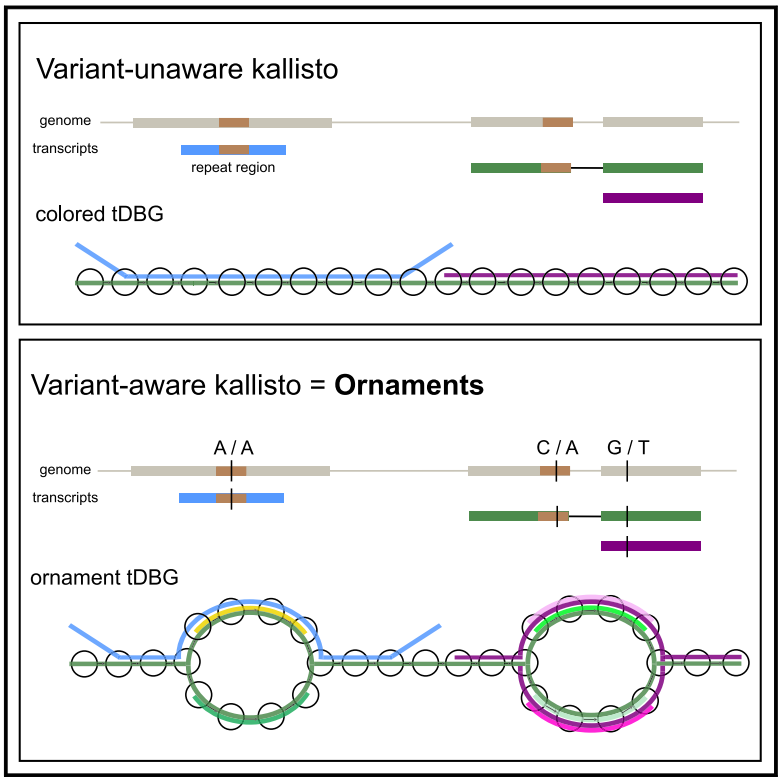
Ornaments for efficient allele-specific expression estimation with bias correction

Graphical abstract



Authors

Abhinav Adduri, Seyoung Kim

Correspondence

sssykim@acm.org

Allele-specific expression is essential for understanding gene regulation in diploid organisms. However, its estimation from RNA-seq is challenging due to allele-specific read mapping bias and computational cost. Here, we introduce Ornaments, a tool that corrects for the bias with high efficiency to estimate expression at heterozygous SNP and indel loci.

Adduri & Kim, 2024, The American Journal of Human Genetics *111*, 1770–1781
August 8, 2024

Ornaments for efficient allele-specific expression estimation with bias correction

Abhinav Adduri[1] and Seyoung Kim[2,*]

Summary

Allele-specific expression plays a crucial role in unraveling various biological mechanisms, including genomic imprinting and gene expression controlled by *cis*-regulatory variants. However, existing methods for quantification from RNA-sequencing (RNA-seq) reads do not adequately and efficiently remove various allele-specific read mapping biases, such as reference bias arising from reads containing the alternative allele that do not map to the reference transcriptome or ambiguous mapping bias caused by reads containing the reference allele that map differently from reads containing the alternative allele. We present Ornaments, a computational tool for rapid and accurate estimation of allele-specific transcript expression at unphased heterozygous loci from RNA-seq reads while correcting for allele-specific read mapping biases. Ornaments removes reference bias by mapping reads to a personalized transcriptome and ambiguous mapping bias by probabilistically assigning reads to multiple transcripts and variant loci they map to. Ornaments is a lightweight extension of kallisto, a popular tool for fast RNA-seq quantification, that improves the efficiency and accuracy of WASP, a popular tool for bias correction in allele-specific read mapping. In experiments with simulated and human lymphoblastoid cell-line RNA-seq reads with the genomes of the 1000 Genomes Project, we demonstrate that Ornaments improves the accuracy of WASP and kallisto, is nearly as efficient as kallisto, and is an order of magnitude faster than WASP per sample, with the additional cost of constructing a personalized index for multiple samples. Additionally, we show that Ornaments finds imprinted transcripts with higher sensitivity than WASP, which detects imprinted signals only at gene level.

Introduction

Allele-specific expression has been used to characterize various biological phenomena in diploid organisms, including gene expression affected by *cis*-acting variants in an allele-specific manner,[1,2] allele-specific nonsense-mediated mRNA decay,[3,4] and monoallelic expression of imprinted genes.[5,6] Allele-specific expression is typically measured as RNA sequencing (RNA-seq) read depths at heterozygous loci. There are two well-known biases introduced in allele-specific read mapping: reference bias and ambiguous mapping bias.[7–9] The reference bias arises from reads with alternative alleles that do not map to the reference transcriptome, leading to an underestimate of the expression of the alternative allele. The ambiguous mapping bias arises from reads that map to a heterozygous site but also to homozygous sites repeated in other genomic locations, since reads with the other allele at the same heterozygous site do not map to the same homozygous sites.

Removing these biases efficiently for accurate allele-specific expression estimation has been a challenging problem. Mapping reads to a diploid personalized transcriptome[10,11] removed only reference bias but not ambiguous mapping bias. WASP,[7] a popular tool for removing ambiguous mapping bias, was not adequate, as it simply discarded ambiguously mapped allele-specific reads, obtained allele-specific read counts only at gene level but not at transcript level, accounted for only SNPs but not indels, and was computationally expensive. RPVG[12] removed reference bias by mapping reads to a pantranscriptome constructed from a haplotype reference panel; however, it did not fully correct for ambiguous mapping bias and was computationally expensive. Kallisto,[13] a widely used tool for rapid transcriptome quantification, probabilistically assigned multi-mapped reads given a diploid transcriptome with known variant phases. However, with kallisto, RPVG, and other related methods,[14,15] inaccurate phasing could lead to inaccurate estimates of allele-specific signals.

Here, we introduce Ornaments, a tool for accurate and efficient estimation of allele-specific transcript expression at unphased heterozygous loci from RNA-seq reads. Ornaments removes reference bias by taking into account sample-specific variant information at SNP and indel loci and ambiguous mapping bias by probabilistically assigning reads to multiple transcripts and variant loci they map to.

Ornaments is a lightweight modification of kallisto[13] that improves upon the accuracy and efficiency of WASP, as well as the accuracy of kallisto, while leveraging the speed of kallisto. Ornaments modifies each of the two stages of kallisto, the read mapping and quantification stages. During the read mapping stage, Ornaments introduces an ornament transcriptome de Bruijn graph (tDBG), a key data structure that represents a personalized transcriptome within an ornament index. An ornament tDBG is obtained by augmenting the colored tDBG of kallisto with ornaments, each with two shades corresponding to the two variant alleles (Figures 1A

[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA; [2]Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA 15261, USA
*Correspondence: sssykim@acm.org
https://doi.org/10.1016/j.ajhg.2024.06.014.

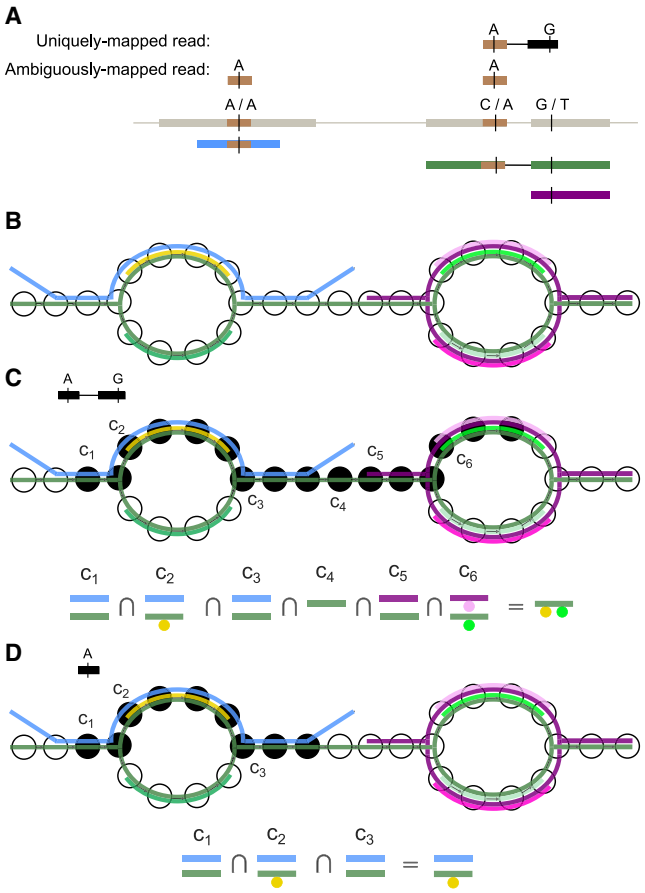


Figure 1. Ornaments overview

(A) The genome (gray) with three transcripts (blue, green, and purple), two heterozygous SNPs, one uniquely mapped allele-specific read (top read) and one ambiguously mapped allele-specific read (bottom read), and the repeat region (brown).

