Q-function $Q(v_d, w_d | \hat{v}_d^{old}, \hat{w}_d^{old})$. Since

$$\begin{aligned} \ln Pr(\mathbf{B}^d, \mathbf{C}^d, v_d, w_d | \mathbf{X}^d, \mathbf{N}^d) &= \sum_{i=1}^I \{ b_{id} [\ln v_d + \ln L_{id1}] + c_{id} [\ln w_d + \ln L_{id2}] \\ &\quad + (1 - b_{id} - c_{id}) [\ln(1 - v_d - w_d) + \ln L_{id0}] \} + \ln v_d \\ &\quad + \ln w_d + \ln(1 - v_d - w_d) + \text{constant} \end{aligned} \quad (\text{S.18})$$

We have

$$\begin{aligned} Q(v_d, w_d | \hat{v}_d^{old}, \hat{w}_d^{old}) &= E_{old}[\ln Pr(\mathbf{B}^d, \mathbf{C}^d, v_d, w_d | \mathbf{X}^d, \mathbf{N}^d)] \\ &= \{ \sum_{i=1}^I E_{old}(b_{id}) + 1 \} \ln v_d + \{ \sum_{i=1}^I E_{old}(c_{id}) + 1 \} \ln w_d \\ &\quad + \{ \sum_{i=1}^I E_{old}(1 - b_{id} - c_{id}) + 1 \} \ln(1 - v_d - w_d) + \text{constant} \end{aligned} \quad (\text{S.19})$$

In the M-step, we find v_d and w_d that maximize $Q(v_d, w_d | \hat{v}_d^{old}, \hat{w}_d^{old})$. By solving

$$\frac{\partial Q(v_d, w_d | \hat{v}_d^{old}, \hat{w}_d^{old})}{\partial v_d} = 0 \quad (\text{S.20})$$

$$\frac{\partial Q(v_d, w_d | \hat{v}_d^{old}, \hat{w}_d^{old})}{\partial w_d} = 0 \quad (\text{S.21})$$

We obtain

$$\hat{v}_d^{new} = \frac{\sum_{i=1}^I Pr_{old}(b_{id} = 1) + 1}{I + 3} \quad (\text{S.22})$$

$$\hat{w}_d^{new} = \frac{\sum_{i=1}^I Pr_{old}(c_{id} = 1) + 1}{I + 3} \quad (\text{S.23})$$

Here

$$\begin{aligned} Pr_{old}(b_{id} = 1) &= Pr(b_{id} = 1 | \mathbf{X}_{id}, \mathbf{N}_{id}, \hat{v}_d^{old}, \hat{w}_d^{old}) = \frac{Pr(b_{id} = 1, \mathbf{X}_{id} | \mathbf{N}_{id}, \hat{v}_d^{old}, \hat{w}_d^{old})}{Pr(\mathbf{X}_{id} | \mathbf{N}_{id}, \hat{v}_d^{old}, \hat{w}_d^{old})} \\ &= \frac{\hat{v}_d^{old} L_{id1}}{\hat{v}_d^{old} L_{id1} + \hat{w}_d^{old} L_{id2} + (1 - \hat{v}_d^{old} - \hat{w}_d^{old}) L_{id0}} \end{aligned} \quad (\text{S.24})$$

$$\begin{aligned} Pr_{old}(c_{id} = 1) &= Pr(c_{id} = 1 | \mathbf{X}_{id}, \mathbf{N}_{id}, \hat{v}_d^{old}, \hat{w}_d^{old}) = \frac{Pr(c_{id} = 1, \mathbf{X}_{id} | \mathbf{N}_{id}, \hat{v}_d^{old}, \hat{w}_d^{old})}{Pr(\mathbf{X}_{id} | \mathbf{N}_{id}, \hat{v}_d^{old}, \hat{w}_d^{old})} \\ &= \frac{\hat{w}_d^{old} L_{id2}}{\hat{v}_d^{old} L_{id1} + \hat{w}_d^{old} L_{id2} + (1 - \hat{v}_d^{old} - \hat{w}_d^{old}) L_{id0}} \end{aligned} \quad (\text{S.25})$$

Using the posterior mode, one can similarly compute $Pr(b_{id}, c_{id} | \mathbf{X}^d, \mathbf{N}^d, v_d, w_d)$ and $\tilde{P}_{id}$ to detect and rank AS SNPs.

References

- [1] **The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.
- [2] Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851–1858.
- [3] **GM12878 genotype** [ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/trio/snps].
- [4] Degner J, Marioni J, Pai A, Pickrell J, Nkadori E, Gilad Y, Pritchard J: **Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.** *Bioinformatics* 2009, **25**:3207–3212.
- [5] Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, Veyrieras J, Stephens M, Gilad Y, Pritchard J: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**:768–772.

- [6] Ji H, Jiang H, Ma W, Johnson D, Myers R, Wong W: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nature Biotechnology* 2008, :1293–1300.
- [7] hg18 Ensemble gene annotation [ftp://ftp.ensembl.org/pub/release-54/gtf/homo_sapiens/].