(B) The ornament tDBG constructed from the transcripts in (A). An ornament tDBG consists of a colored tDBG over *k*-mer nodes colored for transcripts and ornaments over *k*-mer nodes shaded for variant alleles.

(C) The variant-aware pseudoalignment of the uniquely mapped read. The filled *k*-mer nodes overlap with the read. Among the filled nodes, the first node of each contig is marked as $c_1, \ldots, c_6$, where a contig is defined as a sequence of *k*-mers annotated with the same set of [color, shade] pairs. The variant-aware equivalence class for the read is the set of [color, shade] pairs found from the intersection of the sets of colors for contigs and the union of shades associated with the colors in the intersection and is given as {[green color, yellow/lime shades]}.

(D) The variant-aware equivalence class of the ambiguously mapped read is {[blue color, ∅], [green color, yellow shade]}, where ∅ indicates no paired shades.

and 1B). Given the ornament index, Ornaments modifies the pseudoalignment of kallisto to variant-aware pseudoalignment to find a variant-aware equivalence class for each read, which we define as the set of transcripts and variant alleles that the read maps to (Figures 1C and 1D). In the quantification stage, Ornaments uses a variation of the mixture model and expectation-maximization (EM) algorithm of kallisto to obtain expected allele-specific read counts at heterozygous SNP and indel loci for each transcript in addition to transcript quantification. Using simulated and human lymphoblastoid cell-line RNA-seq reads[16] with the genetic variants of the 1000 Genomes Project samples,[17] we demonstrate that per sample, Ornaments is nearly as efficient as kallisto and an order of magnitude faster than WASP, with the additional cost of constructing a personalized index for each additional sample prior to read mapping. In addition, we show that Ornaments improves upon the accuracy of WASP by effectively removing allele-specific read mapping biases at transcript level and the accuracy of kallisto by accounting for genetic variants.

Material and methods

We introduce Ornaments in two stages for read mapping and quantification. We begin by describing how we prepare inputs to Ornaments, including variant information and ornament personalized transcriptome.

Preparing variant information

Given the reference transcriptome, transcript annotation, and sample-specific information on the genomic coordinates and alleles of variants, we prepare variant information as follows. First, we extract SNPs and indels in exonic regions and transform the genomic coordinates of those variants into transcriptomic coordinates. Then, for each variant, we store the information on the transcript it appears in, transcriptomic coordinates of the variant, and allele. A variant with a single genomic coordinate that appears in multiple alternatively spliced transcripts is associated with multiple transcripts and transcriptomic coordinates. If there are multiple overlapping indels at the same genomic coordinate, we keep only one of the indels. In our implementation, we keep the last one, as variants are typically provided in the order of genomic coordinates.

Preparing ornament personalized transcriptome

From the variant information above and reference transcriptome, we prepare an ornament personalized transcriptome, which will be used by Ornaments to build an ornament index over *k*-mer sequences. An ornament personalized transcriptome consists of transcript sequences and ornament sequences. The transcript sequences are set to those in the reference transcriptome with a modification to alternative allele at each alternative-allele homozygous locus and retain their transcript names in the reference transcriptome. Two ornament sequences are added for each pair of a heterozygous variant and transcript containing this variant. Each ornament sequence consists of the variant allele and the flanking sequences of length *k* on each side of the variant in the transcript. In rare cases of *n* heterozygous loci within *k* base pairs, where almost always $n = 2$, an ornament sequence is added for each of the 2^n combinations of the alleles. The ornament sequences are named as a concatenation of the name of the transcript of origin, position of the variant within the transcript, and allele.

The construction of ornament personalized transcriptome is efficient in both time and space. It takes only a few seconds to construct it, and its size is not significantly larger than the size of the reference transcriptome, since ornament sequences are substantially shorter than transcript sequences.

Constructing ornament index

Given the ornament personalized transcriptome, we construct an ornament index, which consists of an ornament tDBG and

ornament hash table. We augment the kallisto index, which consists of a colored tDBG and hash table, to represent variants in the ornament index.

To construct an ornament tDBG, we begin by applying the kallisto algorithm for constructing a colored tDBG to both transcript and ornament sequences. This assigns colors to both transcript and ornament sequences, creates the *k*-mer nodes of a tDBG from these sequences, and annotates each *k*-mer node with the set of colors corresponding to the sequences in which the *k*-mer appears. In our modification, we additionally ensure that the *k*-mer nodes from the ornament sequences are also annotated with the colors of the transcripts that gave rise to the given ornament sequences. Then, the kallisto algorithm proceeds to assign a set of colors to each contig, where a contig is defined as a sequence of *k*-mers annotated with the same set of colors between two junctions of the tDBG. We call the resulting tDBG an ornament tDBG, as the *k*-mers from the ornament sequences form decorative bubble-like structures in the tDBG whose top and bottom halves are additionally colored for the two alleles (Figures 1A and 1B). The colors of the ornament sequences are mapped to shades using an ornament hash table as we describe below.

Next, we construct an ornament hash table that maps a *k*-mer to a set of colors for transcripts and shades for variant alleles. An ornament hash table consists of a kallisto hash table, which maps each *k*-mer to a set of colors, and an auxiliary hash table, which further maps a color to an ornament shade if the color corresponds to an ornament sequence. The auxiliary hash table stores two pieces of information for each ornament shade: the variant allele and location in the transcript. Overall, the ornament hash table maps a *k*-mer to a set of pairs $[t, s]$ of a color t and shade s, where $s = \varnothing$, with $\varnothing$ indicating no ornament shades, if the *k*-mer is not found in any ornament sequences that originated from the transcript t.

Variant-aware pseudoalignment

Using the ornament index, Ornaments performs variant-aware pseudoalignment of reads to the personalized transcriptome. Variant-aware pseudoalignment is a modification of the pseudoalignment of kallisto to assign a read to a variant-aware equivalence class, which we define as the set of possible transcripts and variant alleles of origin for the given read. To obtain the variant-aware equivalence class of a read, we first map each *k*-mer of the read to colors and shades using the ornament hash table, mapping only the first *k*-mer of each contig and skipping to the first *k*-mer of the next contig for speed-up as in kallisto. Then, we combine these colors and shades across all *k*-mers of the read, by taking the intersection of the sets of colors for all *k*-mers as in the kallisto pseudoalignment and taking the union of the shades that are paired with the transcript colors in this intersection. A shade is included in the union, only if the read contains all *k*-mers of the given shade in the tDBG, except when the *k*-mers with the shade are located at either end of the read. The resulting variant-aware equivalence class of the read is a set of pairs $[t, s]$ of a color t and shade s that the read maps to (e.g., {[green color, yellow/lime shades]} for the uniquely mapped read in Figure 1C, and {[blue color, $\varnothing$], [green color, yellow shade]} for the ambiguously mapped read in Figure 1D).

Applying variant-aware pseudoalignment to all reads from a sample results in read counts for each of the variant-aware equivalence classes. These are the sufficient statistics needed for the quantification of transcript expression and allele-specific expression at heterozygous loci.

Quantification

Given the read counts for variant-aware equivalence classes, we modify kallisto to quantify allele-specific expression at heterozygous loci in addition to transcript expression. The key idea behind our modification is to first estimate transcript expression as in kallisto but in a variant-aware manner, followed by inferring expected allele-specific read counts at heterozygous loci. Below, we slightly re-cast the kallisto quantification method to provide the full set-up of a mixture model, as only the objective function for parameter estimation was explicitly stated for kallisto and the model set-up is not immediately obvious from the objective alone. Then, we describe our modification of kallisto in the three components of the statistical method: the model, estimation, and inference.

We describe the kallisto mixture model as a probability model for a random variable E representing the equivalence class of a read, which takes a value $e \in \mathbb{Q}$ for the set $\mathbb{Q}$ of all possible equivalence classes. The kallisto mixture model is

$$P(E = e) = \sum_{t=1}^{n_T} P(E = e|T = t)P(T = t), \qquad \text{(Equation 1)}$$

where T is a latent variable for an unobserved transcript label for the read, taking a value from $\{1, \ldots, n_T\}$ for n_T transcripts. The model above has the mixture proportion $P(T = t) = \theta_t$, where the parameter $\theta_t \in \boldsymbol{\theta} = \{\theta_1, \ldots, \theta_{n_T}\}$ represents unknown expression quantification for the transcript t and satisfies $\sum_{t=1}^{n_T} \theta_t = 1$, and the mixture component model

$$P(E = e|T = t) = \begin{cases} \frac{\ell_e}{\ell_t} & \text{if } t \in e, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{(Equation 2)}$$

where ℓ_t is the effective transcript length representing the possible number of starting positions of a read on the transcript t, ℓ_e is a subset of these starting positions on the transcript $t \in e$ that result in the given equivalence class e, and $\sum_{e \in \mathbb{Q}} P(E = e|T = t) = 1$. Notice that a non-zero $P(E = e|T = t)$ has an identical numerator for all t for a given e and has an identical denominator for all e for a given t. Equation 1 defines a generative model for the equivalence class of a read, where a transcript t is selected with the probability $P(T = t)$ and then given the transcript t, an equivalence class e is selected with the probability $P(E = e|T = t)$. We show below that it is not necessary to obtain ℓ_e explicitly, since these quantities cancel out and do not appear in the update equations of the EM algorithm.

The ornament mixture model has the same parameters $\boldsymbol{\theta}$ for the mixture proportions of n_T transcripts as the kallisto mixture model. However, the ornament mixture model is aware of variants, as its random variable E now represents the variant-aware equivalence class of a read, taking a value from the set $\mathbb{Q}$ of all possible variant-aware equivalence classes, and its mixture component model takes into account SNPs and indels. The ornament mixture component model extends that of kallisto in Equation 2 to

$$P(E = e|T = t) = \begin{cases} \frac{2\ell_e}{\ell_{t,m+p}} & \text{if } [t, \varnothing] \in e, \\ \frac{\ell_e}{\ell_{t,m+p}} & \text{if } [t, s] \in e \text{ for some } s \neq \varnothing, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(Equation 3)}$$

Above, $\ell_{t,m+p}$ is the combined effective transcript length of the maternal and paternal alleles of the transcript t, which

corresponds to the diploid length and reduces to $\ell_{t,m+p} = 2\ell_t$ for transcripts with no indels, and ℓ_e is defined as in Equation 2 but for a variant-aware equivalence class. In the numerator, the diploid length $2\ell_e$ is used, if a read pseudoaligns to both alleles of the transcript t, with no ornament shades paired with the transcript in e (e.g., the pair [blue color, $\varnothing$] in the variant-aware equivalence class of the read in Figure 1D). In contrast, the haploid length ℓ_e is used, if a read pseudoaligns to only one of the two alleles of the transcript t, with some ornament shades paired with it in e (e.g., [green color, yellow/lime shades] in Figure 1C and [green color, yellow shade] in Figure 1D). Overall, although the ornament mixture model does not model allele-specific expression directly as parameters, it is aware of variants through the probability in Equation 3: the probability is doubled for a read mapped to both transcript alleles at homozygous loci compared to a read mapped to only one allele at heterozygous loci and is adjusted based on the transcript lengths for transcripts with indels.

To estimate the parameters $\boldsymbol{\theta}$ for transcript expression, both Ornaments and kallisto use the EM algorithm.[18,19] However, they differ in that Ornaments is aware of variants in the EM algorithm via the modified mixture component model in Equation 3. Let e_i denote the variant-aware equivalence class of a read i, where $i = 1, \ldots, n_R$ for n_R reads. Instead of directly maximizing the data log likelihood, the EM algorithm for the ornament mixture model maximizes the expected complete-data log likelihood:

$$\mathbb{E}\left[\sum_{i=1}^{n_R} \log P(E = e_i|T = t)P(T = t)\right]$$

$$= \sum_{e\in\mathbb{Q}} |e| \sum_{t=1}^{n_T} P(T = t|E = e)\log P(E = e|T = t)P(T = t), \quad \text{(Equation 4)}$$

where $|e|$ represents the number of reads with e, or the sufficient statistics from the variant-aware pseudoalignment, and the expectation is taken with respect to the probability of the unobserved transcript labels for the reads given the observed variant-aware equivalence classes of the reads.

In each iteration of the EM algorithm, the E step computes the posterior probability $P(T = t|E = e)$ in Equation 4 for the unobserved T given the observed E, using the estimate $\widehat{\boldsymbol{\theta}}$ from the previous M step, and the M step maximizes Equation 4 to update $\widehat{\boldsymbol{\theta}}$, using $P(T = t|E = e)$ from the previous E step. Specifically, the E step computes the posterior probability

$$P(T = t|E = e) = \begin{cases} \dfrac{2\widehat{\theta}_t}{D_{t,e}} & \text{if } [t,\varnothing] \in e, \\ \dfrac{\widehat{\theta}_t}{D_{t,e}} & \text{if } [t,s] \in e \text{ for some } s \neq \varnothing, \\ 0 & \text{otherwise,} \end{cases} \quad \text{(Equation 5)}$$

where $D_{t,e} = \ell_{t,m+p} \cdot \left(\sum_{i:[i,\varnothing]\in e} \frac{2\widehat{\theta}_i}{\ell_{i,m+p}} + \sum_{i:[i,s]\in e, s\neq\varnothing} \frac{\widehat{\theta}_i}{\ell_{i,m+p}} \right)$ (see supplemental material and methods for derivation). This posterior probability can be viewed as a soft assignment of a read with the variant-aware equivalence class e to the transcript t. It also provides insights into how the EM algorithm handles an ambiguously mapped read in Ornaments, since the posterior probability of transcripts with no associated ornament shades in e is twice the probability of transcripts paired with ornament shades. Computing this posterior probability amounts to inferring the values of latent variables given data, a task carried out in the E step of the EM algorithm for latent-variable models in general.[18,19]

Given the posterior probabilities from the E step, the M step maximizes Equation 4 and updates the estimate as

$$\widehat{\theta}_t = \frac{1}{n_R} \sum_{e\in\mathbb{Q}} |e| P(T = t|E = e) \quad \text{(Equation 6)}$$

(see supplemental material and methods for derivation). The M-step update in Equation 6 is again aware of variants, as it uses the variant-aware posterior probabilities from the E step. Notice that ℓ_e appears in neither the E-step update in Equation 5 nor the M-step update in Equation 6 and thus is not needed to estimate $\boldsymbol{\theta}$. Convergence is called when the relative change for each θ_t is less than 0.1%.

Because allele-specific expression is not explicitly parameterized in the ornament mixture model, it is not directly estimated by the EM algorithm. Instead, given the estimate $\widehat{\boldsymbol{\theta}}$, it is inferred as the expected allele-specific read count at each heterozygous locus of each transcript, by computing the posterior probability in Equation 5 for each read and adding it across reads mapped to the locus. Specifically, at a heterozygous locus j on a transcript t, we compute the expected read depths d_{t,j_R} for the reference allele j_R and d_{t,j_A} for the alternative allele j_A as

$$d_{t,j_R} = \sum_{e\in\mathbb{Q}} |e| P(T = t|E = e)\mathbb{I}([t, j_R] \in e)$$

$$d_{t,j_A} = \sum_{e\in\mathbb{Q}} |e| P(T = t|E = e)\mathbb{I}([t, j_A] \in e),$$

where $\mathbb{I}(z)$ is an indicator function that outputs 1 if z = true and 0 if z = false.

Results

We benchmarked Ornaments against WASP, RPVG, and kallisto using simulated and lymphoblastoid cell-line RNA-seq reads[16] for 165 individuals with SNP genotypes from the 1000 Genomes Project.[17] These individuals were children in trios with known parental genotypes in the 1000 Genomes Project and thus with known haplotypes. We used the variants and the transcript annotation from GENCODE (version 36, Ensembl 102) to build ornament personalized transcriptomes. Low-quality RNA-seq reads that were too short or contained ambiguous nucleotides were removed using Trimmomatic 0.35.[20]

In all experiments, default settings were used for WASP, RPVG, and kallisto. For WASP, we used STAR for constructing an index[21] and the STAR re-implementation of WASP[22] for mapping and filtering reads. During the initial read mapping with the STAR aligner in WASP, to ensure that reads that map to SNP loci with alternative alleles are not dropped due to reference bias, we allowed reads to multi-map across up to 40 loci. Since the WASP re-mapping pipeline drops reads that are mapped to indels, in our comparison we did not include reads that Ornaments maps to indels and to SNPs within average read-length

distance or 100 bp of an indel. For RPVG, we used vg autoindex to construct a pantranscriptome and index given the genome, phased SNP genotypes, and transcript annotations and used vg mpmap in VG (v1.53.0 Valmontone) to produce multipath alignments. For quantification with RPVG (v1.0), we used the transcript inference mode to estimate transcript expression and the haplotype-transcript inference mode to estimate allele-specific transcript expression as read counts for a given haplotype. For comparison of RPVG against WASP and Ornaments, we divided this haplotype expression estimate from RPVG by the haplotype length and multiplied this by the read length to obtain expression estimates at heterozygous loci, as suggested by the authors.

For WASP, RPVG, and kallisto, we constructed a single index to be shared across all samples for read mapping, using the reference transcriptome for kallisto and WASP, and additionally using the variant information and known haplotypes for RPVG. For Ornaments, we constructed a personalized index for each sample.

Simulation

For simulation study, we selected 10 samples among the 165 children and generated 60 million RNA-seq reads for each sample from the phased diploid transcriptome of the sample and ground-truth allele-specific transcript abundances. These 10 samples spanned multiple ethnicities (The 1000 Genomes Project: HG00405, HG00526, HG00709, and HG00621 for East Asian ancestry; NA12766, NA12335, and NA07029 for European ancestry; and NA18869, NA18930, and NA19211 for African ancestry). The ground-truth allele-specific expression levels and background noise levels were set to the estimates obtained by applying RSEM[23] to the lymphoblastoid cell-line reads that were aligned to the personalized transcriptome using Bowtie 2.0.[24] The background noise was estimated to be 20% on average across samples. During simulation with the RSEM-simulate-reads program, we recorded reads overlapping with variants from which we inferred the ground-truth allele-specific read counts at each heterozygous locus.

We first compared the number of reads dropped by different methods, as these reads can affect accuracy (Figure S1). WASP dropped on average three times as many reads per sample as Ornaments, and kallisto with reference transcriptome dropped 16.3% more reads than Ornaments. RPVG dropped on average 11% more reads than Ornaments, though for half of the samples RPVG dropped fewer reads than Ornaments. Most (95.5%) of the reads dropped by WASP were ambiguously mapped allele-specific reads, whereas most of the reads dropped by kallisto, Ornaments, and RPVG were not allele-specific. Kallisto and Ornaments dropped reads mainly because pseudoalignment requires exact *k*-mer matches.

We compared Ornaments, WASP, and RPVG on the accuracy of allele-specific expression at heterozygous SNP loci. In our comparison of Ornaments with RPVG, transcript-level estimates were used, but in the comparison with WASP, gene-level estimates were used after aggregating transcript-level estimates from Ornaments over transcripts from the same gene. Accuracy was compared at SNP loci with and without ambiguously mapped allele-specific reads. A SNP was considered as involved in ambiguous read mapping if a variant-aware equivalence class from Ornaments that contains the [color, shade] pair for the given transcript/SNP pair also contains other transcript colors with no paired shades.

Ornaments outperformed WASP and RPVG in the accuracy of allele-specific expression, as it can correctly map and apportion ambiguously mapped allele-specific reads. At SNP loci with ambiguously mapped allele-specific reads, Ornaments had significantly lower mean absolute relative difference (MARD) for the estimated allele-specific expression and higher correlation between the true and estimated allelic ratios than both WASP and RPVG (Figures 2A and 2B). This in turn led to slightly higher accuracy for Ornaments at the other SNP loci without ambiguously mapped reads (Figures 2C and 2D). Unlike WASP, RPVG retained ambiguously mapped reads and used probabilistic approach to quantification. However, its quantification method had a limited capability to handle ambiguously mapped reads, as it made additional assumptions such as requiring the multiple transcripts involved in ambiguous read mapping to originate from the same haplotype. As a result, its accuracy was lower at loci that involve ambiguously mapped reads.

Ornaments achieved higher accuracy than WASP and RPVG in downstream analysis of detecting genes and transcripts with differentially expressed alleles (Figure 3). In our comparison of Ornaments with WASP, genes with differentially expressed alleles were identified by applying GeneiASE,[25] a tool that combines allele-specific signals across multiple loci within the same gene with unknown phases, to the gene-level estimates (p value < 0.05). In the comparison with RPVG, allele specifically expressed transcripts were identified by applying a negative-binomial test to the transcript-level estimates (p value < 0.05). Since RPVG requires known phases in the reference panel, the known phases were used to aggregate the estimates from Ornaments across multiple loci within the same transcript. Based on the detected genes and transcripts and the ground-truth allele-specific read counts, sensitivity and specificity were computed for each method. For nearly all samples, Ornaments had higher sensitivity and specificity for genes and transcripts that contained ambiguously mapped reads (Figures 3A and 3B), while this difference in accuracy was less for those that did not contain ambiguously mapped reads (Figures 3C and 3D). This suggests that highly accurate allele-specific signals from Ornaments can lead to higher accuracy in downstream analysis, compared to WASP and RPVG.

Ornaments achieved higher accuracy of transcript expression quantification than RPVG and variant-unaware kallisto, as Ornaments can correctly map and

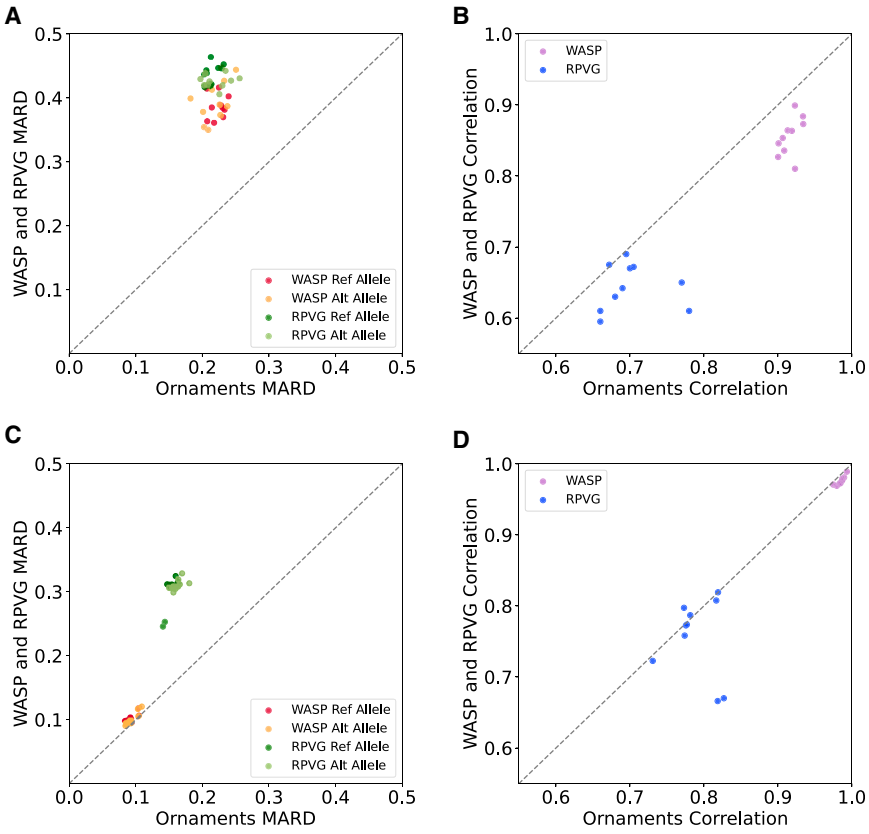


Figure 2. Comparison of Ornaments, WASP, and RPVG on the accuracy of allele-specific expression using simulated reads

Gene-level estimates were used to compare Ornaments with WASP, and transcript-level estimates were used to compare Ornaments with RPVG. At heterozygous loci with ambiguously mapped allele-specific reads, (A) the accuracy of allele-specific expression, measured as mean absolute relative difference (MARD) between the truth and estimate and (B) the accuracy of allelic ratios, measured as correlation between the true and estimated ratios across loci. At heterozygous loci without ambiguous allele-specific read mapping, (C) the accuracy of allele-specific expression and (D) the accuracy of allelic ratios. Each dot represents each of 10 samples. Only loci with true sequencing depth $\geq$10 reads were considered.

apportion ambiguously mapped allele-specific reads (Figure 4). Using the same simulated data above, the accuracy was measured as MARD on expressed transcripts (true abundance $>$ 0) and as mean absolute difference (MAD) on unexpressed transcripts. MAD was used for unexpressed transcripts, as MARD is known to be biased by small ground-truth values.[26] The accuracy was compared on transcripts with and without ambiguously mapped reads as well as with and without variants. Transcripts were considered as overlapping with ambiguously mapped reads, if they contained SNPs involved in ambiguous read mapping as determined by the variant-aware equivalence classes from Ornaments. Ornaments outperformed kallisto when transcripts contained ambiguously mapped allele-specific reads because variant-unaware kallisto cannot correctly map and apportion reads across the heterozygous and homozygous loci in repeat regions (Figures 4A, 4B, S2A, S2B, and S3). Ornaments had only slightly higher accuracy than kallisto for transcripts without ambiguously mapped reads even when the transcripts contained variants (Figures 4A, 4B, S2C, S2D, and S4). This is because kallisto maps reads to the same region regardless of the allele at heterozygous loci, skipping the mapping of the *k*-mer containing the SNP for efficiency if the *k*-mer is located between tDBG junctions. RPVG had lower accuracy than both kallisto and Ornaments for all types of transcripts, especially for transcripts that contained ambiguously mapped allele-specific reads (Figures 4C and 4D).

Next, we compared the computation time of the different methods. Since WASP, kallisto, and RPVG construct a single index for all samples, whereas Ornaments constructs a personalized index for each sample, we evaluated the computation time for both a single sample and multiple samples. Per sample, Ornaments was on average 11 times faster than WASP, 104 times faster than RPVG, and nearly as fast as kallisto (Figure 5A). Specifically, Ornaments required 9.2 min for constructing a personalized index and 8.7 min for quantification per sample, only slightly slower than kallisto, which took 9.2 and 8.4 min, respectively. WASP was significantly slower, taking 76 min for building a STAR index and 120.7 min for read alignment and quantification. RPVG took 33 h to construct a pantranscriptome and index, 3.1 h to align reads, and 18.2 min to quantify transcript haplotypes. Since kallisto, WASP, and RPVG use the same index for all samples, the time taken to construct the index is amortized as the sample size increases. However, even with a shared index, WASP and RPVG had a significantly higher cost of read alignment and quantification, and thus, for 10 samples, Ornaments was approximately 8 times faster than WASP and 20 times faster than RPVG (Figure 5B). For a very large number of samples, Ornaments is expected to retain its substantial advantage in efficiency over WASP and RPVG, approximately 7 times faster than WASP and 11 times faster than RPVG, and to require at most twice as much time as kallisto, since Ornaments and kallisto spend nearly the same amount of time on index construction and on read mapping and quantification. All computation times were obtained using 16 threads on a machine with two Intel Xeon 2.1GHz 8 core processors and 64 GB memory.

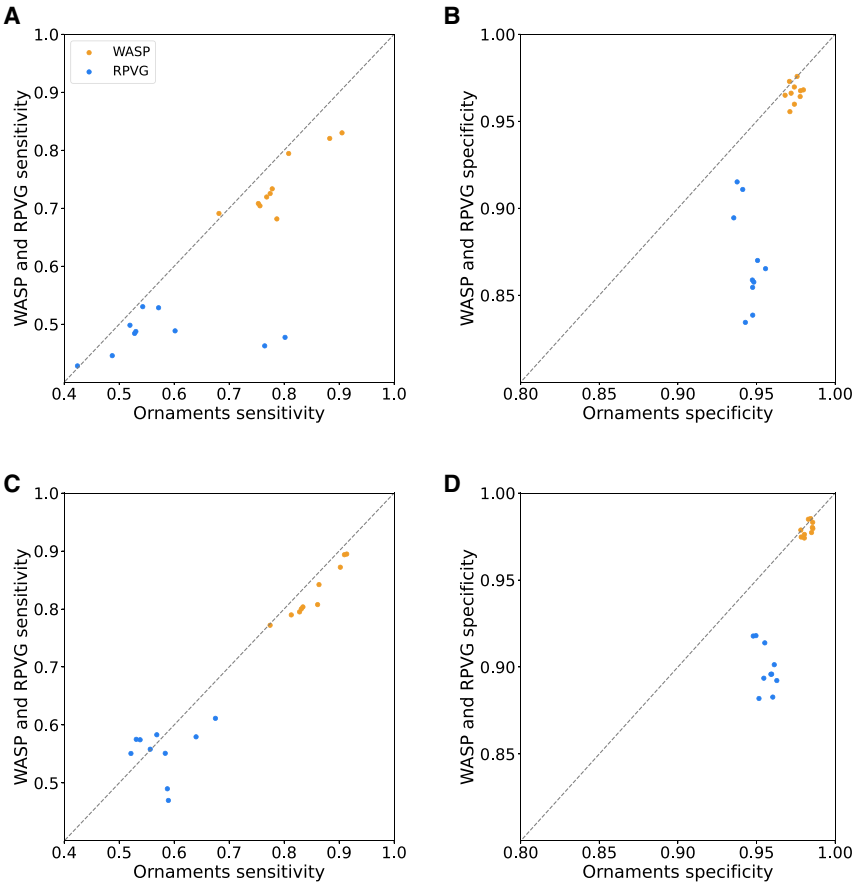


Figure 3. Comparison of Ornaments, WASP, and RPVG on the accuracy of detecting allele specifically expressed genes and transcripts using simulated reads

GeneiASE with gene-level estimates (p value <0.05) were used to compare Ornaments with WASP, and negative-binomial tests with transcript-level estimates (p value <0.05) along with the known haplotypes were used to compare Ornaments with RPVG. The detected genes and transcripts were compared against those obtained from the ground-truth allele-specific read counts. For transcripts and genes with variants that induce ambiguous allele-specific read mapping, (A) sensitivity and (B) specificity of the methods. For transcripts and genes without ambiguously mapped allele-specific reads, (C) sensitivity and (D) specificity of the methods. Each dot represents each of 10 samples.

Human lymphoblastoid cell-line RNA-seq reads

Using the lymphoblastoid cell-line reads and genome sequences for 165 children from the 1000 Genomes Project,[16] we benchmarked Ornaments against WASP. We omitted RPVG in this experiment due to its high computational cost for processing a large number of samples. We compared Ornaments and WASP in terms of the allele-specific expression and allelic ratios at heterozygous loci, using gene-level summaries from Ornaments. For both expression and ratios, the correlation between Ornaments and WASP was lower at the heterozygous loci overlapping with ambiguously mapped reads than at the other heterozygous loci (Figure 6). This result provides evidence for the superior ability of Ornaments to correct for ambiguous mapping bias.

To evaluate the impact of estimates from the different methods on downstream analysis, we compared allele specifically expressed transcripts identified by Ornaments and GeneiASE with allele specifically expressed genes identified by WASP and GeneiASE.[25] The transcripts found by Ornaments included the majority of the genes found by WASP and a large number of additional genes. Specifically, Ornaments found 4,374 genes with differentially expressed alleles in at least one constituent transcript of the gene in at least 10 samples, whereas WASP identified only 1,034 genes in at least 10 samples. Out of the 1,034 genes from WASP, 897 genes were also found by Ornaments. This suggests higher sensitivity of Ornaments in downstream analysis, as Ornaments detects allele-specific signals at transcript level with highly accurate bias correction.

To see if these allele specifically expressed transcripts from Ornaments and genes from WASP can reproduce known biological results, we examined these transcripts and genes that overlap with previously known imprinted genes, where one allele is exclusively expressed over the other. We compiled a set of 157 imprinted genes that are either known to undergo imprinting or found to be imprinted in an independent dataset. Our set included 141 genes in the GeneImprint database,[27] 13 genes identified from analysis of lymphoblastoid cell-line RNA-seq data for 80 individuals with European ancestry[28] and for 63 unrelated individuals,[29] and other known imprinted genes from the literature.[30–32] Ornaments had 34 genes in the overlap, whereas WASP had a smaller overlap of 30 genes (Figure 7). For 24 out of the 29 imprinted genes found by both methods, Ornaments called differential expression in more samples (Table 1). Furthermore, Ornaments detected a subset of transcripts, on average one or two transcripts, per gene as imprinted in a given sample (Table 1; Figures 7, S5, and S6). In many such cases, WASP failed to detect the imprinting signal, as the signal was lost or weaker at the gene level. These findings provide evidence that Ornaments, with its probabilistic approach, can accurately attribute the allele-specific signals to multiple transcripts of the given gene, offering advantages over WASP, which captures signals at gene level.

Discussion

We introduced Ornaments, a computational tool for accurate and efficient quantification of transcript expression

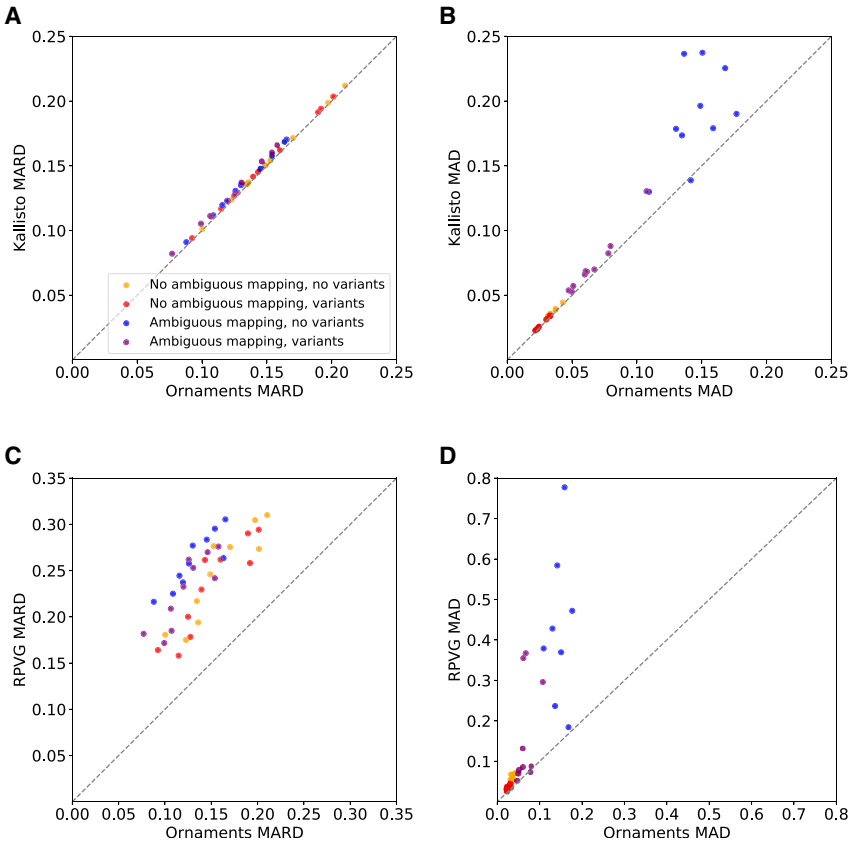


Figure 4. Comparison of Ornaments, kallisto, and RPVG on the accuracy of transcript quantification using simulated reads
The accuracy of Ornaments and kallisto is compared for (A) expressed transcripts in mean absolute relative difference (MARD) and (B) unexpressed transcripts in mean absolute difference (MAD). The accuracy of Ornaments and RPVG is compared for (C) expressed transcripts in MARD and (D) unexpressed transcripts in MAD. Each colored dot represents each of the 10 samples for each of the four types of transcripts (shown with colors) with and without ambiguously mapped reads and variants.

and allele-specific expression at unphased heterozygous loci from RNA-seq reads. Ornaments is an adaptation of kallisto that takes advantage of the speed of kallisto while improving the accuracy of the existing methods by accounting for variants, by correcting for allele-specific read mapping biases, and by capturing allele-specific signals at transcript level rather than at gene level.

One important future direction is to extend Ornaments to construct a single population-level index from multiple samples, rather than a personalized index, to further reduce computation time. Such a population-level index would have a modified ornament tDBG that represents all variants that are heterozygous in one or more samples. Then, for variant-aware pseudoalignment, we could use either a personalized index derived from the population index or the population index directly. We expect these two approaches to have a different trade-off between accuracy and speed. The former approach would have the same accuracy as Ornaments but incur the computational cost for modifying the tDBG to remove the ornaments for variants that are not heterozygous in the given sample. The latter approach could have lower accuracy for a large population with dense polymorphic loci because in genomic regions with densely packed variants, with more junctions in the population-level ornament tDBG, more *k*-mers would be checked for exact sequence matches, leading to more reads being dropped due to sequencing errors. However, this approach could be implemented efficiently with a modified variant-aware pseudoalignment that checks for exact *k*-mer matches at the tDBG junctions around ornaments only if the corresponding locus is heterozygous in the given sample.

Another potential future direction is to extend Ornaments to build an index from a haplotype reference panel as in RPVG. This would enable Ornaments to quantify allele-specific expression for a sample solely using

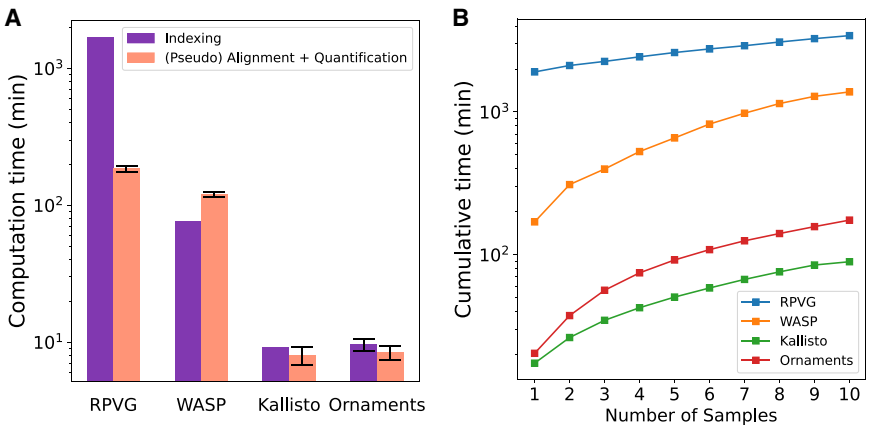


Figure 5. Computation time of Ornaments and other methods
(A) Computation time for a single sample. In RPVG, WASP, and kallisto, the indexing cost (purple) is incurred once, as the same index is re-used for multiple samples, whereas in Ornaments, a personalized index is constructed repeatedly for each sample. In all methods, the cost of alignment and quantification (pink) is incurred for each sample and is shown as an average over 10 simulated samples with error bars for one standard deviation.
(B) Cumulative computation time as the sample size increases.

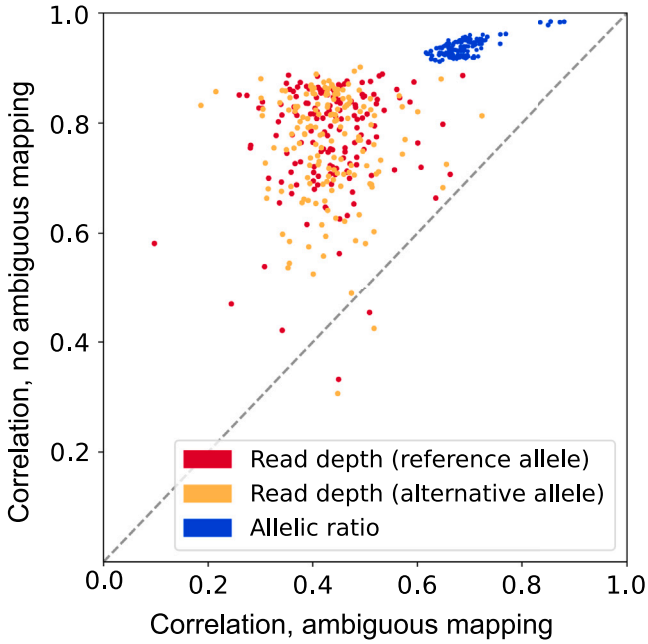


Figure 6. Comparison of Ornaments and WASP on lymphoblastoid cell-line RNA-seq reads

Correlation between Ornaments and WASP estimates across SNP loci with ambiguously mapped reads (x axis) and without ambiguously mapped reads (y axis). Each dot represents the correlation for each of 165 samples.

RNA-seq reads without requiring genotype data. This could be accomplished by extending the ornament tDBG such that a haplotype with multiple variants in the reference panel is assigned a color and is represented as a string with multiple ornaments. Overall, Ornaments is a flexible tool that could be extended in various ways for allele-specific expression quantification.

Data and code availability

This study did not generate datasets. The code for Ornaments that was generated during this study is available at a GitHub repository https://github.com/SeyoungKimLab/Ornaments, along with installation instructions.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.06.014.

Acknowledgments

The authors thank the anonymous reviewers for their insightful comments. This work was supported by NIH-1R21HG011116, NIH-1R21HG010948, and NSF-DBI2154089.

Author contributions

A.A. and S.K. conceived the project, developed the method, designed the experiments, and wrote the paper. A.A. performed the experiments.

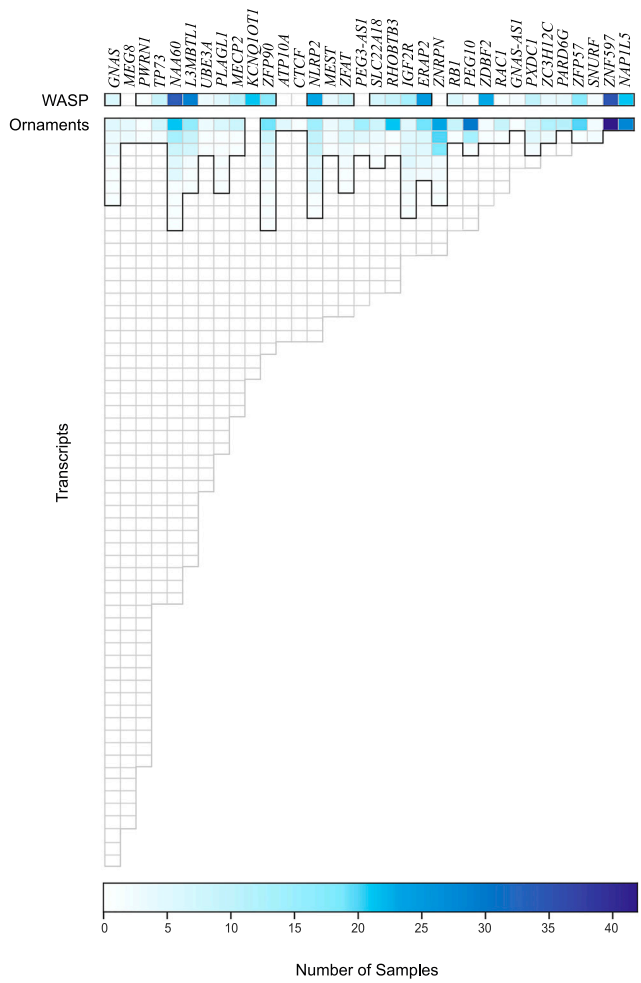


Figure 7. Known imprinted genes overlapping with allele specifically expressed transcripts from Ornaments and genes from WASP in lymphoblastoid cell line

Each cell shows the number of samples in which differential expression between two alleles was found by WASP for the given gene or by Ornaments for the given transcript. The cells with one or more individuals are enclosed in black lines. *SNHG14* with 158 transcripts, omitted in the figure, is a known imprinted gene that had allele specifically expressed transcripts in 82 samples for Ornaments but only in 47 samples for WASP.

Declaration of interests

The authors declare no competing interests.

Received: September 27, 2023
Accepted: June 24, 2024
Published: July 23, 2024

Web resources

1000 Genomes Project, http://www.1000genomes.org
BLAST, https://blast.ncbi.nlm.nih.gov/Blast.cgi
dbSNP, http://www.ncbi.nlm.nih.gov/snp/
Ensembl, http://useast.ensembl.org/Homo_sapiens/Gene
GENCODE, https://www.gencodegenes.org/
OMIM, http://www.omim.org/

Table 1. Known imprinted genes found by WASP and Ornaments in lymphoblastoid cell lines

Gene	WASP	Ornaments	WASP and	Ornaments	Source
	Number of samples	Number of samples[a]	Ornaments (overlap in samples)	Average number of transcripts per sample	
GNAS (MIM: 139320)	6	12	4	1.25	GeneImprint[27]
MEG8 (MIM: 613648)	0	7	0	1.00	Charlier et al.[33,34]
PWRN1 (MIM: 611215)	1	2	1	1.00	Wawrzik et al.[33,35]
TP73 (MIM: 601990)	9	9	2	1.33	GeneImprint[27]
NAA60 (MIM: 614246)	34	42	25	1.33	Jadhav et al.[16]
L3MBTL1 (MIM: 608802)	29	21	9	1.71	Jadhav et al.[16]
UBE3A (MIM: 601623)	3	5	2	1.00	Sadikovic et al.[33,36]
PLAGL1 (MIM: 603044)	5	10	4	1.10	Kas et al.[33,37]
MECP2 (MIM: 300005)	11	10	6	1.00	Nakashima et al.[33,38]
KCNQ1OT1 (MIM: 604115)	21	0	0	0.00	Cagle et al.[33,39]
ZFP90 (MIM: 609451)	17	28	12	1.50	GeneImprint[27]
ATP10A (MIM: 605855)	0	5	0	1.00	Pastinen et al.[29]
CTCF (MIM: 604167)	0	1	0	1.00	Rubio et al.[33,40]
NLRP2 (MIM: 609364)	23	28	10	1.75	Meyer et al.[41]
MEST (MIM: 601029)	4	5	3	1.20	Pastinen et al.[29]
ZFAT (MIM: 610931)	8	17	4	1.12	Pilvar et al.[42]
PEG3 (MIM: 601483)	0	20	0	1.00	GeneImprint[27]
SLC22A18 (MIM: 602631)	7	13	3	1.08	GeneImprint[27]
RHOBTB3 (MIM: 607353)	9	23	7	1.00	GeneImprint[27]
IGF2R (MIM: 604893)	13	23	6	1.17	Kukuvitis et al.[33,43]
ERAP2 (MIM: 609497)	25	29	13	1.20	GeneImprint[27]
SNRPN (MIM: 182279)	0	35	0	1.91	Jadhav et al.[16]
RB1 (MIM: 614041)	6	10	6	1.00	GeneImprint[27]
PEG10 (MIM: 609810)	2	38	1	1.16	Jadhav et al.[16]
ZDBF2 (MIM: 617059)	23	2	1	1.50	GeneImprint[27]
RAC1 (MIM: 602048)	1	10	1	1.00	GeneImprint[27]
GNAS-AS1 (MIM: 610540)	2	1	0	1.00	Jadhav et al.[16]
PXDC1	10	13	7	1.08	GeneImprint[27]
ZC3H12C (MIM: 615001)	4	14	2	1.07	GeneImprint[27]
PARD6G (MIM: 608976)	4	11	3	1.00	GeneImprint[27]
ZFP57 (MIM: 612192)	17	20	15	1.05	Mackay et al.[33,44]
SNURF (MIM: 182279)	2	4	1	1.00	Jadhav et al.[16]
ZNF597 (MIM: 614685)	35	42	28	1.00	Jadhav et al.[16]
NAP1L5 (MIM: 612203)	21	28	21	1.00	GeneImprint[27]
SNHG14 (MIM: 616259)	47	82	32	2.27	Jadhav et al.[16]

[a]Number of samples with at least one imprinted transcript for a gene.

References

1. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. *48*, 206–213. https://doi.org/10.1038/ng.3467.
2. Zhabotynsky, V., Huang, L., Little, P., Hu, Y.-J., Pardo-Manuel de Villena, F., Zou, F., and Sun, W. (2022). eQTL mapping using allele-specific count data is computationally feasible, powerful, and provides individual-specific estimates of genetic effects. PLoS Genet. *18*, e1010076. https://doi.org/10.1371/journal.pgen.1010076.
3. Lappalainen, T., Sammeth, M., Friedländer, M.R., Ac't Hoen, P., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511. https://doi.org/10.1038/nature12531.
4. Lindeboom, R.G., Supek, F., and Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. Nat. Genet. *48*, 1112–1118. https://doi.org/10.1038/ng.3664.
5. Knight, J.C. (2004). Allele-specific gene expression uncovered. Trends Genet. *20*, 113–116. https://doi.org/10.1016/j.tig.2004.01.001.
6. Buchroithner, B., Klausegger, A., Ebschner, U., Anton-Lamprecht, I., Pohla-Gubo, G., Lanschuetzer, C.M., Laimer, M., Hintner, H., and Bauer, J.W. (2004). Analysis of the LAMB3 gene in a junctional epidermolysis bullosa patient reveals exonic splicing and allele-specific nonsense-mediated mRNA decay. Lab. Invest. *84*, 1279–1288. https://doi.org/10.1038/labinvest.3700164.
7. Van De Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063. https://doi.org/10.1038/nmeth.3582.
8. Chen, J., Rozowsky, J., Galeev, T.R., Harmanci, A., Kitchen, R., Bedford, J., Abyzov, A., Kong, Y., Regan, L., and Gerstein, M. (2016). A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat. Commun. *7*, 11101–11113. https://doi.org/10.1038/ncomms11101.
9. Stevenson, K.R., Coolon, J.D., and Wittkopp, P.J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genom. *14*, 1–13. https://doi.org/10.1186/1471-2164-14-536.
10. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. *7*, 522. https://doi.org/10.1038/msb.2011.54.
11. Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*, 195. https://doi.org/10.1186/s13059-015-0762-6.
12. Sibbesen, J.A., Eizenga, J.M., Novak, A.M., Sirén, J., Chang, X., Garrison, E., and Paten, B. (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nat. Methods *20*, 239–247. https://doi.org/10.1038/s41592-022-01731-9.
13. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.
14. Nariai, N., Kojima, K., Mimori, T., Kawai, Y., and Nagasaki, M. (2016). A Bayesian approach for estimating allele-specific expression from RNA-seq data with diploid genomes. BMC Genom. *17*, 2–17. https://doi.org/10.1186/s12864-015-2295-5.
15. Raghupathy, N., Choi, K., Vincent, M.J., Beane, G.L., Sheppard, K.S., Munger, S.C., Korstanje, R., Pardo-Manual de Villena, F., and Churchill, G.A. (2018). Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. Bioinformatics *34*, 2177–2184. https://doi.org/10.1093/bioinformatics/bty078.
16. Jadhav, B., Monajemi, R., Gagalova, K.K., Ho, D., Draisma, H.H., van de Wiel, M.A., Franke, L., Heijmans, B.T., van Meurs, J., Jansen, R., et al. (2019). RNA-seq in 296 phased trios provides a high-resolution map of genomic imprinting. BMC Biol. *17*, 50. https://doi.org/10.1186/s12915-019-0674-0.
17. The 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
18. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B *39*, 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
19. Bilmes, J.A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute *4*, 126.
20. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
21. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.
22. Asiimwe, R., and Dobin, A. (2024). STAR+WASP reduces reference bias in the allele-specific mapping of RNA-seq reads. Preprint at bioRxiv. https://doi.org/10.1101/2024.01.21.576391.
23. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.
24. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
25. Edsgärd, D., Iglesias, M.J., Reilly, S.-J., Hamsten, A., Tornvall, P., Odeberg, J., and Emanuelsson, O. (2016). GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. Sci. Rep. *6*, 1–13. https://doi.org/10.1038/srep21134.
26. Yi, L., Liu, L., Melsted, P., and Pachter, L. (2018). A direct comparison of genome alignment and transcriptome pseudoalignment. Preprint at bioRxiv. https://doi.org/10.1101/444620.
27. Jirtle R.L. GeneImprint. https://www.geneimprint.com.
28. Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., Bibikova, M., Chudin, E., Barker, D.L., Dickinson, T., et al. (2008). Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. PLoS Genet. *4*, e1000006. https://doi.org/10.1371/journal.pgen.1000006.

29. Pastinen, T., Sladek, R., Gurd, S., Ge, B., Lepage, P., Lavergne, K., Villeneuve, A., Gaudin, T., Brändström, H., Beck, A., et al. (2004). A survey of genetic and epigenetic variation affecting human gene expression. Physiol. Genom. *16*, 184–193. https://doi.org/10.1152/physiolgenomics.00163.2003.
30. Reik, W., and Walter, J. (2001). Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. *2*, 21–32. https://doi.org/10.1038/35047554.
31. Meguro, M., Kashiwagi, A., Mitsuya, K., Nakao, M., Kondo, I., Saitoh, S., and Oshimura, M. (2001). A novel maternally expressed gene, ATP10C, encodes a putative aminophospholipid translocase associated with angelman syndrome. Nat. Genet. *28*, 19–20. https://doi.org/10.1038/ng0501-19.
32. Nakabayashi, K., Bentley, L., Hitchins, M.P., Mitsuya, K., Meguro, M., Minagawa, S., Bamforth, J.S., Stanier, P., Preece, M., Weksberg, R., et al. (2002). Identification and characterization of an imprinted antisense RNA (MESTIT1) in the human MEST locus on chromosome 7q32. Hum. Mol. Genet. *11*, 1743–1756. https://doi.org/10.1093/hmg/11.15.1743.
33. Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., et al. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr. Protoc. Bioinformatics *54*, 1.30.1–1.30.33. https://doi.org/10.1002/cpbi.5.
34. Charlier, C., Segers, K., Wagenaar, D., Karim, L., Berghmans, S., Jaillon, O., Shay, T., Weissenbach, J., Cockett, N., Gyapay, G., and Georges, M. (2001). Human–ovine comparative sequencing of a 250-kb imprinted domain encompassing the callipyge (clpg) locus and identification of six imprinted transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8. Genome Res. *11*, 850–862. https://doi.org/10.1101/gr.172701.
35. Wawrzik, M., Spiess, A.-N., Herrmann, R., Buiting, K., and Horsthemke, B. (2009). Expression of SNURF–SNRPN upstream transcripts and epigenetic regulatory genes during human spermatogenesis. Eur. J. Hum. Genet. *17*, 1463–1470. https://doi.org/10.1038/ejhg.2009.83.
36. Sadikovic, B., Fernandes, P., Zhang, V.W., Ward, P.A., Miloslavskaya, I., Rhead, W., Rosenbaum, R., Gin, R., Roa, B., and Fang, P. (2014). Mutation update for UBE3A variants in angelman syndrome. Hum. Mutat. *35*, 1407–1417. https://doi.org/10.1002/humu.22687.
37. Kas, K., Voz, M.L., Hensen, K., Meyen, E., and Van de Ven, W.J. (1998). Transcriptional activation capacity of the novel PLAG family of zinc finger proteins. J. Biol. Chem. *273*, 23026–23032. https://doi.org/10.1074/jbc.273.36.23026.
38. Nakashima, N., Yamagata, T., Mori, M., Kuwajima, M., Suwa, K., and Momoi, M.Y. (2010). Expression analysis and mutation detection of DLX5 and DLX6 in autism. Brain Dev. *32*, 98–104. https://doi.org/10.1016/j.braindev.2008.12.021.
39. Cagle, P., Qi, Q., Niture, S., and Kumar, D. (2021). KCNQ1OT1: an oncogenic long noncoding RNA. Biomolecules *11*, 1602. https://doi.org/10.3390/biom11111602.
40. Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin*105*, pp. 8309–8314. https://doi.org/10.1073/pnas.0801273105.
41. Meyer, E., Lim, D., Pasha, S., Tee, L.J., Rahman, F., Yates, J.R., Woods, C.G., Reik, W., and Maher, E.R. (2009). Germline mutation in NLRP2 (NALP2) in a familial imprinting disorder (beckwith-wiedemann syndrome). PLoS Genet. *5*, e1000423. https://doi.org/10.1371/journal.pgen.1000423.
42. Pilvar, D., Reiman, M., Pilvar, A., and Laan, M. (2019). Parent-of-origin-specific allelic expression in the human placenta is limited to established imprinted loci and it is stably maintained across pregnancy. Clin. Epigenet. *11*, 94. https://doi.org/10.1186/s13148-019-0692-3.
43. Kukuvitis, A., Georgiou, I., Syrrou, M., Andronikou, S., Dickerman, Z., Islam, A., McCann, J., and Polychronakos, C. (2004). Lack of association of birth size with polymorphisms of two imprinted genes, IGF2R and GRB10. J. Pediatr. Endocrinol. Metab. *17*, 1215–1220. https://doi.org/10.1515/JPEM.2004.17.9.1215.
44. Mackay, D.J., Callaway, J.L., Marks, S.M., White, H.E., Acerini, C.L., Boonen, S.E., Dayanikli, P., Firth, H.V., Goodship, J.A., Haemers, A.P., et al. (2008). Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. Nat. Genet. *40*, 949–951. https://doi.org/10.1038/ng.187.